AN ABSTRACT OF THE THESIS OF

Leigh Ann H. Starcevich for the degree of Doctor of Philosophy in Statistics presented on August 9, 2010.
Title: Propensity Score Methodology for Nonignorable Nonresponse

Abstract approved:

Virginia M. Lesser

When data are not missing at random, approaches to reduce nonresponse bias include subsampling nonresponding units and modeling. The objective of this thesis is to develop unbiased and precise model-assisted estimators of the population total that are applicable to data from a complex survey design with nonignorable nonresponse. When information from a nonrespondent subsample is available, weighting methods for missing-at-random data may be modified to reduce bias from nonignorable missingness in estimates of population totals. Propensity score methodology for nonignorable missingness is developed for use with the weighting class adjustment and with the Horvitz-Thompson estimator to account for the dependence between the outcome of interest and the response mechanism. The novel propensity score techniques for nonignorable nonresponse are applied to a binary outcome subject to nonignorable missingness from a complex survey of elk hunters and are also examined with simulation.

Propensity Score Methodology for Nonignorable Nonresponse

by
Leigh Ann H. Starcevich

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented August 9, 2010
Commencement June 2011

Doctor of Philosophy thesis of Leigh Ann H. Starcevich presented on August 9, 2010.

APRROVED:

Major Professor, representing of Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Leigh Ann H. Starcevich, Author

ACKNOWLEDGEMENTS

The author expresses sincere appreciation to the following:

Ginny Lesser for making sampling so interesting, for being persistent and patient, and for encouraging me to not miss out on life while achieving this goal.

My other committee members from the statistics department (Dave Birkes, Alix Gitelman, and Paul Murtaugh) for the great contribution each of them makes to our department.

Peter McEvoy for being my Graduate School representative and for sending me Evrim.

N. Scott Urquhart for his mentorship and support of my statistical consulting career.

New Mexico Department of Game and Fish personnel Darrel Weybright and Kerry Mower, for their funding of the nonresponse bias study, and Stewart Liley, for providing estimates from the mandatory elk licensee survey.

Kirk Steinhorst and Kathryn Irvine for being a great and stimulating consulting team and for their extra help during the final push.

The Babes for monthly sisterhood and lots of laughs.

My parents, A.W. and Ann Harrod, for never doubting that I would finish.

My husband, Steve Starcevich, for his constant support and encouragement, wonderful meals, exuberant fathering, and endless patience.

My daughter, Lola, for dancing every day.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Survey nonresponse occurs when a complete response cannot be obtained for a portion of the sample (Lessler and Kalsbeek, 1992; Lohr, 1999; Little and Rubin, 2002). In a survey involving the use of questionnaires, missing data may result from the inability to contact a person chosen for the sample or from refusal to participate in the survey. In an ecological survey, selected sites may be inaccessible due to rough terrain or because a landowner denies access. If the reason for the missingness is unrelated to the survey design, the outcome of interest, or any covariates, then ignoring the nonresponse should still provide valid inference (Little and Rubin, 2002). However, if the reason for the missingness is related to any of these, ignoring the nonresponse may lead to invalid inference in the form of biased estimates and poor confidence interval coverage (Särndal, Swensson, and Wretman, 1992).

When the mechanism causing the missingness is related to the outcome of interest, the nonresponse is called *nonignorable* (Little and Rubin, 2002). In this case, the information obtained from respondents is inadequate to provide unbiased estimation. Model-based approaches are used to explicitly describe the relationship between the response mechanism and the outcome of interest under the assumption of nonignorable nonresponse. However, model-based inference adds more complexity to the analysis and relies on accurate model specification (Särndal et al., 1992). When feasible, a nonrespondent subsample provides information about the subpopulation of nonrespondents (Bartholomew, 1961; Hansen and Hurwitz, 1946; Rao, 1983).

In many large governmental surveys with available list frames for the population units, surveys are conducted periodically to monitor parameters of interest. However, nonrespondent subsamples require additional effort and cost that may not be feasible for every survey. Intermittent nonrespondent subsampling could be used to accurately model the relationship between the response mechanism and the outcome of interest. It may be feasible to apply the nonresponse model to data from similar surveys conducted on the same population but at different times. These studies are also subject to nonresponse but not augmented by a nonrespondent subsample. The objective of the work in this thesis is to obtain a model-assisted estimator of the population total that minimizes mean square error and is applicable to data from a complex survey design with nonignorable nonresponse.

In Chapter 2, a general background on nonresponse error is first provided, and the mechanisms generating missing data are described. Methods for nonresponse adjustment are discussed with an emphasis on nonresponse weighting. The primary approach used in this work, propensity score methodology, is reviewed in detail.

Propensity score methodology is extended to the case of nonignorable nonresponse in Chapter 3, providing a new application of propensity score methodology. Assuming a binary outcome of interest and estimation of the population total, propensity scores estimated from the combined information from respondents and subsampled nonrespondents are used to create adjustment classes for a weighting technique called the *weighting class adjustment* (Oh and Scheuren, 1983). Other approaches for forming weighting classes are examined under a range of assumptions and compared to the novel approach incorporating nonignorable nonresponse. The performances of the estimators are compared with data from a survey of about 40,000 New Mexico elk hunters.

In Chapter 4, the propensity score methodology from Chapter 3 is applied to the weighting approach examined by Cassel, Särndal, and Wretman (1983) in which the reciprocal of the propensity score is used to weight respondent outcomes for nonresponse. Propensity score methodology for nonignorable nonresponse is applied to a modified Horvitz-Thompson (1952) estimator, referred to as the *propensity score adjustment estimator,* using three approaches to model the response probability estimation. The estimators are compared to the weighting class adjustment estimators from Chapter 3 using the New Mexico case study data by comparing the relative bias, confidence interval coverage, and root mean square error of the estimates.

Simulations comparing all estimators discussed in Chapters 3 and 4 are summarized in Chapter 5. In contrast to the large population size observed in the case study, the simulations assess the performance of the estimators for a smaller population of 1,000 units assuming a binary response. Simulation scenarios incorporate a range of response rates, success rates, and odds ratios of response relative to successful and unsuccessful units. Furthermore, response rates and/or success rates are allowed to vary between the modeling data set and the verification data set so that model robustness may be examined. The performances of the estimators are compared using relative bias, confidence interval coverage, and root mean square error.

In Chapter 6, the results of the pilot data analysis and the simulations are compared and discussed. In recent years, the New Mexico Department of Game and Fish has penalized nonrespondents by prohibiting participation in the subsequent year's hunt. This response incentive has nearly tripled the response rate and provides more complete data for interpretation of the pilot data analysis results. Conclusions regarding the utility and feasibility of this work are stated, and directions for future work are proposed.

CHAPTER 2: LITERATURE REVIEW

2.1 INTRODUCTION

Two types of survey error affect the accuracy and precision of estimates: sampling error and nonsampling error (Lessler and Kalsbeek, 1992; Lohr, 1999; Särndal et al., 1992; Thompson, 1992). Sampling error is caused by the uncertainty that is generated from surveying only a portion of the population, and nonsampling error arises from the imperfect execution of a sampling design. Nonresponse error is a type of nonsampling error that occurs when data are missing from a survey. Nonresponse error may cause the nonrandom exclusion of a portion of the target population (Cassel et al., 1983). The failure to obtain a complete response from every sampling unit may produce biased estimates if the missing outcomes differ substantially from the observed outcomes. In this chapter, a review of relevant literature provides the motivation for the nonresponse bias adjustment approaches proposed in subsequent chapters.

Nonresponse error affects inference from surveys across many subject areas such as wildlife management (MacDonald and Dillman, 1968; Pendleton, 1992), economics

(Detlefsen, 2007), and the National Crime Survey of the U.S. Census Bureau (Bailey, 1986). Nonresponse rates as low as 10% can affect inference (Lohr, 1999). An estimated undercount of 1.18% in the 2000 U.S. Census prompted nonresponse adjustments to account for disparities in response rates among different minority groups (US Census Monitoring Board Presidential Members, 2001). Biased inference due to nonresponse may adversely affect policy and management decisions (Kauff, Olsen, and Fraker, 2002; Pickreign and Gabel, 2005). The only assurance that a researcher has to be certain that nonresponse error does not cause misleading inference is to fully observe the sample (Kalton and Kasprzyk, 1986). However, avoiding nonresponse is not always possible. The goal of this review is to summarize the research on nonresponse error and to establish the basis for proposing new methods of error reduction for unbiased estimation. Nonsampling error sources will be briefly addressed followed by a more detailed discussion of nonresponse error and nonresponse adjustment techniques. One such nonresponse adjustment technique, propensity score methodology, will be explored more fully to provide a foundation for the research proposed in the subsequent chapters of this thesis.

2.2 NONSAMPLING ERROR

Nonsampling error arises when the sampling design is not perfectly implemented. Only a perfectly designed survey, executed perfectly, could avoid generating nonsampling errors. While little can be done to avoid nonsampling error completely, measures may be taken to minimize the effects of nonsampling error. The bias introduced by nonsampling error may be reduced by careful planning prior to survey execution and by applying appropriate techniques during the analysis phase. The source of the nonsampling error must be identified to assess the impact of that error source and determine the appropriate adjustments.

Nonsampling error is categorized in several different ways. Lessler and Kalsbeek (1992) classify nonsampling error into three types: frame error, measurement error, and nonresponse error. Lohr (1999) defines nonsampling error as any error that cannot be attributed to variation among samples, such as selection bias and response inaccuracy. By Lohr's definition, measurement error and nonresponse error are classified as types of selection bias. Särndal et al. (1992) categorize nonsampling error as “errors due to nonobservation” and “errors in observations.” The former principally include errors due to undercoverage and nonresponse, while the latter consist of data collection and data processing errors. The categorization by Lessler and Kalsbeek (1992) provides the most detailed breakdown of nonsampling error among those reviewed and will be used to guide the current discussion.

Frame error is a type of nonsampling error originating from problems with the sampling frame. The sampling frame is the inventory of all units in the population from which the sample is drawn (Lessler and Kalsbeek, 1992). The target population is the group of population units to which the researcher wants to make inference (Kish 1978). The associations between the sampling frame and the target population are referred to as linkages (Särndal et al., 1992). A one-to-one linkage exists when every unit in the target population is associated with a single unit in the sampling frame, and every unit in the sampling frame is associated with a single unit in the target population. A many-to-one linkage occurs when every unit in the sampling frame is linked to only one unit in the target population, but a unit in the target population may have more than one link to a unit in the sampling frame. A one-to-many linkage exists when a unit in the sampling frame is associated with more than one unit in the target population, but every unit in the target population is linked to a single unit in the sampling frame. Frame error arises when a one-to-one linkage does not occur. For instance, frame error occurs when some target population units are not linked to any units in the sampling frame (undercoverage); when some units in the sampling frame are linked to units that are not members of the target population (overcoverage); when some target population units are linked to multiple frame units (duplicate listings); when the frame contains incorrect information that is used for stratification, probability-proportional-to-size selection, or regression estimation; and when the frame is obsolete or too coarsely-grained to provide linkages to every target population member (Särndal et al., 1992). Sampling frames may be subject to one or more of these problems. These types of frame errors can occur for many types of surveys including aerial surveys of animal populations and surveys of human populations.

Consider a target population of elk within a specific habitat type within a park boundary and a GIS vegetation coverage as a sampling frame. Undercoverage may occur if the sampling frame excludes small unknown pockets of habitat within the park. Frame overcoverage may exist if observers are unsure of park boundaries and count elk outside the park. Errors in the GIS coverage could cause inaccurate stratification that adds to overall error by inflating variance estimates. For a survey that involves completing a questionnaire, duplicate listings of individuals in the target population could be obtained from merging sampling frames obtained from multiple sources (known as *dual-frame sampling*). The target population is therefore perceived to be larger than it truly is, and estimates of the population total may be inflated.

Similarly, an obsolete frame may contain individuals that no longer meet the definition of the target population, and additional error may be generated by making inference to an inflated population.

Measurement error, another nonsampling error source, occurs when an inaccurate response is obtained for a unit (Lessler and Kalsbeek, 1992; Thompson, 1992). Särndal et al. (1992) identify three types of measurement error sources in surveys: the use of an inaccurate instrument, an inaccurate measure provided by a respondent, and an interviewer influencing the response. One type of measurement error is the use of an inaccurate instrument, such as an uncalibrated thermometer, poorly constructed questionnaire, or an inadequately-trained observer who miscounts elk during an aerial survey. A second type of measurement error, often called response error, can be generated when survey respondents give false information, change their behavior due to survey involvement, or cannot recall relevant details. In the elk aerial survey example, response error may occur when elk are classified by the observer into the wrong age group which is a key variable of interest in the study. Elk age groups are useful for assessing elk productivity and predicting future population changes. The third type of measurement error may result when the experience and personal characteristics of the interviewer/observer or survey timing affect the outcome. For example, aerial survey observers could have very different levels of experience and may introduce differing levels of error into data collection. In regard to error associated with the timing of the survey, vegetation surveys could be conducted before or after peak vegetation occurs, causing underestimation of true vegetative cover or abundance. Methods to reduce measurement error have been documented in areas of study such as aerial surveys of animal populations (Buckland, Anderson, Burnham, and Laake, 1993), water quality surveys (Helsel, 2005), and cognitive studies (Salvucci, Walter, Conley, Fnk, Saba, and Kaufman, 1997).

The third type of nonsampling error, called nonresponse error, occurs when incomplete responses are obtained from sampling units (Lessler and Kalsbeek, 1992; Lohr, 1999; Thompson 1992). When the outcomes of nonrespondents are substantially different from outcomes of respondents, the survey results based on only the survey respondents can greatly distort inference to the target population. Nonresponse error is further classified based on the amount of missing information. If all outcomes for the entire sampling unit are not observed, then the data are subject to unit nonresponse. If some information is collected on the sampling unit but at least one outcome of interest is missing, then the unit has experienced item nonresponse (Little and Rubin, 2002). In a survey using questionnaires, for example, unit nonresponse would occur if a person was solicited for survey involvement but declined to participate. However, a person might decide to answer the questionnaire but leave several items blank. These omissions result in item nonresponse. Nonresponse error is specifically addressed in this research and is discussed in more detail below.

2.3 NONRESPONSE ERROR

Many factors can lead to missing data and generate nonresponse error. A sampling frame, for example, may contain remote areas where sites cannot be accessed due to observer constraints. A poorly planned survey design can include, for instance, a telephone solicitation effort scheduled when people are rarely home. If the solicitation is only conducted for a brief period of time, sample members have fewer opportunities to respond and unit nonresponse will result (Dillman, Eltinge, Groves, and Little, 2002). Including features in the survey operation such as a reference to the sponsoring agency's reputation in the introduction, response incentives, and follow-up procedures to contact nonrespondents generally increases the level of cooperation. The survey mode may also influence a person's level of participation. For example, Dillman (2000) determined that face-to-face interviews generate higher response rates than telephone surveys, and telephone surveys successfully obtain more respondents than do mail questionnaires. However, more recent research found that a mail survey with a small monetary incentive obtained a response rate nearly twice that observed from a telephone survey (Dillman, Phelps, Tortora, Swift, Kohrell, Berck, et al., 2009; Lesser, Dillman, Carlson, Lorenz, Mason, and Willits, 2001). Some researchers have observed declines in telephone survey response rates in the previous decade or more due to increases in refusals and noncontacts (Curtin, Presser, and Singer, 2005; Keeter, Kennedy, Dimock, Best, and Craighill, 2006). Regardless of the effort expended to avoid nonresponse, missingness is often a factor in many surveys.

Nonresponse occurs commonly in large-scale surveys conducted by federal and state agencies. In many cases, survey research organizations consider the consequences of nonresponse error so critical as to warrant the application of adjustment techniques. The Office of Management and Budget recommends nonresponse adjustments such as weighting adjustments and imputation in the analysis of surveys exhibiting response rates of 80% or less or item nonresponse at or exceeding 30% (Standards and Guidelines for Statistical Surveys, 2006). The U.S. Census Bureau uses weighting methods (Bailey, 1986; Bailey, 2005) and imputation (Bailey, Jansto, and Smith, 1991) to reduce the impact of nonresponse error on estimates from its Survey of Income and Program Participation (SIPP). The U.S. Census Bureau's economic surveys generally obtain response rates ranging from 75% to 79%. Therefore, imputation and nonresponse weighting adjustments are used to adjust for nonresponse error (Detlefsen, 2007). The Centers for Disease Control (CDC) conduct annual telephone surveys to track behaviors and conditions related to health. An example of one of these surveys is the Behavioral Risk Factor Surveillance System. In 2007, state-level response rates ranged from 15% to 70% (2007 Behavioral Risk Factor Surveillance System Summary Data Quality Report, 2008). Post-stratification adjustments are applied to these data to account for the nonresponse. Post-strata cells are created using demographics such as age, gender, race, and the geographic area of the country. New Mexico Department of Game and Fish annually censuses hunters licensed in New Mexico to hunt elk; however, only 6% to 45% of licensees returned questionnaires for any year between 1988 and 2003 (Harrod, 2007). Subsamples of nonrespondents indicate that estimates based on using only the returned questionnaires overestimate the elk harvest by 28% in both 2001 and 2003. Overestimation of the annual elk harvest results in reductions of future hunt licensing, which may stimulate undesired growth in elk populations in the state of New Mexico. Adjustment techniques incorporating the nonrespondent subsample data will be discussed in later chapters.

When the outcomes of nonrespondents are substantially different from outcomes of respondents, survey results based on a sample may be unrepresentative of the target population and subject to nonresponse error. Therefore, understanding how nonresponse error affects the survey outcome is important. Missing data are often categorized by the pattern of missingness exhibited by the data (Little and Rubin, 2002). Identifying this pattern is often helpful when selecting an appropriate analysis strategy because many methods require assumptions on the characteristics of the missingness.

The pattern of missingness may be defined by the number of outcomes affected by the missingness. Univariate nonresponse occurs when all of the missingness is limited to one outcome of interest. With multivariate missingness, the missingness occurs for more than one outcome. Monotone missingness is a pattern of missingness occurring in clinical trials where repeated measures are taken over time from the same individuals. Missingness is considered monotone when no outcomes are measured for a unit once the initial missingness occurs. Consider a clinical trial where subjects with a particular disease are administered treatments and followed through time. Monotone missingness would occur when the subjects that are cured of a disease drop out of the study and do not return. Missingness that is not strictly monotonic may easily be made monotone by dropping all responses after the first occurrence of missingness.

Missingness that follows no specific pattern is classified as a general missingness pattern. An example of general missingness is a data set obtained from a survey subject to both unit and item nonresponse. The missing data from item nonresponse often differs among respondents and no specific pattern emerges.

Nonresponse error is a ubiquitous sample survey problem generated from many uncontrollable sources. Robust and effective analytical methods to account for nonresponse error are needed in many areas of research using surveys where data are missing. The work included in this thesis will address analytical methods to reduce nonresponse error for unit nonresponse. In order to determine which analytical methods might be appropriate for the analysis of a particular data set, the mechanism causing the missingness must first be examined.

2.4 MISSINGNESS MECHANISMS

The type of data missingness influences the approach taken to account for nonresponse. Missing data are categorized into three types: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) (Dillman et al., 2002; Little and Rubin, 2002; Lohr, 1999). These three categories are defined by the mechanisms which generate the missingness. Let **Y** be the complete data outcome for the sample; let **R** be the vector of missing data indicators for the sample; let **X** be the complete set of covariates associated with **Y** and **R**; and let φ be a set of unknown parameters. Then the conditional distribution of **R**, $f\left(\mathbf{R}\middle|\mathbf{Y},\mathbf{X},\varphi\right)$, may be used to characterize the missing data mechanism (Little and Rubin, 2002). This parameterization provides a framework for the relationship of the outcome of interest to the covariate information and response mechanism.

When data are MCAR, the missingness mechanism does not depend on either the outcome of interest or any covariates (Little and Rubin, 2002). In this case, the response probabilities are mutually independent and the response indicators are conditionally independent of each other given the responding sample size (Lohr, 1999). Therefore, the conditional distribution of the missingness indicator reduces to $f\left(\mathbf{R}\middle|\mathbf{Y},\mathbf{X},\varphi\right)=f\left(\mathbf{R}\middle|\varphi\right)$ for all Y and φ. The missingness pattern itself may not be random, but the mechanism behind the missingness is random. An example of MCAR missingness would be an observer not interviewing a sampling unit due to vehicle malfunctioning or a laboratory worker accidentally dropping a water chemistry sample. If these events are unrelated to the outcome of interest or any related covariates, then the missing data are MCAR. For the analyses, the sample of respondents is treated as representative of the original sample that was selected (Lohr, 1999). Standard analysis techniques are used with the reduced sample and unbiased estimates of population parameters may be obtained.

MAR missingness occurs when the data missingness is related to covariates (**X**) but not the missing outcomes (**Y**) (Little and Rubin, 2002; Lohr, 1999). Little and Rubin (2002) partition the outcome of interest **Y** as $\left(\mathbf{Y}^o, \mathbf{Y}^m\right)$ where $\mathbf{Y}^o$ represents the vector of observed outcomes and $\mathbf{Y}^m$ represents the vector of missing outcomes. Under this parameterization, the conditional distribution of the response indicator under MAR missingness is $f\left(\mathbf{R} \middle| \mathbf{Y}, \mathbf{X}, \varphi\right) = f\left(\mathbf{R} \middle| \mathbf{Y}^o, \mathbf{X}, \varphi\right)$ for all $\mathbf{Y}^m$ and φ. Therefore, the missingness depends on observed outcomes which can be explained by related covariates. For MAR missingness, nonresponse adjustments may be made on the vector of observed outcomes, $\mathbf{Y}^o$, if the covariates related to the missingness (**X**) are available (Lohr, 1999). This type of missingness is also referred to as ignorable missingness because the response mechanism is MCAR once the nonresponse adjustment model is applied, not because the nonresponse bias may be ignored. An example of MAR missingness might include the situation in which an observer cannot survey a site because it is physically inaccessible. If accessibility is unrelated to the response outcome but is related to covariates such as elevation or slope, then these covariates may be used to model the missing outcomes. Let $\mathbf{X} = \left(\mathbf{X}_1, \mathbf{X}_2\right)$, where $\mathbf{X}_1$

represent a subset of available covariates that are associated with the response mechanism, and let $\mathbf{X}_2$ represent the complement of covariates that are unassociated with the response mechanism. A slightly stronger assumption of MAR missingness is the case when $f\left(\mathbf{R}\middle|\mathbf{Y},\mathbf{X},\varphi\right) = f\left(\mathbf{R}\middle|\mathbf{X}_1,\varphi\right)$, for all $\mathbf{X}_2$ and φ. This assumption is related to the assumption of quasi-randomization proposed by Oh and Scheuren (1983) and identifies a subset of covariates for which nonresponse adjustment provides unbiased estimation. Furthermore, unconfounded missingness is defined as missingness that depends only on the design variables, **X**, but does not rely on either the observed outcomes or missing outcomes. This slightly stronger assumption of MAR requires that $f\left(\mathbf{R}\middle|\mathbf{Y},\mathbf{X},\varphi\right) = f\left(\mathbf{R}\middle|\mathbf{X},\varphi\right)$ for all **Y** and φ (Dillman et al., 2002; Little and Rubin, 2002).

Of the three types of nonresponse, NMAR missingness can be the most challenging to correct because the response mechanism is related to the unknown values of $\mathbf{Y}^m$. For this type of missingness, the conditional distribution of the response indicator cannot be simplified from the form $f\left(\mathbf{R}\middle|\mathbf{Y},\mathbf{X},\varphi\right)$ (Lohr, 1999; Little and Rubin, 2002; Dillman et al., 2002). This type of missingness is also called nonignorable nonresponse. In this case, further survey effort must be expended to obtain information from nonrespondents (Bartholomew, 1961; Elliott, Little, and Lewitsky, 2000; Hansen and Hurwitz, 1946; Rao, 1983) or strong modeling assumptions must be made to adjust observed outcomes for nonresponse bias (Rotnitzky and Robins, 1995; Stasny, 1991). An example of NMAR missingness is the case of an observer who has been denied access to a survey site by a private landowner. If the landowner is influencing the outcome of interest at the site by environmental degradation, then the missing outcomes are substantially different from the observed outcomes with no environmental degradation. Therefore, further information is needed to model this difference accurately. Modeling the nonresponse on only the covariates and observed outcomes or ignoring the nonresponse completely will produce biased estimates and statements on confidence will be invalid (Lessler and Kalsbeek 1992; Särndal et al. 1992; Thompson, 1992; Lohr, 1999).

2.5 APPROACHES FOR NONRESPONSE ADJUSTMENT

Choosing a nonresponse adjustment approach requires identifying the type of missingness encountered by the data. If data are MCAR, then no adjustments are necessary. If the data are MAR, then respondent data may be used to adjust for nonresponse bias using covariate information related to the missingness. If the missingness is NMAR, then information on the relationship between the outcome of interest and the response mechanism is needed to adjust for the nonresponse. This additional information will be collected either from a subsample of nonrespondents or from assumptions on the missingness. Approaches to correct for nonresponse bias are subsequently discussed for the cases of ignorable and nonignorable nonresponse.

2.5.1 Ignorable missingness

Analysis methods to account for missing data must be appropriate for the missing data mechanism. For MCAR missingness, standard analysis tools using no adjustment methods may be used with the sample of respondents to obtain unbiased estimates (Lohr, 1999). When the missingness mechanism is MAR, several adjustment methods are available which include weighting adjustments, imputation, and parametric models.

2.5.1.1 Weighting adjustments

Weighting adjustment methods employ weights so that observed responses are extrapolated appropriately to the sampled population. Weighting adjustment procedures include weighting class adjustment, post-stratification adjustment, raking, and response propensity adjustment (Holt and Smith, 1979; Little and Rubin, 2002; Lohr, 1999; Oh and Scheuren, 1983). The goals of these methods are to weight observed outcomes so that discrepancies between the sample and the population may be minimized to obtain estimates that are unbiased under repeated sampling (Gelman and Carlin, 2002). These methods require covariate information to form weighting classes. Within each weighting class, adjusted weights are computed to reduce the distortion of nonresponse so that nearly-unbiased estimates may be obtained.

2.5.1.1.1 Weighting class adjustment

The weighting class adjustment requires information from the sample rather than the entire population to account for nonresponse (Lessler and Kalsbeek, 1992). Covariate information is used to form weighting classes within which nonresponse adjustments are made (Lohr, 1999). The weighting class adjustment is biased, and the bias is a function of the difference between the estimate of interest for respondents differs and the estimate of interest for nonrespondents (Lessler and Kalsbeek, 1992). However, the weighting class adjustment estimator is conditionally unbiased given the intended and obtained sample sizes, and this added level of uncertainty requires an additional variance component to account for the potential bias when the variance is calculated without conditioning on the observed sample sizes (Oh and Scheuren, 1983).

The weighting class adjustment requires the "quasi-randomization" assumption introduced by Oh and Scheuren (1983), which handles the missing data problem like a two-stage sample. Sample inclusion is treated as the first-stage. In the second stage, each unit within a subpopulation is subject to an independent Bernoulli response process with a common positive probability of response within each weighting class. When the weighting class variables are chosen so that the quasi-randomization assumption is met, the missingness mechanism is MCAR within each weighting class (Little and Rubin, 2002). The quasi-randomization approach also requires the assumption that units in different subpopulations have independent response mechanisms. The weighting classes, created to approximate these subpopulations, are formed by levels of variables that are associated with the response mechanism. The weights to adjust estimates for nonresponse are calculated within each weighting class.

Oh and Scheuren (1983) suggest that a smaller, more robust set of classification variables be used so that weighting class sample sizes are large and variances within weighting classes are controlled. Even if the response mechanism can be perfectly modeled by a large set of covariates, the small class sizes may cause variance instability and violate assumptions of normality for means within weighting classes which affects confidence interval coverage. Therefore, the authors recommend that an estimator with small bias and stable variance is preferred over an unbiased estimator with unstable variance and a non-normal distribution.

Response rates are estimated within each adjustment class in one of two ways. Oh and Scheuren's (1983) weighting class adjustment approach assumes that the response mechanism is a simple random sample within each weighting class. Therefore, the response rate within each adjustment class is simply the proportion of respondents for sampling units within that adjustment class. When the sampling design is complex, design weights may be incorporated into the estimation of within-class response rates (Lessler and Kalsbeek 1992; Little and Vartivarian, 2003). These design weights are known at the time of survey execution and may be used to obtain accurate estimates of the response rate given the complex survey design (Gelman and Carlin, 2002). The two methods of computing these adjustment weights are referred to as post-stratification weighting and inverse-probability weighting, respectively. The quasi-randomization assumption of Oh and Scheuren (1983) assumes that the missingness mechanism is a simple random sample within each subpopulation. This assumption may simplify the computational requirements but may not precisely reflect the response rates in the corresponding subpopulations when complex surveys designs are used (Lessler and Kalsbeek, 1992).

2.5.1.1.2 Post-stratification adjustment

Post-stratification (Holt and Smith, 1979) was originally developed to obtain domain estimates for covariates that were not used as design strata. For example, a researcher may want to make inference by gender from a sampling design that did not include gender as a stratification variable. Because sample sizes are not predetermined in unplanned post-strata, the unconditional variance of post-strata estimators is increased by a variance component that accounts for random sample sizes. The benefits of post-stratification include conditional unbiasedness and gains in efficiency. Like weighting class adjustments, post-stratification adjustments require the quasi-randomization assumption to define the response mechanism. Furthermore, response rates may be estimated with post-stratification weighting or inverse-probability weighting within each adjustment class. Post-stratification adjustments differ from weighting class adjustments by the amount of covariate information needed. When covariate data used to create weighting classes is known for the entire population, post-stratification adjustments may be used.

2.5.1.1.3 Raking adjustment

Raking (also called ratio raking adjustment, iterative proportional fitting, or multiplicative weighting) is another weighting adjustment to account for MAR missingness (Bethlehem, 2002). Raking is an iterative post-stratification technique used when more than one post-stratum is needed but only marginal totals of each post-stratum are known (Lohr, 1999). The sums of the sampling weights are calculated within each weighting class and then the weights are adjusted by the ratio of the true marginal total divided by the estimated marginal total calculated from the weights. This process is repeated for each weighting variable until the estimated marginal totals calculated from the adjusted weights converge to the true marginal totals (Little and Rubin, 2002). The sum of the weights within each weighting class is an asymptotically unbiased estimate of the true parameter within each weighting class (Oh and Scheuren, 1983). The raking procedure requires the quasi-randomization assumption (response mechanisms are independent Bernoulli processes with a common response probability within post-strata and are independent among post-strata) as well as the additional assumption of no interaction among weighting class variables with the response mechanism. Benefits of raking adjustment include the ability to incorporate outside data sources when only marginal totals are available. In addition, using the marginal totals rather than weighting classes with small sample sizes common with post-strata reduces the problem of variance inflation from small cell sizes. Disadvantages of the raking adjustment include lack of convergence to the true marginal total if some of the cell totals are zero. In addition, the variance may be inflated if the weighting classes are weakly correlated with the within-class means (Lohr, 1999). Oh and Scheuren (1983) also warn that raking adjustments are computationally more complex and may result in unstable estimates that do not converge to marginal totals.

2.5.1.1.4 Response propensity adjustment

Other methods of estimating the probability of response are used in adjustments for MAR nonresponse. Politz and Simmons (1949) propose a method of estimating response probabilities based on information related to the number of times a nonrespondent was available during a survey period. This approach assumes that multiple contacts were planned in the survey protocol as a mechanism to reduce nonresponse. Assuming that the most accessible respondents are overrepresented in a sample, respondents who are less accessible are given more weight to represent the inaccessible nonrespondents (Lohr, 1999).

The probability of response conditional on sample inclusion, also called the propensity score, may be directly estimated to account for nonresponse. Propensity score methodology, first proposed by Rosenbaum and Rubin (1983), was originally used to match treatment and control units in a nonrandomized treatment so that the treatment effect could be estimated within levels of potentially-confounding covariates. Logistic or probit regression is used to directly estimate the probability of response from related covariates (Steinhorst and Samuel, 1989; Cassel et al., 1983). A benefit of this approach is that response probabilities may be modeled for each combination of covariates. However, the inverse of each response probability is used to weight for nonresponse. If very small propensity scores are estimated for response, then the corresponding weights will increase variance estimators. Therefore, variance estimates may be unstable when very small propensity scores are obtained. The resulting variance inflation may cause an overall increase in the mean squared error (MSE) despite the reduction in bias. This method also requires accurate logistic regression model specification (Little and Rubin, 2002). Model misspecification may not reduce the bias or may even increase the bias of estimates adjusted with the erroneous propensity scores.

Propensity score classification is an extension of propensity score methodology and was first used to compare treatment and control groups within subclasses of related covariates in observational studies (Rosenbaum and Rubin, 1983). In this application, the propensity score measures the probability that, conditional on observed covariates, a unit belongs to the treatment group. A sample is "balanced" when treatment and control units are grouped so that the effects of confounding covariates are diminished and the treatment effect may be estimated. Samples may be balanced retrospectively by creating groups based on covariates that influence the treatment effect. When a large number of covariates are available, the number of distinct covariate combinations may be reduced by forming groups defined by quantile of the propensity score. Within subclasses based on the propensity score, the distribution of the covariates conditional on the propensity score is the same for treatment and control units, or is balanced. Furthermore, the covariates and the treatment indicator are independent, conditional on the propensity score.

Propensity score methodology was first applied in a missing data context by David, Little, Samuhel, and Triest (1983). The goal of their work is to avoid stratifying on the complete set of covariates in order to simplify estimation. By categorizing the propensity score into groups with similar response propensities, the response mechanism within each class is independent of the outcome of interest. Therefore, the propensity score stratification creates subpopulations within which the distribution of the covariates is the same for respondents and nonrespondents. The benefits of this approach are that any number of covariates related to the response mechanism can be summarized by one explanatory variable, the propensity score. The propensity score is relatively easy to calculate, requiring logistic or probit regression, quantile estimation, and the Horvitz-Thompson (1952) estimator. Subsequent chapters of this thesis will focus on propensity score stratification and its application in missing data settings, and a more thorough discussion of the relevant literature will be discussed in a later section of this chapter.

All of the weighting adjustments discussed require the formation of weighting classes within which adjustment weights are calculated. Weighting classes should be constructed so that each class contains a "reasonable" number of respondents. Lessler and Kalsbeek (1992) propose at least 20 respondents per class and Lohr (1999) adds the additional requirement of a response rate of 50% or more within each weighting class. Weighting classes should be formed for post-stratification adjustments so that measures of the population size within each post-stratum are accurate. Choosing weighting classes that are associated with the response mechanism reduces bias.

The simplicity of using a single set of weights to account for nonresponse comes at the cost of variance inflation in the weighted estimates (Little and Rubin, 2002). Variances from weighting class adjustments will be inflated due to the variation among the adjustments from different weighting classes (Lessler and Kalsbeek, 1992). Assuming simple random sampling within weighting classes and constant variance, the inflation factor due to weighting is one plus the squared coefficient of variation of the weights for all responding units. While nonresponse weighting decreases bias, the introduction of variable weights increases the variance. Imputation is a technique used for nonresponse adjustment that does not inflate the variance of the estimator.

2.5.1.2 Imputation

Imputation is a MAR nonresponse adjustment technique that works by filling in missing values to create complete data sets (Little and Rubin, 2002). Imputation may be used for unit nonresponse or item nonresponse, and imputed data may be generated once (single imputation) or many times (multiple imputation). Advantages of using imputation include the ability to use complete-case analysis methods and its use with complex surveys. Imputation is used widely for nonresponse adjustment by agencies such as the U.S. Census Bureau (Bailey et al., 1991) and the U.S. Department of Justice (Kennickell, 1997). Imputation has the advantage of controlling variance and bias. However, estimates from imputation may be susceptible to bias depending on the imputation procedure used. In addition, accounting for imputation variance should be incorporated into the mean square error.

The two general approaches for imputation are explicit modeling and implicit modeling (Little and Rubin, 2002). In explicit modeling, a formal predictive distribution is assumed. Mean imputation, regression imputation, and stochastic regression imputation are examples of explicit modeling imputation approaches. Implicit modeling employs an algorithm to obtain imputed values and is used in approaches such as hot deck imputation, cold deck imputation, and substitution. Furthermore, several imputation methods may be combined in one analysis.

Mean imputation, in which means of cells or subclasses are substituted for missing values, is an explicit modeling method similar to weighting class adjustments in that the subclass means are substituted for missing values. Regression imputation uses an explicit regression model to predict missing values, and stochastic regression imputation incorporates an added residual to account for the uncertainty in the prediction. Mean imputation may distort the distribution of outcomes within a weighting cell because all of the missing data are assigned the same imputed value (Little, 1986). Imputation using draws from a predictive distribution is another form of explicit modeling. This approach is superior to mean imputation because the imputed data sets reflect the variance of the sample. This extension of the mean imputation methodology motivates the use of Bayesian models in imputation adjustments (Little and Rubin, 2002; Zhang, 2003).

In contrast to the predictive model used in explicit modeling, implicit modeling employs an algorithm to obtain imputed values. Several methods of implicit imputation have been developed. Hot deck imputation substitutes observed values from responding units with similar characteristics in place of the missing values. Cold deck imputation uses information from external sources to obtain complete data sets. Substitution is used to replace missing observational units by substituting a nonresponding unit with a unit not previously included in the sample. Little and Rubin (2002) warn that complete data sets from substitution are still affected by missingness after the imputations are made because the substitutions are not the actual outcomes. Therefore, the additional error introduced by the imputation should be incorporated in estimates of precision.

Each of these implicit and explicit modeling imputation methods produces a data set without missing outcomes. Comparisons of implicit and explicit models for the U.S. Census Bureau's Current Population Survey (CPS) revealed that the use of hot deck imputation corrected more error than stochastic regression imputation methods and was only slightly less efficient than regression imputation methods (David, Little, Samuhel, and Triest, 1986). However, matching nonrespondents and respondents with similar characteristics is difficult when extensive covariate information is available, because no respondents are available for specific combinations of variables. CPS researchers concluded that the modeling approaches would be superior to hot deck imputation because these methods can incorporate any number of variables and only the significant interactions need to be included in the model.

Single imputation substitutes only one value for each missing unit; the drawback of this approach is its inability to account for all of the variability in the outcome or the additional variability due to the correction with imputation (Little and Rubin, 2002; Wang, Sedransk, and Jinn, 1992). Multiple imputation involves creating several complete data sets to provide an estimate of imputation error. Multiple imputations may be derived from draws from the posterior predictive distribution of the missing outcomes or within adjustment cells defined by propensity scores when the missingness pattern is monotone (Zhang, 2003). The variance of an imputation estimator is a function of the variance of the outcome within each imputed data set and the variance of the estimator among the imputation data sets. Other methods of calculating variances for imputation estimates include explicit variance estimators such as that used for the weighting class adjustment, imputation modification that preserves the complex survey design, and resampling methods such as the bootstrap and jackknife. While multiple imputation is somewhat model-specific, the procedure is easy to implement and valid even for small samples.

Weighting adjustments and imputation techniques may incorporate the survey design while making weak model assumptions to account for the nonresponse. These model-assisted techniques incorporate modeling approaches with design-based estimation and allow adjustment for nonresponse with consideration for survey complexity. Model-based approaches require more stringent assumptions on the nature of the nonresponse and do not directly incorporate the inclusion weights. However, model-based analyses for nonresponse are especially useful when data are NMAR.

2.5.1.3 Model-based approaches

Weighting adjustments and some types of imputation (e.g. mean imputation or imputation based on propensity score classification) represent randomization approaches in which the outcomes are treated as fixed and the mechanisms of sample inclusion and response are considered random (Little, 1982; Rubin, 1983). Modeling approaches to nonresponse treat the outcomes as random with distributions specified by some model. Model-assisted and model-based approaches differ because design weights are ignored in model-based approaches.

Approaches to modeling nonresponse vary depending on the missing data patterns. When data exhibit monotone missingness, the likelihood may be factored for complete and incomplete data (Little, 1982). If the parameters corresponding to the complete and incomplete data are distinct, then maximum likelihood methods may be used to model them separately. For general missingness patterns, the EM algorithm may be used with regression imputation used in the expectation step and an adjustment to the covariance in the maximization step to account for additional variation from the imputation (Dempster, Laird, and Rubin, 1977). Drawbacks of likelihood approaches to modeling nonresponse include the possibility of non-unique maximum likelihood solutions and the effects on bias due to assumptions on the distribution of the outcome that are not met (Lessler and Kalsbeek, 1992). However, likelihood methods are desirable because the analysis is straightforward after the set of assumptions on the nature of the nonresponse are made. These methods are especially useful when data are NMAR and the outcomes of nonrespondents cannot be accounted for by covariate information alone.

2.5.2 Nonignorable missingness

When the missingness mechanism is NMAR, then either further effort is needed to obtain information from nonresponding units or a model must be used. Hansen and Hurwitz (1946) first proposed subsampling nonrespondents to account for missing data. Extensions of this work incorporate the survey mode and number of survey attempts to obtain responses from sample members who originally did not respond to the survey (Bartholomew 1961; Rao 1983; Elliott et al. 2000). Singh and Sedransk (1983) outline a Bayesian approach to estimate the mean outcome and provide information on optimal choices of sample sizes for initial samples and subsamples.

Often, nonrespondents cannot be subsampled. If a site is inaccessible to an observer crew or a landowner denies access to a site on private property, then no amount of additional effort will make it possible to include those sites in the sample. In this case, a modeling approach must be used to account for the nonresponse bias. Nonignorable response models offer the advantage of formally incorporating subjective information about the response mechanism, but a drawback of these methods is the added complexity of inference and increased model dependence (Lessler and Kalsbeek, 1992).

A variety of models is available to account for nonignorable nonresponse. The EM algorithm may be used to obtain maximum likelihood estimates regardless of the type of missing data pattern encountered (Ibrahim, Chen, and Lipsitz, 2001; Little, 1982). Copas and Farewell (1998) reduce the effects of nonignorable item nonresponse by incorporating an "enthusiasm to respond" variable that accounts for nonrespondents' reluctance to discuss personal information. This variable is modeled as a propensity score from covariates that might affect response, such as the age of the respondent, the perceived level of embarrassment from sensitive questions, and the respondent's ability to understand the questions. Scharfstein, Rotnitzky, and Robins (1999) use semiparametric response models and assume that the selection bias parameter is known. Similarly, Qin, Leung, and Shao (2002) use a semiparametric maximum likelihood approach but the response indicator is modeled parametrically in this application. Efficiency is also increased by incorporating auxiliary information. Birmingham and Fitzmaurice (2002) apply the multivariate logistic model with pattern-mixture models to account for nonignorable nonresponse. Tang, Little, and Raghunathan (2003) use the marginal distribution of the covariates in a multivariate regression analysis in which the missingness mechanism does not need to be specified. Regardless of the parameterization of the missingness, nonignorable nonresponse models are highly sensitive to error from model misspecification. A range of model assumptions must be explored to assess model sensitivity (Birmingham and

Fitzmaurice, 2002; Little, 1982; Scharfstein et al., 1999). Assumptions of missingness may also be evaluated with tests of the available data.

2.5.3 Tests for Missingness Types

The data available from the sampling frame and respondents are used to evaluate the type of missingness and identify covariates for analysis. If covariate data are available for both respondents and nonrespondents, then a logistic regression analysis of the response indicator is used to determine if nonresponse adjustment procedures should be used. If any covariates are significant explanatory variables in the model of response probabilities, then data are not likely to be MCAR (Lohr, 1999).

Chi-square goodness-of-fit tests may be used to test the null hypothesis that the levels of an auxiliary variable are not associated with response rates. If response differs significantly by the levels of the auxiliary variable, then that auxiliary variable might be an appropriate weighting class variable for a MAR adjustment technique. Regressing the respondent outcomes on the estimated propensity score provides information to test the hypothesis of MAR data (Little, 1986). For example, assume that the response propensity can be effectively modeled from a suite of available covariates. Then the estimated propensity score can be used as a predictor in a model of the outcome of interest. If the coefficient for the propensity score is significantly different from zero, then the outcome of interest is related to the response mechanism and nonresponse adjustment is recommended. The inability to detect variables for which response rates vary does not imply that no covariates are associated with the response mechanism. In this case, even a simple nonresponse adjustment across all nonrespondents may reduce nonresponse bias. These tests are most useful when available covariate information is highly correlated with both the missingness mechanism and the outcome of interest.

2.5.4 Summary

The vast majority of literature about nonresponse adjustment has focused on the case of MAR missingness. However, many surveys encounter NMAR missingness (Birmingham and Fitzmaurice, 2002; Ibrahim et al., 2001; Kott, 2005) and inference may be misleading if adjustment methods are not incorporated. A drawback of many nonignorable nonresponse models is that they are very complex and difficult to apply unless one is experienced in statistical theory and software use. Propensity score methodology offers a simpler approach that requires the use of logistic or probit regression, the calculation of quantiles, and the use of Horvitz-Thompson (1952) estimators. Propensity score methodology is reviewed more thoroughly in the subsequent section, and this background will be used to construct the theory for propensity score methodology for NMAR missingness when a subsample of nonrespondents is available.

2.6 PROPENSITY SCORE METHODOLOGY

2.6.1 Background

Propensity score methodology was first developed by Rosenbaum and Rubin (1983) to estimate a treatment effect in an observational study setting. In this setting, the propensity score is defined as the probability that a unit is assigned to a specific treatment group. In a randomized clinical trial, the propensity score is known and is part of the experimental design. In a nonrandomized survey, assignment to the treatment group is not randomized. The propensity score must be estimated from data to understand the mechanism behind treatment assignment. The treatment effect may be confounded with many variables associated with treatment assignment. By matching treatment subjects with control subjects that exhibit similar covariate information, the effect of the treatment is more accurately estimated. The covariate information used for matching may be summarized by the propensity score, simplifying the matching process and ensuring that sample sizes within each matching class are adequate.

Several terms are defined to help in the discussion of propensity score methodology. When treatment and control units are grouped so that the effects of confounding covariates on the treatment effect are diminished, the sample is considered "balanced". A *balancing score* is defined as a function of the covariates such that conditioning on the balancing score yields conditional independence between the covariates and the treatment assignment. Therefore, the information contained in the balancing score is adequate in grouping treatment and control units so that treatment effects are accurately estimated. *Strongly ignorable missingness* is missingness in which, given a set of covariates, the outcome of interest is independent of the treatment assignment. This assumption is analogous to the MAR assumption in the nonresponse setting.

In the observational clinical study setting, Rosenbaum and Rubin (1983) use large-sample theory to propose three theorems which provide the basis for propensity score methodology. The first theorem is that, conditional on the propensity score, treatment assignment and observed covariates are independent. In this case, the information contained in the observed covariates is summarized by the propensity score. The second theorem states that a function of the covariates is a balancing score if and only if the propensity score can be expressed as a function of the balancing score. In the third theorem, if the treatment assignment is strongly ignorable given a set of covariates, then treatment assignment is strongly ignorable given any balancing score that is a function of that set of covariates. These three theorems form the arguments for the techniques of pair matching and subclassification. Pair matching on balancing scores occurs when the treatment effect is estimated by conditioning on the propensity score. Subclassification on balancing scores requires forming subclasses with similar treatment propensity scores so that unbiased estimates of treatment effects may be obtained within each group.

Applications of propensity score methodology are used extensively for matched case-control studies (Dehejia and Wahba, 2002; Foster, 2003; Imbens, 2000). In clinical research, propensity score matching is useful to account for the effects of many covariates while utilizing a simple adjustment tool. Rubin and Thomas (2000) extend the theory so that variables highly correlated with the outcome may be used in addition to the propensity score for matching. Lu (2005) uses propensity score matching to balance time-related covariates in longitudinal studies. Variable numbers of matched controls may be obtained through propensity score matching to increase the proportion of bias removed (Ming and Rosenbaum, 2000). Clinical studies often involve a large number of covariates; selecting a subset for nonresponse adjustment can be difficult. Propensity score methodology offers an approach that allows the incorporation of many covariates and simple model selection tools may be used to obtain a highly-correlated subset of covariates for matching. The application of propensity score methodology to missing data problems does not focus on matching but rather on the grouping of respondents with similar scores.

2.6.2 Propensity score methodology in nonresponse adjustment

In nonresponse adjustment, inference on treatment assignment is replaced by inference on the response indicator, with the propensity score measuring the propensity to respond for a given unit. The inverse of the estimated response propensity is used to obtain a modified Horvitz-Thompson (1952) estimator of the parameter of interest. This estimator may inflate the variance of estimates when the estimated response propensity is very small (Little, 1986). Stratification on the response propensity score can approximately balance the outcomes between respondents and nonrespondents if the stratification on the propensity score is "fine" enough (David et al., 1983). The stratification must be fine enough so that outcomes for respondents and nonrespondents are approximately equal within each stratum but coarse enough so that sample sizes are adequate within each stratum. The classification of propensity scores into five or six quantiles provides the basis for defining weighting adjustment classes called response propensity strata within which missingness is approximately MCAR (Little and Rubin, 2002). Weights within each stratum are usually estimated nonparametrically as the inverse of the observed response rate within each adjustment cell. This approach to weighting is similar to Oh and Scheuren's (1983) post-stratification approach assuming quasi-randomization if the weighting classes are structured appropriately on the propensity score.

A benefit of propensity score stratification is observed when the full set of covariates is prohibitively large for adjustment. The propensity scores computed from the full set of covariates are highly model-dependent and may lack respondents within every level of the suite of variables. If a large number of covariates are available, a reduced set of covariates must be identified for propensity score estimation. For logistic regression, this reduced set of covariates may be obtained from model selection using drop-in-deviance tests or information criteria.

A similar approach to forming adjustment classes is predictive mean stratification (Little, 1986). In this approach, the outcome of interest is modeled as a function of related covariates. The predictions are grouped into intervals which define adjustment classes. Values of the outcome are approximately constant within each adjustment class, so the independence of the outcome and the response mechanism holds approximately within each adjustment class when data are MAR.

Nonresponse adjustment cells may be formed on quantiles of the propensity score or the predicted mean, and each approach has advantages and disadvantages. Nonresponse adjustment within cells formed by the propensity score is effective for reducing bias when missingness is MAR (David et al., 1983). However, weighting methods inflate variances, especially when the propensity score is associated with covariates that are unrelated to the outcome of interest (Little, 1986). Adjustment cells based on the predicted mean are effective in mean imputation approaches to control variance but may be biased. Weighting by the inverse of the response rate within each cell is equivalent to mean imputation within adjustment cells for estimates of totals, means, and domain means. Classification on both classes formed by the response propensity and classes formed by the predicted mean, called *joint classification,* has the benefit of both gains in efficiency and bias reduction (Vartivarian and Little, 2003).

Instrumental variable regression is an estimation approach which models the mean of the outcome of interest as a function of the response propensity score (David et al., 1983). However, the response propensity is not modeled as a function of the outcome because only respondent data are used and nonrespondent outcomes are not available. Assuming missing data are MAR, the regression model includes effects for the propensity score and auxiliary data associated with the outcome mean. To avoid collinearity problems, different covariates should be used in the predicted mean model than are used in the response propensity model. The instrumental variable regression model includes the response propensity in the case that the mean is poorly predicted by the model covariates.

When a subsample of nonrespondents is available, joint classification and instrumental variable regression may be adapted for nonignorable nonresponse adjustments. Joint classification involves modeling the predicted mean and the response propensity independently which is not practical with nonignorable nonresponse because the outcome of interest and the response mechanism are correlated. Instrumental variable regression can be extended so that, in addition to modeling the predicted mean from response propensities, the response propensity score may be modeled from the predicted mean or some function of the predicted mean. Then the relationship between the outcome and the response mechanism inherent in nonignorable nonresponse may be reflected in the estimated response propensities.

2.7 RESEARCH DIRECTIONS

Joint classification by both response propensity scores and predicted means, along with instrumental variable regression, motivate the use of propensity score methodology in the case of NMAR missingness. The correlation between the response indicator and the outcome of interest may be used to model predicted means as a function of response. Response propensities may be modeled as a function of the outcome when nonrespondent information is available from a subsample. A benefit of this approach is the ability to develop a predictive mean model that may be applied in repeated surveys of the same population. This predictive model will be applied to binary outcomes to obtain a success propensity score that is used to model response when the outcome of interest is unavailable for nonrespondents. Propensity score methodology and predictive mean theory are used to form weighting classes for two nonresponse adjustment approaches.

In Chapter 3, propensity score methodology is extended to the case of nonignorable nonresponse. This new methodology is applied to the weighting class adjustment (Chapter 4) and the Cassel et al. (1983) estimator (Chapter 5) using pilot data from a survey of New Mexico elk licensees. In Chapter 6, the estimators are compared for a wider range of scenarios with simulations.

CHAPTER 3: PROPENSITY SCORE METHODOLOGY FOR NONIGNORABLE MISSINGNESS

Propensity score methodology has been used primarily in the MAR missingness setting. In this chapter, propensity score theory is extended to the case of nonignorable nonresponse for a new application of propensity score methodology. Notation is first established, and then the motivation and assumptions for the methodology are discussed. Properties of the NMAR propensity score methodology are proposed for nonignorable nonresponse (Chapter 3) and then examined with pilot data for two estimators (Chapters 4 and 5) and simulations (Chapter 6).

3.1 NOTATION

Consider a finite population $\mathcal{U} = \{u_1, \dots, u_N\}$. For each unit u_i, there exists a real-valued outcome Y_i and a vector $\mathbf{X}_i = \left(X_{1i}, \dots, X_{pi}\right)'$ of covariates. Let $\mathbf{Y} = \left(Y_1, \dots, Y_N\right)$ and $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_N\}$. Denote a nonempty set s such that $s \subseteq \mathcal{U}$ refers to an unordered sample. Let n, the number of elements of S, represent the intended sample size before nonresponse occurs, and assume that n is fixed. Let the set of all sets s be denoted as $\mathscr{T}$. Define a function $p(s)$ on $\mathscr{T}$ such that $p(s) > 0$, for all $s \in \mathscr{T}$, and $\sum_{\mathscr{T}} p(s) = 1$, where the summation occurs across all possible sets s in $\mathscr{T}$. Then $p(s)$ is the function that specifies the sampling design (Cassel, Särndal, and Wretman, 1977).

Let **D** = $(D_1, D_2, \ldots, D_N)$ describe the implementation of the sampling design, where D_i is the indicator that unit i is included in the sample, *i.e.*

$$D_i = \begin{cases} 1, & \text{unit } i \text{ included in the sample} \\ 0, & \text{otherwise} \end{cases}.$$

Let **D** consist of a vector of independent and identically distributed binary random variables where $D_i \sim Bernoulli(\pi_i)$ and $\pi_i = P(D_i = 1)$ is the probability of sample inclusion for unit i. Furthermore, let $\pi_{ij} = P(D_i = 1, D_j = 1)$ be the joint inclusion probability for units i and j. Note that the inclusion probabilities and joint inclusion probabilities may depend on the individual covariate values as well as the full set of covariate information, *i.e.* $\pi_i = \pi^{(1)}(\mathbf{X}_i, \mathbf{X})$ and $\pi_{ij} = \pi^{(2)}(\mathbf{X}_i, \mathbf{X}_j, \mathbf{X})$, where $\pi^{(1)}$ and $\pi^{(2)}$ are known functions. Let $\mathcal{I}_D = \{i : D_i = 1\}$.

We assume that the covariates $\mathbf{X}_i$ are observed for every unit of the sample. However, when nonresponse occurs, the outcome of interest, Y_i, is not observed for every unit in the sample. For $i \in \mathcal{I}_D$, let R_i be the indicator that unit i responds, *i.e.*

$$R_i = \begin{cases} 1, & \text{unit } i \text{ responded} \\ 0, & \text{otherwise} \end{cases}.$$

Let $\mathbf{R}$ denote a Bernoulli random vector of missing data indicators for the sample with indices in increasing order and define $\mathcal{I}_R = \{i \in \mathcal{I}_D : R_i = 1\}$. Let $p_{ri} = P(R_i = 1 | D_i = 1)$ be the probability of response for unit i given sample inclusion. When data are MAR, we can model the response propensity as $p_{ri} = p(\mathbf{X}_i; \beta)$ where p is a known function and β is a vector of unknown parameters. Under nonignorable nonresponse, we must assume that $p_{ri} = p(Y_i, \mathbf{X}_i; \beta)$ because the response indicator and the outcome of interest are not independent. Let $\mathbf{Y}^o$ denote the vector of observed Y_i, $i \in \mathcal{I}_R$, such that such that $\mathbf{Y}^o = \{Y_i : D_i = 1, R_i = 1\}$ for $i = 1,\ldots n$ with indices in increasing order. Let $m = \sum_{i=1}^{n} R_i$ be the number of responding units, where $m < n$ when nonresponse occurs.

The conditional distribution of R_i, $f\left(R_i \middle| Y_i, \mathbf{X}_i\right)$, is used to describe missing data mechanisms and their inherent assumptions (Little and Rubin, 2002). When nonresponse is nonignorable, the conditional distribution of R_i cannot be simplified. Nonresponse adjustment techniques must account for the correlation between the response indicator and the outcome of interest.

3.2 MISSINGNESS

Survey nonresponse is a potentially serious source of error in estimation from survey data (Lessler and Kalsbeek, 1999). Nonresponse occurs when a unit selected for the sample does not provide a complete response. In surveys using questionnaires, nonresponse may occur when a person selected for the survey is not home when called or refuses to participate in the survey. Nonresponse is found in ecological surveys when a sample site is inaccessible or located on private property where a landowner denies access. The inability to obtain complete responses for all sampled units can produce biased estimates of parameters and variance, leading to erroneous inference (Lohr, 1999).

Two types of nonresponse behavior may be observed: unit nonresponse and item nonresponse (Little and Rubin, 2002). Unit nonresponse occurs when the entire set of variables for a unit is missing. When a partial response is obtained, then the sample is said to be subject to item nonresponse. Nonresponse bias impacts estimation when the nonresponse rate is substantial and missing outcomes differ substantially from the observed outcomes (Lessler and Kalsbeek, 1999). Nonresponse bias may cause invalid inference due to variance inflation and biases in point and precision estimates (Dillman et al., 2002). Methods to reduce bias from nonresponse error have been developed to correct this problem (David et al., 1983; Holt and Smith, 1979; Little, 1986; Oh and Scheuren, 1983). In selecting the appropriate method with which to reduce nonresponse bias, it is helpful to understand the mechanism that generates the data missingness. We will restrict further discussion to the case of unit nonresponse.

The conditional distribution of $\mathbf{R}$, $f\left(\mathbf{R} \mid \mathbf{Y}, \mathbf{X}, \varphi\right)$, is used to describe missing data mechanisms and their inherent assumptions, where φ represents all unknown parameters related to response (Little and Rubin, 2002). Missingness is classified into three categories (Dillman et al., 2002; Little and Rubin, 2002; Lohr, 1999): missing-completely-at-random (MCAR), missing-at-random (MAR), and not-missing-at-random (NMAR) missingness. These three categories are distinguished by the conditional distribution of the response indicator, **R**. When MCAR missingness occurs, the mechanism generating the missing data is not related to the outcome of interest, any covariates, or any unknown parameters. In this case, the conditional distribution of **R** reduces to $f\left(\mathbf{R}\middle|\mathbf{Y},\mathbf{X},\varphi\right)=f\left(\mathbf{R}\middle|\varphi\right)$ for all **Y** and φ. The realized sample is considered a random subsample of the full sample and statistical analysis without any corrections for nonresponse may be conducted.

When data are MAR, then the missingness mechanism is related to the outcome of interest through covariates. If these covariates can be identified, then the distribution of the response indicator conditional on the observed outcomes and related covariates is independent of the missing outcomes and reduces to $f\left(\mathbf{R}\middle|\mathbf{Y},\mathbf{X},\varphi\right)=f\left(\mathbf{R}\middle|\mathbf{X},\varphi\right)$ for all **Y** and φ. Many methods are available to correct nonresponse bias from MAR missingness. Adjustment methods include weighting methods (Holt and Smith, 1979; Oh and Scheuren, 1983), which modify design weights to account for missing outcomes, and imputation (Little and Rubin, 2002; Wang et al., 1992), which uses various methods of substitution to obtain complete-case data sets not subject to missingness.

When data are NMAR, the missingness mechanism is related to the missing outcomes. In this case, conditioning on the observed data does not remove the dependence on missing outcomes. The conditional distribution of the response indicator cannot be simplified from its fullest form, $f\left(\mathbf{R}\middle|\mathbf{Y},\mathbf{X},\varphi\right)$. Also called nonignorable nonresponse, this type of missingness requires further survey effort to obtain information from nonrespondents (Bartholomew, 1961; Elliott et al., 2000; Hansen and Hurwitz, 1946; Rao, 1983) or stronger modeling assumptions (Rotnitzky and Robins, 1995; Stasny, 1991). Correct specification of the missingness mechanism is necessary to obtain unbiased inference.

3.3 INFERENCE

Inference will be made on the population total, $\tau = \sum_{i=1}^{N} y_i$ for values y_i of variables Y_i, $i = 1, \ldots, N$. Using design weights π_i for units in the sample, the Horvitz-Thompson (1952), or HT, estimator of the population total, τ, is:

$$\hat{T} = \sum_{i=1}^{N} \frac{D_i y_i}{\pi_i} = \sum_{i=1}^{n} w_i y_i \ , \tag{3.1}$$

where $w_i = \pi_i^{-1}$. In the case of nonresponse, the following modified Horvitz-Thompson estimator (Little and Rubin, 2002) is used:

$$\tilde{T} = \sum_{i=1}^{N} \frac{D_i R_i y_i}{\pi_i p_i} = \sum_{i=1}^{m} w_i \theta_i y_i \ , \tag{3.2}$$

where $\theta_i = p_i^{-1}$ is the response weight to account for nonresponding units. Several methods are used to estimate the response probability. Weighting class adjustments and post-stratification employ adjustment cells within which the response probability is estimated using post-stratification weighting or inverse-probability weighting (Gelman and Carlin, 2002). Logistic regression may also be used to estimate the response propensity for each unit (Cassel et al., 1983).

Note that $D_i R_i$ in equation (3.2) is an indicator for the event that unit *i* is both included in the sample and is observed, where

$$P\left(D_i R_i = 1\right) = P\left(D_i = 1, R_i = 1\right) = P\left(D_i = 1\right) P\left(R_i = 1 \middle| D_i = 1\right) = \pi_i p_i \ .$$

Therefore, $\tilde{T}$ may be regarded as a Horvitz-Thompson estimator and is unbiased for the true population total, *i.e.* $E\left(\tilde{T}\right) = T$. To calculate $\tilde{T}$, π_i and p_i are needed. The inclusion probabilities, $\pi_i = \pi^{(1)}\left(\mathbf{X}_i, \mathbf{X}\right)$, may be calculated given a specified sampling design. Three difficulties exist with the calculation of $p_i = p\left(y_i, \mathbf{X}_i; \beta_r\right)$. First, p_i depends on the unknown parameter, β_r. Therefore, p_i cannot be calculated, only estimated. Second, when nonresponse is nonignorable, the probability of response depends on the outcome of interest, *i.e.* $p\left(y_i, \mathbf{X}_i; \beta_r\right) = P\left(R_i = 1 \middle| y_i, \mathbf{X}_i; \beta_r\right)$. The response probability p_{ri} cannot be accurately estimated from the observed data because the observed outcomes are insufficient to provide information on missing outcomes when nonresponse is nonignorable. Third, suppose we obtain an estimate $\hat{\beta}_r$ for β_r. Then we can estimate p_i by $\hat{p}_i = p\left(y_i, \mathbf{X}_i; \hat{\beta}_r\right)$. However, when y_i is missing, p_i cannot be calculated.

The second difficulty may be dealt with by obtaining a subsample of nonrespondents. Let $\mathscr{I}_b = \left\{i \in \mathscr{I}_D : R_i = 0\right\}$. Let $s^{(b)}$ denote a set of size n_b that is randomly selected from all units indexed by i such that $i \in \mathscr{I}_b$, where $n_b \leq (n - m)$. Let $\mathbf{Y}^{(b)}$ represent the outcomes for the set $s^{(b)}$, and $\mathbf{X}^{(b)}$ represents the covariate matrix for the subsampled nonrespondents. The nonrespondent subsample is assumed to come from a probabilistic design obtained by randomization. The outcomes in the nonrespondent subsample may be used to estimate p_{ri} to alleviate the second difficulty described above. A logistic regression model of the response indicator based on the combined data from the original survey and the nonrespondent subsample may provide estimates of the response probability for each unit based on related covariates. Methods to obtain unbiased estimates of the response probabilities for NMAR missingness are discussed in this chapter and in Chapters 4 and 5. The third difficulty is addressed by assuming a superpopulation model for the unknown parameter, $\boldsymbol{\beta}_r$. The model for $\boldsymbol{\beta}_r$ may be constructed from information in the subsample and applied to data from similar surveys with nonresponse but without a nonrespondent subsample. The estimate of $\boldsymbol{\beta}_r$ will depend on both the initial respondents and a subsample of nonrespondents, *i.e.* $\hat{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}}_r\left(\mathbf{R}, \mathbf{Y}^o, \underset{\sim}{\mathbf{X}}, \mathbf{Y}^{(b)}, \underset{\sim}{\mathbf{X}}^{(b)}\right)$. Because the nonrespondent subsample is used only in the superpopulation setting, the nonrespondent subsampling design is left unspecified.

3.4 SUPERPOPULATION MODEL

Under the superpopulation model setting, assume that the vector $\mathbf{y} = (y_1, y_2, ..., y_N)$ is a realization of the random variable $\mathbf{Y} = (Y_1, Y_2, ..., Y_N)$. Let $\mathbf{X}_i$ be the complete set of covariates associated with unit *i,* and let $\mathbf{X}$ represent the covariate data set for all units. Assume that $\mathbf{X}_1, ..., \mathbf{X}_N$ are independent with density $f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\psi})$ for some unknown parameter $\boldsymbol{\psi}$ and assume that the pairs $(Y_i, \mathbf{X}_i)$ are independent. Further assume that the outcomes are independent and identically distributed conditionally on the covariates with densities $f_Y(y_i; \mathbf{X}_i, \varsigma)$. Assume that the sample size *n* is fixed for all *s*. Now the inclusion probabilities and joint inclusion probabilities are defined as:

$$P\left(D_i = 1 \middle| \mathbf{Y}, \mathbf{X}\right) = \pi^{(1)}\left(\mathbf{X}_i, \mathbf{X}\right) = \pi_i$$

and

$$P\left(D_i = 1, D_j = 1 \middle| Y_i, \mathbf{X}_i\right) = \pi^{(2)}\left(\mathbf{X}_i, \mathbf{X}_j, \underset{\sim}{\mathbf{X}}\right) = \pi_{ij} .$$

Similarly, under the superpopulation model, define the response probability as:

$$P\left(R_i = 1 \middle| Y_i, \mathbf{X}_i\right) = p_r\left(Y_i, \mathbf{X}_i; \beta_r\right) = p_{ri}.$$

We also assume that the sampling design and the response mechanism are independent conditional on the outcome of interest and the covariate matrix, *i.e.*

$$D_i \perp R_i \left| Y_i, \mathbf{X}_i \right. .$$

Define a covariate matrix $\mathbf{X}_s$ that contains all the columns of **X**, the response indicator vector, **R**, and any relevant interactions between **R** and columns of **X**. Similarly, define a covariate matrix $\mathbf{X}_r$ that contains all the columns of **X**, the outcome vector, **Y**, and any relevant interactions between **Y** and columns of **X**. Then the covariate matrix, $\mathbf{X}_s$ may be used to model the outcome of interest and X_r may be used to model the response probability, p_{ri}.

Assume that the outcome of interest is a binary indicator of success, where $P\left(Y_i = 1 \middle| R_i, \mathbf{X}_i\right) = p_s\left(R_i, \mathbf{X}_{si}; \beta_s\right) = p_{si}$. The distributions of the outcome of interest and the response indicator are explicitly defined under the superpopulation model by logistic regression models for the binary outcome of interest and the response indicator:

$$E\left(Y_i \middle| R_i, \mathbf{X}_i\right) = p_{si} = \left[1 + \exp\left(-\mathbf{X}'_{si}\beta_s\right)\right]^{-1} \tag{3.3}$$

$$E\left(R_i \middle| Y_i, \mathbf{X}_i\right) = p_{ri} = \left[1 + \exp\left(-\mathbf{X}'_{ri}\beta_r\right)\right]^{-1}, \tag{3.4}$$

where $\boldsymbol{p}_s$ is the vector of "conditional" success propensity scores, β_s is the set of unknown regression coefficients for the success propensity model, $\boldsymbol{p}_r$ is the vector of "conditional" response propensity scores, and β_r is the vector of unknown regression coefficients. The term "conditional" is used to describe $\boldsymbol{p}_s$ and $\boldsymbol{p}_r$ to emphasize their conditionality on the response mechanism and outcome of interest, respectively. We may also express the models for the conditional success and response propensity scores, respectively, as follows:

$$\text{logit}\left[p_s\left(R_i, \mathbf{X}_i; \beta_s\right)\right] = \mathbf{X}_{si}{}'\beta_s,$$

and

$$\text{logit}\left[\boldsymbol{p}_r \left(Y_i, \mathbf{X}_i; \beta_r \right) \right] = \mathbf{X}_{ri}{}' \beta_r \,,$$

where $\text{logit}(a) = \log\left(\dfrac{a}{1-a} \right)$.

3.5 PROPENSITY SCORE METHODOLOGY

When data are MAR, the probability of response depends on the outcome of interest only through covariates related to response; i.e.

$$P\left(R_i \middle| \mathrm{Y}_i, \mathbf{X}_i \right) = P\left(R_i \middle| \mathbf{X}_i \right).$$

The propensity score was originally defined by Rosenbaum and Rubin (1983) under MAR missingness as:

$$e\left(\mathbf{X}_i \right) = P\left(R_i = 1 \middle| \mathbf{X}_i \right),$$

where, in this setting, the response mechanism represents the treatment. Rosenbaum and Rubin (1983) state that $e(\mathbf{X}_i)$ functions as a balancing score, meaning that the distribution of $\mathbf{X}_i$ conditional on $e(\mathbf{X}_i)$ is the same for responding (treated) units and nonresponding (untreated) units. Therefore, the missingness within each level of the balancing score is no longer dependent on the covariates, **X**, so missingness is MCAR within these levels. The finest balancing score is the set of covariates, $\mathbf{X}_i$, and the coarsest balancing score is the propensity score, $e(\mathbf{X}_i)$. We see that, because $\mathbf{X}_i$ is finer than $e(\mathbf{X}_i)$:

$$\begin{aligned} P\left(R_i = 1 \middle| e(\mathbf{X}_i)\right) &= E\left[P\left(R_i = 1 \middle| \mathbf{X}_i, e(\mathbf{X}_i)\right) \middle| e(\mathbf{X}_i)\right] \\ &= E\left[e(\mathbf{X}_i) \middle| e(\mathbf{X}_i)\right] \\ &= e(\mathbf{X}_i) \end{aligned}.$$

Furthermore, because $\mathbf{X}_i$ is finer than $e(\mathbf{X}_i)$, we have that:

$$P\left(R_i = 1 \middle| \mathbf{X}_i, e(\mathbf{X}_i)\right) = P\left(R_i = 1 \middle| \mathbf{X}_i\right) = e(\mathbf{X}_i).$$

Therefore, $P\left(R_i = 1 \middle| \mathbf{X}_i, e(\mathbf{X}_i)\right) = P\left(R_i = 1 \middle| e(\mathbf{X}_i)\right)$ and $\mathbf{X}_i \perp R_i \left| e(\mathbf{X}_i)\right.$. This result implies that the covariates are independent of the response mechanism conditional on the propensity score. Therefore, the distribution of the covariates conditional on the propensity score is the same for respondents and nonrespondents. This balance allows the construction of adjustment classes containing units with similar response rates. Therefore, the independence of the response mechanism and covariates holds approximately, and the MCAR missingness may be assumed within each adjustment class. Quintiles are used to group propensity scores into adjustment classes (Little and Rubin, 2000). The coarsening of the propensity score reduces model dependence while balancing the responding and nonresponding subpopulations relative to covariates. The quintiles can be used to create groups within which response rates are modeled.

When data are NMAR, the probability of response is dependent on both related covariates and the outcome of interest, so $P\left(R_i \middle| Y_i, \mathbf{X}_i, e(\mathbf{X}_i)\right)$ cannot be further simplified without additional information about the missingness mechanism. Therefore, the MAR definition of the propensity score given by Rosenbaum and Rubin (1983) cannot be used to balance the covariate distribution with respect to the response indicator. For NMAR data, the response propensity score is analogous to the probability of response defined earlier as:

$$p_{ri} = p_r\left(Y_i, \mathbf{X}_i\right) = P\left(R_i = 1 \middle| Y_i, \mathbf{X}_i\right) \tag{3.5}$$

Using this new definition of the response propensity score for NMAR missingness, its properties as a balancing score are shown in the proof of the following lemma.

Lemma: The NMAR response propensity score, $p_{ri} = p_r\left(Y_i, \mathbf{X}_i\right)$, is a balancing score for the response mechanism when nonresponse is nonignorable.

Proof: Because $\left(Y_i, \mathbf{X}_i\right)$ is finer than $p_r\left(Y_i, \mathbf{X}_i\right)$, the information provided by $p_r\left(Y_i, \mathbf{X}_i\right)$ is redundant when combined with $\left(Y_i, \mathbf{X}_i\right)$. Using the definition of $p_r\left(Y_i, \mathbf{X}_i\right)$ we have:

$$P\left(R_i = 1 \middle| Y_i, \mathbf{X}_i, p_r\left(Y_i, \mathbf{X}_i\right)\right) = P\left(R_i = 1 \middle| Y_i, \mathbf{X}_i\right) = p_r\left(Y_i, \mathbf{X}_i\right).$$

Conditioning on only the NMAR response propensity score, applying the Law of Iterated Expectations (Billingsley, 1995, p. 448, Theorem 34.4), and using the result from above, we have that:

$$P\left(R_i = 1 \middle| p_r\left(Y_i, \mathbf{X}_i\right)\right) = E\left[E\left(R_i = 1 \middle| p_r\left(Y_i, \mathbf{X}_i\right), Y_i, \mathbf{X}_i\right) \middle| p_r\left(Y_i, \mathbf{X}_i\right)\right]$$

$$= E\left[p_r\left(Y_i, \mathbf{X}_i\right) \middle| p_r\left(Y_i, \mathbf{X}_i\right)\right]$$

$$= p_r\left(Y_i, \mathbf{X}_i\right).$$

Therefore, $P\left(R_i = 1 \middle| Y_i, \mathbf{X}_i, p_r\left(Y_i, \mathbf{X}_i\right)\right) = P\left(R_i = 1 \middle| p_r\left(Y_i, \mathbf{X}_i\right)\right)$ and $\left(Y_i, \mathbf{X}_i\right) \perp R_i \,\middle|\, p_r\left(Y_i, \mathbf{X}_i\right)$. The NMAR response propensity score balances the distribution of the covariates between respondents and nonrespondents, and we have the following result from Rosenbaum and Rubin's Theorem 1 in the notation of Dawid (1979):

$$\left(Y_i, \mathbf{X}_i\right) \perp R_i \,|\, p_r\left(Y_i, \mathbf{X}_i\right). \tag{3.6}$$

□

Let $p_{ri}^{*} = p_{r}^{*}\left(Y_i, \mathbf{X}_i\right)$ be the quintile of the response propensity score associated with the i^{th} unit. Balance holds approximately when conditioning on quintiles, $p_{r}^{*}\left(Y_i, \mathbf{X}_i\right)$, of the response propensity score under nonignorable missingness (Vartivarian and Little, 2003). If the covariates, $\mathbf{X}_i$, are good predictors of response, then the units within each adjustment group based on quintiles of the NMAR response propensity score should have approximately equal response probabilities.

When nonresponse occurs, the outcome of interest is not available. When nonignorable nonresponse occurs, the missing outcomes may be substantially different from the respondent outcomes. Modeling the NMAR response propensity score from the outcome of interest is not possible unless a nonrespondent subsample or a model for the outcome is available. Often, obtaining a subsample of nonresponding units is not possible, such as when a landowner denies access to a survey site. In some cases, such as in large governmental surveys with list frames, a nonrespondent subsample may be obtained. However, regularly obtaining a nonrespondent subsample for periodic surveys may not be feasible due to the extra time, effort, and expense required for each survey. In this case, nonrespondent subsamples conducted every few years may provide the information to create a predictive model for the outcome of interest. This predictive model of success may be applied during years when a nonrespondent subsample is not available.

Assume that the outcome of interest is a binary indicator of success, where $Y_i = 1$ indicates a success. Propensity score methodology is used to obtain estimates of success propensity from a predictive model of success so that the information obtained from the nonrespondent subsample can be used for similar surveys for which a nonrespondent subsample is not available. An implicit assumption is that the success mechanism does not differ substantially between the modeling data set and the estimation data set.

Define the conditional success propensity score as:

$$p_{si} = p_s\left(R_i, \mathbf{X}_i\right) = P\left(Y_i = 1 \middle| R_i, \mathbf{X}_i\right).$$

The conditional success propensity score is shown to be a balancing score for the outcome of interest so that the joint distribution of the success covariates and the response mechanism is the same for successful and unsuccessful units.

Lemma: The conditional success propensity score, $p_s\left(R_i, \mathbf{X}_{si}\right)$, is a balancing score for $\boldsymbol{Y}$.

Proof: Adding $p_s\left(R_i, \mathbf{X}_i\right)$ to the terms of conditioning in the conditional success propensity score, we can simplify the expression because $\left(R_i, \mathbf{X}_i\right)$ is finer than $p_s\left(R_i, \mathbf{X}_i\right)$:

$$P\left(Y_i = 1 \middle| R_i, \mathbf{X}_i, p_s\left(R_i, \mathbf{X}_i\right)\right) = P\left(Y_i = 1 \middle| R_i, \mathbf{X}_i\right) = p_s\left(R_i, \mathbf{X}_i\right).$$

Furthermore, conditioning on only the conditional success propensity score, applying the Law of Iterated Expectation (Billingsley, 1995, p. 448, Theorem 34.4), and using the result from above, we have that:

$$\begin{aligned} P\left(Y_i = 1 \middle| p_s\left(R_i, \mathbf{X}_i\right)\right) &= E\left[E\left(Y_i = 1 \middle| R_i, \mathbf{X}_i, p_s\left(R_i, \mathbf{X}_i\right)\right) \middle| p_s\left(R_i, \mathbf{X}_i\right)\right] \\ &= E\left[p_s\left(R_i, \mathbf{X}_i\right) \middle| p_s\left(R_i, \mathbf{X}_i\right)\right] \\ &= p_s\left(R_i, \mathbf{X}_i\right). \end{aligned} \tag{3.7}$$

□

Therefore, the conditional success propensity score balances the distribution of the covariates between successful and unsuccessful units, rendering the response indicator and related covariates redundant. We may obtain the following result from Rosenbaum and Rubin's Theorem 1, using the notation of Dawid (1979):

$$\left(R_i, \mathbf{X}_i\right) \perp Y_i \,\middle|\, p_s\left(R_i, \mathbf{X}_i\right). \tag{3.8}$$

Let $p_{si}^* = p_s^*\left(R_i, \mathbf{X}_i\right)$ represent the success propensity quintile corresponding to the i^{th} unit. These results hold approximately for quintiles of the conditional success propensity score, $p_s^*\left(R_i, \mathbf{X}_i\right)$, so that $\left(R_i, \mathbf{X}_i\right) \perp Y_i \,\middle|\, p_s^*\left(R_i, \mathbf{X}_i\right)$ approximately.

Often a nonrespondent subsample is not available for every survey exhibiting nonresponse. For these surveys, the outcome of interest is unknown for nonrespondents. One approach is to replace the outcome of interest with its expectation conditional on the response indicator and the covariates. When a nonrespondent subsample is not possible for every survey occasion, information from a previous nonrespondent subsample from a similar population may be used to estimate the expectation of the outcome of interest. Then predictions from the conditional success propensity model may be used as covariates in the response propensity model to account for NMAR missingness when a nonrespondent subsample is not available. Define the conditional response propensity score $p_{cri} = p_{cr}\left(p_{si}, \mathbf{X}_i\right)$ as:

$$p_{cr}\left(p_{si}, \mathbf{X}_i\right) = P\Big(R_i = 1 \Big| E\left(Y_i \middle| R_i, \mathbf{X}_i\right), \mathbf{X}_i\Big)$$

$$= P\Big(R_i = 1 \Big| p_s\left(R_i, \mathbf{X}_i\right), \mathbf{X}_i\Big). \tag{3.9}$$

Notice that $p_r\left(Y_i, \mathbf{X}_i\right)$ conditions on the outcome of interest but $p_{cr}\left(p_{si}, \mathbf{X}_i\right)$ conditions on the conditional success propensity score. The term "conditional" is used to describe the response propensity score to emphasize the dependence on the conditional success propensity score rather than the outcome of interest. Note that under nonignorable missingness, the success propensity for respondents differs from that of nonrespondents for a given set of covariates that are related to the success mechanism. In other words,

$$p_s\left(R_i = 1, \mathbf{X}_i\right) \neq p_s\left(R_j = 0, \mathbf{X}_j\right) \tag{3.10}$$

for $i \neq j$ but for $\mathbf{X}_i = \mathbf{X}_j$.

Now the conditional response propensity score may be expressed as:

$$p_{cri} = p_{cr}\left(p_s\left(R_i, \mathbf{X}_i\right), \mathbf{X}_i\right)$$

$$= P\left(R_i = 1 \middle| p_s\left(R_i, \mathbf{X}_i\right), \mathbf{X}_i\right)$$

for all $r, \mathbf{x}$ with positive probability. By (3.10) we have that $p_s\left(R_i, \mathbf{x}\right) = p_s\left(r, \mathbf{x}\right)$ $\Rightarrow R_i = r$. Therefore, because $\left(R_i, \mathbf{X}_i\right)$ is finer than $p_s\left(R_i, \mathbf{X}_i\right)$ we have that:

$$p_{cri} = p_{cr}\left(p_s\left(R_i = r, \mathbf{X}_i = \mathbf{x}\right), \mathbf{X}_i = \mathbf{x}\right)$$

$$= p_{cr}\left(R_i = r, \mathbf{X}_i = \mathbf{x}\right)$$

$$= P\left(R_i = 1 \middle| R_i = r, \mathbf{X}_i = \mathbf{x}\right)$$

$$= r$$

for all $r, \mathbf{x}$ with positive probability (Birkes, 2009). Since this result holds for all values of R_i, we have that $p_{cri} = R_i$.

The fact that conditioning on the conditional success propensity score reduces to $p_{cri} = R_i$ appears to be a trivial result. However, equation (3.9) defines a useful expression for R_i that leads us to the formation of bias-reducing adjustment cells. We can consider p_{cri} a balancing score for the response indicator because, for $p_{cri} = R_i$, we have that $\left(R_i, \mathbf{X}_i\right) \perp R_i \left| p_{cri} \right.$ because R_i is constant conditional on itself and therefore independent of all random variables. An alternate proof is provided in Appendix B. Balance holds approximately for quintiles of the conditional response propensity score, $\boldsymbol{p}_{cr}^*$.

Define the conditional response propensity score modeled from the quintiles of the conditional success propensity score as $p_{qcr}\left(p_{si}^*, \mathbf{X}_i\right) = P\left(R_i = 1 \left| p_{si}^*, \mathbf{X}_i \right.\right)$. When the quintiles of the conditional success propensity score (p_{si}^*) are used to predict response, units within each class formed by the quintiles are assigned the same success probability, i.e. $p_s^*\left(R_{hi} = 1, \mathbf{X}_{hi}\right) = p_s^*\left(R_{hi} = 0, \mathbf{X}_{hi}\right)$ for all units *i* in success propensity quintile class *h*. Therefore, the trivial solution found above is not obtained when quintiles of the conditional success propensity score are used to predict response because each quintile is not associated with a single response disposition. Both respondents and nonrespondents can occur within each quintile, so coarsening the conditional success propensity score as a predictor of response prevents simplification of the response propensity score to the observed response. In addition to avoiding the issue of the trivial solution, coarsening the conditional success propensity scores into quintiles may reduce success model dependence for the response propensity model. Note that the actual quintile values are use for response model prediction rather than a class designation because the relative quintile values are informative. Let $p^*_{qcr}\left(p^*_{si}, \mathbf{X}_i\right)$ represent the quintiles of $\mathbf{p}_{qcr}$. Balancing properties of the quintiles of $\mathbf{p}_{qcr}$ with respect to $\mathbf{R}$ are expected to be affected by the coarsening of $\mathbf{p}_s$ as a response model predictor.

In summary, an estimator of the success propensity score $\left(p_{si} = p_s\left(R_i, \mathbf{X}_i\right)\right)$ and three estimators of the response propensity are proposed under nonignorable nonresponse: the NMAR response propensity score $\left(p_{ri} = p_r\left(Y_i, \mathbf{X}_i\right)\right)$, the conditional response propensity score $\left(p_{cri} = p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right)$, and the conditional response propensity score from the quintiles of the conditional success propensity score $\left(p_{qcri} = p_{qcr}\left(p^*_{si}, \mathbf{X}_i\right)\right)$. The estimates are obtained in a model-based setting and represent a new application of propensity score methodology. In subsequent chapters, the success and response models are assumed known and applied in model-assisted inferential approaches that permit use of inclusion probabilities from complex survey designs. The performance of these new approaches for nonignorable nonresponse bias adjustment is evaluated for two model-assisted estimators with pilot data for the weighting class adjustment in Chapter 4 and with a novel approach called the *propensity score adjustment estimator* in Chapter 5. In Chapter 6 all estimators are examined with simulations.

CHAPTER 4: WEIGHTING CLASS ADJUSTMENT FOR NMAR DATA WHEN THE OUTCOME IS BINARY AND A NONRESPONDENT SUBSAMPLE IS AVAILABLE

4.1 INTRODUCTION

Superpopulation approaches treat the outcome of interest as a random variable. In the design-based setting, the outcome of interest for unit i, Y_i and the related covariates $\mathbf{X}_i$ are treated as fixed. In model-assisted estimation, adjustments for nonsampling errors are made by assuming models for the error structure while applying design-based tools to account for the survey design (Särndal et al., 1992). In this thesis, model-assisted approaches are used to account for nonsampling error due to nonresponse for data from complex survey designs.

The modified Horvitz-Thompson estimator in equation (3.2) weights each respondent outcome by the inverse of its propensity to respond (Little and Rubin, 2002). This estimator is considered a model-assisted estimator because the superpopulation model is assumed known for the response propensity score. Superpopulation parameters, specifically logistic regression coefficients for the models of the outcome and the response mechanism, are treated as known through conditioning. The weighting class adjustment further incorporates the quasi-randomization assumption to estimate the response propensity within each adjustment class.

First, the motivation for the model assisted setting is established. Then, methodology for the weighting class adjustment and propensity score methodology under nonignorable missingness are reviewed. Finally, the NMAR weighting class adjustment approach is applied to pilot data from a survey of elk hunters in New Mexico and the results are discussed.

4.2 MOTIVATION

Ecological agencies often experience survey nonresponse. For NMAR missingness, inference based only on respondent outcomes may be biased. When a nonrespondent subsample is feasible, the information from nonrespondents may be used in an adjustment procedure for nonresponse called double sampling for stratification. Nonrespondent subsampling often requires considerable effort and cost and may not be possible for every survey that is subject to nonresponse. When the factors affecting nonresponse are consistent over time, information from a nonrespondent subsample may be used to develop a predictive model to reduce nonresponse bias when a nonrespondent subsample is infeasible.

This work adapts MAR methods for NMAR missingness when a nonrespondent subsample is available. First, the MAR adjustment approaches, weighting class adjustment and propensity score classification, are discussed. Second, given that a nonrespondent subsample is available, propensity score classification based on results from Chapter 3 is applied to the weighting class adjustment to adjust for NMAR missingness when the outcome is a binary variable. Finally, a case study is examined in which a survey of elk hunters is found to generate biased estimates of elk harvest when no nonresponse adjustment is incorporated in the analysis.

4.3 MAR WEIGHTING ADJUSTMENT APPROACHES

Weighting adjustment methods are used to adjust the sampling design weights to account for missing data (Little and Rubin, 2002; Lohr, 1999; Oh and Scheuren, 1983). When units are missing, the sampling weights from responding units do not sum to the population size. The observed units must be weighted so that the entire population is represented and estimation is unbiased. MAR weighting adjustment methods are used to weight observed outcomes so that the sample is extrapolated appropriately to the larger population and resulting estimates are unbiased under repeated sampling (Gelman and Carlin, 2002). An extension of the HT estimator for MAR missingness (equation 3.2) weights each outcome by the inverse of its propensity to respond, p_i (Little and Rubin, 2002). This estimator is approximately design-unbiased for the population total if the missingness is MCAR within adjustment cells defined by related covariates (Little and Vartivarian, 2003).

When covariates related to the response mechanism are available, the response propensity (p_i) can be estimated directly by logistic or probit regression (Little and Rubin, 2002). Adjusting the design weights in equation (3.2) with the inverse of $\hat{p}_i$ from the regression model will help to reduce nonresponse bias (Cassel et al., 1983). However, inadequate response within levels of covariate combinations may generate estimated response propensities that are unstable. For example, if estimated response propensities are very small, the variance of the estimator will be inflated.

To avoid unstable estimates from direct estimation, the response propensities may be estimated within coarser groupings of related covariates called adjustment classes. These classes are structured so that response probabilities for subjects within each group are approximately equal. Weighting classes are formed from covariates related to the propensity to respond, the outcome of interest, or a combination of both sets of variables. The weighting class adjustment is used for each weighting class to adjust for nonresponse. Weighting class adjustment and propensity score methodology are two MAR adjustment techniques that employ adjustment classes. These two methods are discussed in a MAR context and then adapted for NMAR missingness when a nonrespondent subsample is available.

4.3.1 Weighting class adjustment

Weighting class adjustment is a nonresponse adjustment originally developed for domain estimation for covariates not used as design strata (Oh and Scheuren, 1983). This estimator is similar to the post-stratification adjustment (Holt and Smith, 1979), except in this case the population sizes are unknown for each adjustment class. The unconditional variance of the estimator includes an additional variance component for random sample sizes, but this is often ignored due to its relatively small effect on the variance (Gelman and Carlin, 2002). The weighting class adjustment estimator is biased to the degree that the estimated population sizes for each adjustment class are biased for the true population sizes for each adjustment class.

Let $\underset{\sim}{h}$ index the set of adjustment classes, $h = 1, \ldots, H$.. Let N_h, n_h, and m_h be the population total, sample size, and number of respondents, respectively, within the h^{th} adjustment class. A general form of the weighting class estimator of the total is given by:

$$\hat{T}_{WC} = \sum_{h=1}^{H} \sum_{i \in S_h} \frac{y_{hi}}{\pi_{hi} \hat{p}_h}, \tag{4.1}$$

where S_h is the set of responding units within weighting class h, y_{hi} is the outcome of interest for unit i with in weighting class h, π_{hi} is the inclusion probability of the i^{th} unit with weighting class h, and $\hat{p}_h$ is the estimated response probability within adjustment cells h =1,..., H.

Response propensities are calculated within each adjustment class assuming the quasi-randomization assumption. Quasi-randomization treats the response mechanism as the second-phase of a two-phase sampling design. In this setting, the first phase is represented by selection in the sample. The second phase is assumed to be a stratified random sample of respondents with weighting classes functioning as strata. The response indicator, **R**, conditional on first-stage sample inclusion, represents an independent Bernoulli process with a common positive probability of response, p_h, $h = 1, \ldots, H$. The response mechanism is assumed uniform within subpopulations and independent among subpopulations. Therefore, within each subpopulation, the missingness mechanism is MCAR. When the quasi-randomization assumption is met, estimation is unbiased within these subpopulations when the estimated subpopulation sizes are unbiased.

Nonresponse adjustment weights, w_i, are computed within each weighting class by one of three methods: post-stratification weighting, weighting class weighting, and inverse probability weighting (Gelman and Carlin, 2002). Post-stratification weighting involves calculating adjustment weights based on known sample and population sizes within each weighting class. Within any given adjustment class, the post-stratification weight is calculated proportional to the number of population units divided by the number of population units within that class, i.e. w_i is proportional to $\frac{N}{N_h}$ for a unit in post-stratum h. When post-stratum information is not available for all units in the population but is available for all units in the sample, the weighting class adjustment is used rather than the post-stratification adjustment. In this case, the adjustment weight is proportional to the number of sampled units divided by the number of sampled units in each class, i.e. w_i is proportional to $\frac{n}{n_h}$ for a unit in post-stratum *h*. The weights for the post-stratification adjustment and the weighting class adjustment are also proportional to the observed response rate within each class, $\frac{n_h}{m_h}$. Inverse probability weighting incorporates the inverse of design probabilities to compute weighting adjustments, such that inverse probability weights are a function of π_i^{-1} . This approach requires that inclusion probabilities are known at the time of survey execution.

Oh and Scheuren (1983) assume a simple random sampling design for the following form of the weighting class adjustment estimator of the population total:

$$\hat{T}_{WC-OS} = \sum_{h=1}^{H} \frac{N}{n} \frac{n_h}{m_h} \tilde{y}_h = \sum_{h=1}^{H} \frac{\hat{N}_h}{m_h} \tilde{y}_h , \tag{4.2}$$

where $\hat{N}_h = \frac{Nn_h}{n}$ is the estimated size of the subpopulation in weighting class *h* and $\tilde{y}_h = \sum_{i=1}^{m_h} y_{i(h)}$. The theoretical variance is approximated as:

$$Var\left(\hat{T}_{WC-OS} \middle| \underset{\sim}{h}\right) \doteq \left(\frac{N}{n}\right)\left(\frac{N-n}{N-1}\right)\sum_{h=1}^{H} N_h \left(\overline{Y}_h - \overline{Y}\right)^2$$

$$+\sum_{h=1}^{H} N_h^2 \left(1 - \frac{\overline{m}_h}{N_h}\right)\frac{V_h}{\overline{m}_h} + \sum_{h=1}^{H} N_h^2 \left(1 - \frac{\overline{m}_h}{\overline{n}}\right)\frac{V_h}{\overline{m}_h^2}, \quad (4.3)$$

where $\overline{n} = \frac{nN_h}{N}$ is the expected overall sample size in cell *h*, $\overline{m}_h = n_h \phi_h$ is the expected number of respondents in cell *h*, ϕ_h is the uniform response probability in weighting class *h*, and V_h is the variance of the outcome of the m_h responding units in weighting class *h*. This estimator uses the weighting class approach to weighting when subpopulation sizes, N_h, are unknown.

The weighting class estimator for inverse-probability weighting (Lessler and Kalsbeek, 1992) is given by:

$$\hat{T}_{WC-IP} = \sum_{h=1}^{H}\sum_{i=1}^{m_h} w_{hi}^* y_{hi}, \quad (4.4)$$

where $w_{hi}^* = \frac{w_{hi}}{p_h}$, $w_{hi} = \pi_{hi}^{-1}$, and $p_h = \frac{\sum_{i=1}^{n_h} w_{hi}}{\sum_{i=1}^{m_h} w_{hi}}$. Here p_h is estimated by inverse-probability weighting. This method provides a more general approach for complex survey designs (Lessler and Kalsbeek, 1992). Variance estimators are not as straightforward to compute with inverse probability weighting, and linearization methods must be used to approximate the variance as a function of the inverse probability weights, w_i (Wolter, 2007; Woodruff, 1971). Little and Vartivarian (2003) caution that estimators based on inverse probability weighting are not necessarily unbiased and suggest that a more robust approach is to use post-stratification weights within weighting classes formed from variables associated with the response mechanism as well as design variables.

Särndal et al. (1992) propose the *response homogeneity group* (RHG) model which, in practice, is equivalent to the quasi-randomization assumption of Oh and Scheuren (1983). Särndal et al. (1992) provide the following general variance estimator for the weighting class estimate of the total under a complex survey design:

$$\hat{V}_{Gen}\left(\hat{T}_{WC}\right)=\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}p_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}+\sum_{h=1}^{H}n_h^2\left(\frac{n_h-m_h}{n_h m_h}\right)S_h^2, \tag{4.5}$$

where S_h^2 is the variance among design-weighted outcomes $\frac{y_{hi}}{\pi_{hi}}$ for respondents in the h^{th} stratum. This variance estimator is a general form that must be tailored to reflect the specific sampling design used. Särndal et al. (1992) give the corresponding variance estimator for the weighting class adjustment under simple random sampling:

$$\hat{V}_{SRS}\left(\hat{T}_{WC}\right) = N^2\left(\frac{N-n}{Nn}\right)\left[\sum_{h=1}^{H}\frac{n_h}{n}\left(1-\frac{n-n_h}{(n-1)m_h}\right)s_h^2 + \frac{n}{n-1}\sum_{h=1}^{H}\frac{n_h}{n}\left(\overline{y}_h - \hat{\overline{y}}\right)^2\right]$$

$$+N^2\sum_{h=1}^{H}\left(\frac{n_h}{n}\right)^2\left(\frac{n_h-m_h}{n_h m_h}\right)s_h^2, \qquad (4.6)$$

where $\overline{y}_h = \frac{1}{m_h}\sum_{i=1}^{m_h} y_{hi}$, $\hat{\overline{y}} = \frac{\sum_{h=1}^{H} m_h \overline{y}_h}{\sum_{h=1}^{H} m_h}$, and s_h^2 is the variance among the outcomes in the h^{th} weighting class. This variance form, which assumes weighting within adjustment classes for unknown subpopulation sizes, will be used for subsequent analyses. Note that this variance estimator does not account for the randomness of the propensity scores on which weighting classes are based.

The weighting class adjustment provides estimation within levels of variables that may or may not be incorporated into the survey design. Stratification reduces the variation of estimates of the population total when outcomes are similar within strata and different among strata (Cochran, 1977). However, stratification may increase the overall variance if the stratification variables are unrelated to the outcome of interest.

Like stratification, weighting class adjustments works best when the variance among adjustment cells is high relative to the variance within adjustment cells (Little, 1986).

Weighting class adjustments depend on the appropriate choice of adjustment variables from which to form adjustment classes. When the dimension of **X** is large, calculation of propensity scores for every covariate combination may result in subgroups that do not contain both treatment and control units. To remedy this problem, one method of obtaining a coarse partition of adjustment is classes is propensity score classification (Rosenbaum and Rubin, 1983). Recall that the estimated propensity score is used in non-randomized studies of a treatment effect to classify treatment and control units into subclasses with similar covariate distributions so that unconfounded inference on the treatment effect may be made. Cochran (1968) suggests that five covariate subclasses are effective in removing roughly 90% of bias. Therefore, quintiles of estimated propensity score from a logistic or probit regression are used to define the subclasses. Within these subclasses, the distribution of the covariates for treatment and control units is approximately equal, i.e. balanced. If the assumption of strongly ignorable missingness can be made, then estimation of the average treatment effect within each subclass is unbiased. Any residual bias is due to the heterogeneity of propensity scores within subclasses.

Propensity score classification is an extension of the propensity score methodology to missing data applications (David et al., 1983). If the propensity score is estimated from covariates that are related to the response mechanism, then grouping on the propensity score creates classes within which the independence of the response mechanism and the covariates holds approximately. A benefit of this approach is the ability to use any number of covariates since they will be summarized by a single variable, the propensity score. This ensures that sample sizes within adjustment classes will be reasonably large for estimation of within-class response rates and variance estimates.

4.3.2 MAR adjustment class construction

The response propensity may be modeled with logistic regression (David et al., 1983). Quantiles of the propensity score form $H = 5$ adjustment classes used in the weighting class adjustment estimators in (4.3) and (4.4). Using design-weighted logistic regression to estimate the propensity scores was found to have little benefit compared to ordinary logistic regression (Little and Vartivarian, 2003). Bias is slightly increased when the sampling weights are included as compared to results from logistic regression modeling that did not incorporate design weights. Therefore, all subsequent discussion of logistic regression models and propensity scores refers to unweighted logistic regression.

One method to form adjustment classes is to create groups based on a classification of the predicted mean (David et al., 1983). Related covariates are used to model the outcome of interest, and then quantiles of the fitted values are used to form adjustment classes. This approach is called "predicted mean classification."

Response propensity classification and predicted mean classification possess different benefits (Little, 1986). Forming adjustment cells from the response propensity produces approximately unbiased estimates of domain and cross-class means (means calculated across levels of adjustment classes), but weighting inflates the variance. However, weighting within adjustment cells based on the predicted mean controls bias and variance for overall means and domain means, but cross-class means may be substantially biased from weighting within groups formed on the predicted mean. Combining response propensity classification and predicted mean classification may be more beneficial than using either approach. Instrumental variable regression and joint classification are two methods that combine predicted mean classification with response propensity classification for MAR data.

In the joint classification approach, adjustment cells are formed by cross-classifying on classes formed from the predicted mean and classes formed from the predicted response probability. Joint classification controls bias, increases efficiency, and allows unbiased estimation of cross-class means (Vartivarian and Little, 2003). With MAR missingness, joint classification enjoys a "double robustness" property where correct specification of either the predicted mean or the response propensity provides benefits. If the predicted mean model is correctly specified but the response propensity model is misspecified, then joint classification controls bias of the overall mean and improves efficiency compared to single classification on the response propensity alone. If only the response propensity model is correctly specified, then joint classification controls bias for overall means and cross-class means. Note that joint classification approaches require the use of 25 adjustment classes rather than the five classes used by single classification approaches because adjustment classes are formed from quintiles of the response propensity score as well as quintiles of the success propensity score.

Instrumental variable regression employs a function of the estimated propensity score to model the predicted mean of the outcome of interest (David et al., 1983). Examples of such functions include predictive mean modeling within classes formed by the propensity score or using standard normal densities of the propensity score as a covariate in a regression model of the outcome. The modeled response propensity score is used as a covariate in the predicted mean model. Including the response propensity score as a predictor of success protects against bias introduced by the misspecification of the relationship between the outcome and covariates. When the predicted mean is accurately modeled within classes formed by the response propensity, the distribution of covariates for respondents and nonrespondents is the same. Then the bias is corrected within adjustment classes where covariate information is balanced even if the model of the outcome is misspecified. Models for the response and for the outcome should be constructed from different covariates to avoid multicollinearity problems (David et al., 1983).

Joint classification and instrumental variable regression for MAR data motivate the development of a methodology that can be used to reduce nonresponse bias when data are NMAR. If information on response from a nonrespondent subsample can be obtained, then the information from that survey may be used to develop models for response that account for its conditional dependence on the outcome of interest. MAR adjustment models that incorporate the subsample data are examined to determine if the data from a nonrespondent subsample is sufficient to reduce bias from nonignorable missingness. Two response propensity models for NMAR data are proposed. A case study involving an annual survey of elk hunters in New Mexico is examined. First, approaches for variable selection and model selection techniques are discussed.

4.3.3 NMAR adjustment class construction

When data are NMAR, the response mechanism is dependent on the outcome of interest, and the distribution of the response mechanism cannot be simplified from the form $f\left(\mathbf{R}\middle|\mathbf{Y},\mathbf{X},\varphi\right)$. To accurately model the outcome, the response indicator must be included as a covariate. Since outcomes are not available for survey nonrespondents, modeling the mean is possible only if nonrespondents can be subsampled or if further modeling assumptions are made.

The goal of this work is to develop a method for obtaining an unbiased estimate of the population total for a binary outcome subject to nonignorable nonresponse. If a subsample of nonrespondents is obtained, then a predictive mean model can provide information on the outcomes of nonrespondents and may be used with similar data sets when nonrespondent subsamples are not available. New techniques are proposed for forming adjustment cells under nonignorable missingness by incorporating the subsample information so that MCAR missingness is achieved within subgroups.

In Chapter 3, three estimators of the response propensity were proposed for nonignorable nonresponse: the NMAR response propensity score $\left(p_{ri} = p_r\left(Y_i, \mathbf{X}_i\right)\right)$, the conditional response propensity score $\left(p_{cri} = p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right)$, and the conditional response propensity score from the quintiles of the conditional success propensity score $\left(p_{qcri} = p_{qcr}\left(p_{si}^{*}, \mathbf{X}_i\right)\right)$. In Chapter 3, the three response propensity score estimators were shown to balance the response mechanism for successful and unsuccessful units when the outcome of interest is a binary response. These three response propensity score estimators are used as the basis for NMAR classification in approaches referred to, respectively, as *NMAR response propensity classification* (indexed by "R"), *conditional response propensity classification* (indexed by "CR"), and *conditional response propensity classification based on quintiles of the conditional success propensity score* (indexed by "QCR"). The conditional success propensity score is also used to form adjustment cells in an approach referred to as *conditional success propensity classification* (SUCC), which reflects predicted mean stratification for NMAR missingness. Additionally, classes formed from each of the response propensity scores are cross-classified with those formed by the conditional success propensity score in the spirit of joint classification (Vartivarian and Little, 2003) and are referred to as *NMAR joint response propensity classification* (JR), *joint* *conditional response propensity classification* (JCR), and *joint conditional response propensity classification based on quintiles of the conditional success propensity score* (JQCR), respectively. The MAR approaches for weighting class formation motivated by David, et al. (1983) and Vartivarian and Little (2003) and the eight NMAR approaches to weighting class formation are described in Table 4.1. These ten weighting class formation approaches are applied to pilot data from annual surveys of elk hunters in New Mexico.

4.4 CASE STUDY

Annual surveys of elk hunters in the state of New Mexico are subject to nonresponse. A subsample of nonrespondents is used to determine if the estimates of annual harvest are biased. These data will be used as a case study to compare models for selecting adjustment classes for weighting class adjustments when data are NMAR.

4.4.1 Background

New Mexico Department of Game and Fish (NMDGF) oversees licensing for annual elk hunts. To maintain a sustainable harvest each year, game managers need accurate estimates of elk harvest. The elk hunter questionnaire is attached to every elk hunting permit sold so that the population of elk licensees is censused. However, since survey return is not mandatory, response rates are rather low (Table 4.2). Game managers feel that successful hunters are more likely to return surveys than licensees who did not harvest an elk. If this claim is true, then annual harvest is overestimated when the results obtained from respondents are extrapolated to the total population. To assess this hypothesis, NMDGF conducted a subsample of nonrespondents to obtain information on elk harvest. From this subsample, model-assisted estimators of elk harvest based on the double sample may be compared to unadjusted estimates from the original sample to evaluate the impact of nonresponse on the elk questionnaire results.

When applying for an elk hunt license, applicants are required to specify the area of the hunt, land ownership type, bag limit, and weapon type of the hunt. Licensee applicants must also provide information on the demographic variables such as age, residency, and gender. Appendix C provides descriptions of all of the covariates available for the data analysis process.

Table 4.1: Success and response model covariates and adjustment cell formation for the weighting class adjustment

WC Adjustment Approach	Missingness mechanism	Success propensity scores used for adjustment classes	Response propensity scores used for adjustment classes	Description
JC	MAR	$\hat{\mathbf{p}}_{sMAR}(\mathbf{X})$	$\hat{\mathbf{p}}_{rMAR}(\mathbf{X})$	Joint classification (Vartivarian and Little, 2003)
IVR	MAR	$\hat{\mathbf{p}}_{sIVR}(\mathbf{p}_{rMAR},\mathbf{X})$	$\hat{\mathbf{p}}_{rMAR}(\mathbf{X})$	Instrumental variable regression (David, et al., 1983)
INT	MAR	$\hat{\mathbf{p}}_{s}(\mathbf{R},\mathbf{X})$	$\hat{\mathbf{p}}_{rMAR}(\mathbf{X})$	Intermediate approach between MAR and NMAR
R	NMAR	-	$\hat{\mathbf{p}}_{r}(\mathbf{Y},\mathbf{X})$	Classification on the NMAR response propensity
CR	NMAR	-	$\hat{\mathbf{p}}_{cr}(\mathbf{p}_{s},\mathbf{X})$	Classification on the conditional response propensity
QCR	NMAR	-	$\hat{\mathbf{p}}_{qcr}(\mathbf{p}_{s}^{*},\mathbf{X})$	Classification on the conditional response propensity based on quintiles of the conditional success propensity
SUCC	NMAR	$\hat{\mathbf{p}}_{s}(\mathbf{R},\mathbf{X})$	-	Motivated by definition of NMAR missingness
JR	NMAR	$\hat{\mathbf{p}}_{s}(\mathbf{R},\mathbf{X})$	$\hat{\mathbf{p}}_{r}(\mathbf{Y},\mathbf{X})$	Joint classification
JCR	NMAR	$\hat{\mathbf{p}}_{s}(\mathbf{R},\mathbf{X})$	$\hat{\mathbf{p}}_{cr}(\mathbf{p}_{s},\mathbf{X})$	Novel approach
JQCR	NMAR	$\hat{\mathbf{p}}_{s}(\mathbf{R},\mathbf{X})$	$\hat{\mathbf{p}}_{qcr}(\mathbf{p}_{s}^{*},\mathbf{X})$	Novel approach

Table 4.2: Elk hunter harvest survey return rates by year

Hunt Year	**Elk Harvest Survey Return Rate (%)**
1988-1989	29.7
1989-1990	30.9
1990-1991	30.1
1991-1992	31.7
1992-1993	32.8
1993-1994	30.4
1994-1995	26.9
1995-1996	26.1
1996-1997	27.8
1997-1998	6.1*
1998-1999	40.2
1999-2000	44.9
2000-2001	24.4
2001-2002	29.5
2002-2003	NA**
2003-2004	24.2

* Questionnaires for the 1997-98 hunt year appear to have been lost at the post office.
** Respondent information unavailable to author.

The population of licensees for the 2001-02 hunts consists of 38,209 licensees. A total of 11,258 licensees responded to the tear-off survey questionnaire via mail or the Department's web site for a total response rate of 29.5%. A total of 4,181 respondents reported bagging an elk for a reported harvest rate of 37.1%. Using a simple extrapolation of sample results to the entire population, an estimated 14,190 elk were harvested statewide during the 2001-02 hunt year with a 95%-confidence interval of (13849, 14531).

From the population of 26,951 nonrespondents, a two-stage cluster sample was selected. Hunts served as primary sampling units and licensees within hunts acted as secondary sampling units. A total of 190 hunts were randomly selected from the population of 572 hunts for which nonresponse occurred. This first-stage sample of primary sampling units represented a total of 9,113 nonrespondents for the 190 hunts. The second stage sample was conducted employed unequal probability sampling for different hunt size classes. For small hunts (≤ 30 licensees) selected in the first-stage sample, the second-stage licensees were censused. Medium-sized hunts (31 to 170 licensees) were sampled at a rate of 0.30 and large hunts (> 170 licensees) were sampled at a rate of 0.15. The second-stage sample consisted of 3,078 nonresponding licensees. Mail surveys were undeliverable for 188 licensees, reducing the effective sample size to 2,890 licensees. The sample was stratified by Weapon Type, Landowner Type, and Hunt Size. An overall 82.2% response rate was obtained from three mailing/internet waves followed by a telephone survey. This response rate was adjusted for undeliverable mail surveys. Most of the remaining nonresponse in the subsample was due to incorrect or out-of-service telephone numbers. Response status was not tracked by license number, so the effective sample size could not be decreased for invalid telephone numbers.

For the 2003-04 hunt season, a total of 40,503 elk hunt licenses were issued in New Mexico and 24.2% of the questionnaires were returned. The unadjusted harvest estimate from the 2003-04 questionnaire survey was calculated as 15,446 elk with a

95%-confidence interval of (15056, 15835). A nonrespondent subsample from the 30,693 nonrespondents was also obtained for the 2003-04 hunt season, and similar design and analysis methods were used. A total of 175 hunts were randomly selected from the population of 555 hunts for which nonresponse occurred. This first-stage sample of primary sampling units represented a total of 10,965 nonrespondents for the 175 hunts. A second-stage subsample of 3,019 licensees was selected. Mail surveys were undeliverable for 271 licensees, decreasing the effective subsample size from 3,019 to 2,748 licensees. An overall nonrespondent subsample response rate of 81.0% was observed from three mailing and internet waves and a telephone follow-up. Again, the response rate was adjusted for undeliverable mail surveys but response status was not tracked by license number so response rates cannot be adjusted for inoperable telephone numbers.

Contingency tables summarizing the distribution of response and success for each year are provides in Table 4.3 Chi-square tests of homogeneity between harvest and success rates indicate that response is not independent of success in 2001 (χ^2 =204.45, df = 1, p-value < 0.0001) or in 2003 (χ^2 =174.61, df = 1, p-value < 0.0001). The odds ratio of the odds of response for successful licensees against the odds of response for unsuccessful licensees is 2.12 for 2001 and 2.35 for 2003. An odds ratio that is greater than one indicates that successful units are more likely to respond than unsuccessful units. The result of a Mantel-Haenszel test for a common odds ratio of 1 indicates that the odds of response for successful licensees compared to unsuccessful licensees is significantly different from 1 (χ^2 =379.49, df = 1, p-value < 0.0001) across the two surveys. This implies that the difference in response rates between successful licensees and unsuccessful licensees is not likely due to chance and that nonresponse bias may be a persistent issue in the NMDGF elk hunter survey. Notice that the success rates are fairly consistent for the respondents and nonrespondents to the initial survey between the two years. This suggests that nonresponse bias may also be consistent in its effect on estimates of total elk harvest.

Table 4.3: Contingency table for response and success for the 2001 and 2003 NMDGF elk licensee surveys and nonrespondent subsamples

	2001		2003	
	Respondents to initial survey	Nonrespondents to initial survey contacted in the nonrespondent subsample	Respondents to initial survey	Nonrespondents to initial survey contacted in the nonrespondent subsample
Successful	4181 (0.37)	517 (0.22)	3746 (0.38)	519 (0.23)
Unsuccessful	7077 (0.63)	1858 (0.78)	6069 (0.62)	1708 (0.77)
TOTAL	11258	2375	9815	2227

4.4.2 Double-sampling for stratification

The nonrespondent subsample was conducted using double-sampling for stratification, a sampling design that uses a two-phase sampling approach to obtain outcomes from first-phase nonrespondents (Thompson, 1992). The design-based estimator provides an unbiased estimate of the population total with which the estimates from the model-assisted approaches may be assessed. The estimate of the population total for the double-sampling for stratification is given by:

$$\hat{T}_d = N\overline{y}_d = N\left(\frac{n_1'}{n'}\overline{y}_1 + \frac{n_2'}{n'}\overline{y}_2\right), \tag{4.7}$$

where $\overline{y}_d$ is the double-sampling for stratification estimator of the population mean, n' is the phase 1 sample size, n_1' is the number of first-phase respondents, and n_2' is the number of first-stage nonrespondents where $n_2' = n - n_1'$ (Thompson, 1992). The variance estimate of the estimate of the population total for the double-sampling for stratification is:

$$\hat{Var}\left(\hat{T}_d\right) = N^2 Var\left(\overline{y}_d\right)$$

$$= N^2\left(\frac{N-1}{N}\right)\left[\left(\frac{n_1'-1}{n'-1} - \frac{n_1'-1}{N-1}\right) w_1 \hat{Var}\left(\overline{y}_1\right) + \left(\frac{n_2'-1}{n'-1} - \frac{n_2-1}{N-1}\right) w_2 \hat{Var}\left(\overline{y}_2\right)\right]$$

$$+\left(\frac{N-n'}{N\left(n'-1\right)}\right)\left[w_1\left(\overline{y}_1 - \overline{y}_d\right)^2 + w_2\left(\overline{y}_2 - \overline{y}_d\right)^2\right] \qquad (4.8)$$

where N is the population size, n_2 is the size of the phase-two sample, $w_h = \frac{n_h'}{n'}$, $\hat{Var}\left(\overline{y}_1\right)$ is the estimated variance of phase 1 outcome mean, and $\hat{Var}\left(\overline{y}_2\right)$ is the estimated variance of phase-two outcome mean. In the NMDGF case study, the licensees are censused each year, so the phase 1 sample size $n' = N$. Censusing at the first phase simplifies the variance estimator to the following form:

$$
\begin{aligned}
Var\left(\hat{T}_d\right) &= N^2\left(\frac{N-1}{N}\right)\left[\left(\frac{n_2'-1}{n'-1}-\frac{n_2-1}{N-1}\right)w_2\,\hat{Var}\left(\bar{y}_2\right)\right] \\
&= N^2\left(\frac{N-1}{N}\right)\left[\left(\frac{n_2'-1}{N-1}-\frac{n_2-1}{N-1}\right)\frac{n_2'\,\hat{Var}\left(\bar{y}_2\right)}{N}\right] \\
&= \left(n_2'-n_2\right)n_2'\,\hat{Var}\left(\bar{y}_2\right)
\end{aligned}
$$

The second-phase nonrespondent subsample for the NMDGF elk hunter survey employed a two-stage cluster sample of hunts stratified by weapon type, landowner type, and hunt size with second-stage licensees selected within first-stage hunts. In this setting, the *population* refers to the population of nonresponding units to the first-phase sample. The mean of the outcome from a two-stage stratified sampling design is given by:

$$
\bar{y}_2 = \frac{1}{n_2'}\sum_{l=1}^{L}\frac{N_l}{n_l}\sum_{i=1}^{n_l}M_{li}\bar{y}_{li}\,,
$$

where l indexes the L strata in the second phase of the double sample such that l =1, …, L; N_l is the number of hunts in the population in stratum l, n_l is the number of sampled hunts in stratum l, M_{li} is the number of licensees in the i^{th} hunt of stratum l, and $\bar{y}_{li}$ is the mean of the outcome within stratum l and hunt i. The estimated variance of the phase-two mean from a two-stage stratified sampling design is:

$$\hat{Var}(\bar{y}_2) = \frac{1}{N^2}\sum_{l=1}^{L}\left(N_l\left(N_l - n_l\right)\frac{s_l^2}{n_l} + \frac{N_l}{n_l}\sum_{i=1}^{n_l} M_{li}\left(M_{li} - m_{li}\right)\frac{s_{li}^2}{m_{li}} \right),$$

where m_{li} is the number of licensees in the sample for the i^{th} hunt of stratum l,

$s_l^2 = \frac{1}{n_l - 1}\sum_{i=1}^{n_l}\left(\bar{y}_{li} - \bar{y}_l\right)^2$, s_{li}^2 is the sample variation for the i^{th} hunt of stratum l,

$\bar{y}_l = \frac{1}{n_l}\sum_{i=1}^{n_l}\bar{y}_{li}$, and $\bar{y}_{li}$ is the sample mean among licensees in the i^{th} hunt of stratum l.

The estimates from the double-sampling for stratification (DSS) analysis demonstrate that the design-based estimator provides a much smaller estimate of harvest than the unadjusted estimates obtained by simple extrapolation of the respondent outcomes to the population (Table 4.4). Dividing the 2001 and 2003 estimates of the total elk harvest from double sampling for stratification by the total number of licensees (38209 and 40503, respectively) yields estimates of success rates of 0.28 and 0.29, respectively.

Table 4.4: Double-sampling for stratification and unadjusted estimates of total elk harvest for 2001 and 2003

Estimation Approach	**Metric**	**2001**	**2003**
DSS	Est. Total 95%-CI	10523 (10299, 10747)	11672 (11339, 12005)
Unadjusted estimates	Est. Total 95%-CI	14190 (13849, 14531)	15446 (15056, 15835)

NMDGF personnel would like to develop a method to adjust survey return information so that unbiased estimates of harvest might be obtained for years in which a subsample was not conducted. The NMAR models to obtain weighting adjustment classes for weighting class adjustments are examined with the data from the two elk hunt years. The 2001-02 data will serve as the modeling data set. For approaches that do not require a nonrespondent subsample complement to every survey, performance is evaluated with the 2003-04 data set. The double sampling for stratification estimates that combine the original surveys and nonrespondent subsamples will be assumed to be the "true" values and will be used to assess the performance of the NMAR weighting class adjustments.

These data sets are relatively rich in covariate information. Identification of the appropriate predictors of success and response are integral to unbiased estimation. Variable and model selection techniques are reviewed and the results of the models are compared for the two hunt years for the weighting class adjustment.

4.4.3 Variable selection

Because all elk hunts in New Mexico limit the hunter to a total of one elk, the outcome of interest is a binary indicator of elk harvest. Therefore, the logistic regression model may be used to model both the response propensity and the success propensity. Following recommendations for logistic regression modeling from Hosmer and Lemeshow (2000), variable selection begins with a univariate analysis of all potential covariates. Pearson's chi-square test of association is used to establish which variables are associated with response and success (Table 4.5).

Table 4.5: Pearson chi-square tests of association for indicators of response and success

Variable	Response	Success
Weapon Type	29.06 ($p < 0.0001$)	379.84 ($p < 0.0001$)
Hunt Size	636.82 ($p < 0.0001$)	4.22 ($p = 0.1261$)
Age Class	146.61 ($p < 0.0001$)	22.90 ($p = 0.0001$)
Resident	44.15 ($p < 0.0001$)	449.29 ($p < 0.0001$)
Bag	66.23 ($p < 0.0001$)	668.95 ($p < 0.0001$)
Area	139.60 ($p < 0.0001$)	275.91 ($p < 0.0001$)
Land Type	2.34 ($p = 0.1261$)	1146.49 ($p < 0.0001$)
Month	117.85 ($p < 0.0001$)	410.56 ($p < 0.0001$)
Sex	0.1123 ($p = 0.7375$)	0.0405 ($p = 0.8405$)

David et al. (1983) recommend that response propensity models and predicted mean models should employ mutually exclusive sets of covariates to avoid multicollinearity problems. Based on the results of the chi-square tests of association, the variables Weapon Type, Residency, Bag Limit, Area, and Month exhibit significant associations with both the response indicator and the success indicator. However, the test results for these variables indicate higher significance for tests of association with success. Age Class also exhibits associations with both response ($p < 0.0001$) and success ($p = 0.0001$) but the p-value is smaller for the association with response. Hunt Size is associated with response ($p < 0.0001$) but not success ($p = 0.1261$). Land Type is associated with success ($p < 0.0001$) but not response ($p = 0.1261$). Sex is not associated with either response (0.1123) or success ($p = 0.8405$), so this variable is not used for response propensity or success propensity modeling. Therefore, success will be modeled from Weapon Type, Residency, Bag Limit, Area, Land Type, and Month while response will be modeled from Hunt Size and Age Class. The covariate for licensee sex does not appear closely associated with either outcome, so this covariate is not used for multivariable modeling of either success or response.

4.4.4 Regression estimation and model selection

The goal of model selection in this exercise is to obtain a robust predictive model that is applicable among several years of survey data. If the model is overfit, it may not apply well in other survey years. If the model is underspecified, then it may not be sensitive enough to produce accurate predictions. Design-weighted logistic regression was ineffective in providing estimated propensity scores that were effective in reducing bias. Little and Vartivarian (2003) also find that logistic regression weighted by design weights was not beneficial. Therefore, unweighted logistic regression modeling is used to model success and response propensities. Maximum penalized likelihood estimates (Firth, 1993) of regression coefficients are used to ensure that the propensity scores are unbiasedly estimated.

Shrinkage estimation was considered for the estimation of parameters from logistic regression. Steyerberg, Eijkeman, Harrell, and Habbema (2000) recommend shrinkage estimators for regression coefficients when data sets from clinical trials are small and prediction is the goal. However, shrinkage will not improve mild overfitting when fitted values are used for ranking (Harrell, Lee, and Mark, 1996). This caveat may then also apply to the formation of propensity score classes since this approach involves ranking. Shrinkage is not deemed necessary when the ratio of the number of model coefficients to the number of observations is less than 0.02 (Shtatland, Kleinman, and Cain, 2004). This ratio is roughly 0.0019 for the pilot data set, so estimation with shrinkage is unnecessary for these data.

Several tests and information criteria are available to select the appropriate logistic regression models for success and response. Hosmer and Lemeshow (2000) suggest stepwise selection using the likelihood ratio test to compare models. The p-value for variable entry into the model should range from 0.15 to 0.20. The p-value for removing a variable should exceed the p-value for entry to avoid repeated entry and removal of the same variable. For a very full model, the removal p-value may be as high as 0.9 while a more parsimonious model may employ a removal p-value only slightly larger than the entry p-value. The likelihood ratio test performs poorly when the probabilities of entry and removal are low (Steyerberg et al., 2000). While Hosmer and Lemeshow (2000) prefer the likelihood ratio test for stepwise procedures, their example analyses produced the same models when score and Wald tests were used.

Model selection with AIC is equivalent to backwards elimination with a probability of removal of 0.157 (Steyerberg et al., 2000). Benefits of AIC include its ability to optimize model parsimony, estimation accuracy, testing of both nested and unnested models, and its asymptotic equivalence to cross-validation techniques (Shtatland, Moore, Dashevsky, Miroshnik, Cain, and Barton, 2000). AIC penalizes new terms less severely than both likelihood ratio tests and Wald tests (Venables and Ripley, 2002). This tendency toward overfitting may lead to bias when applied to small samples because AIC loses its asymptotically optimal properties. AIC is not consistent, so the probability of selecting the wrong model does not decrease as the number of predictors increases, as is true with the BIC. For this large pilot data set, use of the AIC should satisfy asymptotic properties.

The best subset method is another model selection procedure that allows identification of a specified number of models that meet a selection criterion for a specified number of variables. This model selection strategy is implemented in SAS (2002) PROC Logistic using the score test statistic as the selection criterion for the "best" models. Hosmer and Lemeshow (2000) prefer Mallows (1973) C_p statistic and show how this criterion may be calculated from the score test results provided in the output from SAS Proc Logistic. The pilot data set, which includes all main effects and second-order interactions, consists of 71 explanatory variables. This large number of potential covariates motivates the use of other model selection methods to reduce the dimension of the data set for best subsets modeling and to suggest a potential range for the number of explanatory variables. Best subset models will be obtained from the reduced data set for the range of possible covariate sizes.

Based on the available literature, the model selection approach will begin with two stepwise model selection approaches. The first approach will use the score test for variable entry and the Wald test for variable removal. This approach will be implemented in SAS Proc Logistic. The entry and removal p-values of 0.15 and 0.20, respectively, will produce a more restricted model. A stepwise selection procedure using AIC as the evaluating criterion will also be examined using the *stepAIC* function from the R MASS package. Given the tendency for AIC to overfit, this stepwise selection approach should produce a fuller model than that obtained from SAS Proc Logistic. The significant covariates from both stepwise approaches will be used in the "best subsets" procedure in SAS Proc Logistic. For a range of sizes of model covariate data sets, Mallows C_p statistic is computed and used to compare models. Mallows C_p statistic is calculated for each of the best subsets models as well as for the models selected with stepwise regression, and the model with the lowest C_p statistic is selected. Goodness of fit is assessed with the Pearson chi square goodness-of-fit test.

The best subsets model selection approach in SAS produces models that are improper in the sense that interactions are included for excluded main effects. Both stepwise selection procedures produce proper models. The final model is selected among the set of proper models obtained from the two stepwise procedures, the proper model with the lowest C_p statistic from the best subsets approach, and the model with the lowest C_p statistic from the best subsets approach with main effects added to make the model proper. In almost every case, the lowest C_p best subsets model that is made proper involves fewer design variables and removes more bias than the proper model with the lowest C_p statistic.

Three success models and four response models are considered for adjustment cell formation. All models incorporate data from the original sample and the subsample. Success is modeled as MAR in the joint classification approach (Little, 1986; Vartivarian and Little, 2003). In instrumental variable regression, the response propensity score (p_{rMAR}) under MAR missingness is used as a predictor in the success model (David et al., 1983). For the intermediate approach and all of the NMAR approaches, the success model is a function of the response indicator and is only possible when a nonrespondent subsample is available.

The response model for the joint classification approach (Little, 1986; Vartivarian and Little, 2003), instrumental variable regression (David et al., 1983), and the intermediate model is obtained independently from the success model assuming MAR missingness. The NMAR response model (indexed by R) employs the nonrespondent subsample to model response as a function of the outcome of interest. The conditional response model (indexed by CR) includes the conditional success propensity score as a predictor rather than the outcome of interest. The model indexed as QCR incorporates the conditional response propensity score modeled from quantiles of the conditional success propensity score. The variables selected in the model selection approach are given in Appendix D for each success and response propensity model.

4.4.5 Results and discussion

The weighting class adjustment estimates, 95%-confidence intervals, root mean squared error (RMSE), and relative bias from the ten different weighting class adjustment approaches are given in Table 4.8 with the double sampling for stratification and unadjusted estimates for comparison. Relative bias is calculated relative to the design-based estimate obtained from the double sample for stratification. Despite this estimator's design-unbiasedness, note that this estimate may still exhibit some bias due to nonresponse. When adjustment cells contain fewer than 20 respondents, these adjustment cells are absorbed into other adjustment cells so that response rates within adjustment cells may be accurately estimated (Lohr, 1999). The weighting class adjustment is then computed for the set of adjustment cells for which at least 20 respondents were obtained.

The two MAR estimators, joint classification and instrumental variable regression, are ineffective in nonresponse adjustment for these data as compared to the double sampling for stratification estimator. The joint classification approach exhibits relative bias of 0.32 for both 2001 and 2003. These estimates are only slightly less biased than the unadjusted estimates but still indicate that the MAR joint classification is inappropriate for these data. Similarly, MAR instrumental variable regression generates estimates of total elk harvest with relative bias of 0.31 for 2001 and 0.34 for 2003 compared to the double sampling for stratification estimators. These results underscore that appropriate models must be used for adjustment cell formation when data are NMAR, even when information from the nonrespondent subsample is used.

Table 4.6: 2001 and 2003 weighting class adjustment estimates

Estimator	Metric	2001	2003
WC Adj. (Joint classification) (JC)	Est. Total	13888	15404
	95%-CI	(13618, 14158)	(15058, 15751)
	RMSE	3368	3736
	Rel. Bias	0.32	0.32
WC Adj. (Instrumental variable regression) (IVR)	Est. Total	13805	15593
	95%-CI	(13537, 14074)	(15247, 15939)
	RMSE	3285	3925
	Rel. Bias	0.31	0.34
WC Adj. (Intermediate) (INT)	Est. Total	11409	13525
	95%-CI	(11010, 11809)	(12944, 14107)
	RMSE	909	1877
	Rel. Bias	0.08	0.16
WC Adj. (R)	Est. Total	9460	12207
	95%-CI	(9234, 9686)	(11912, 12502)
	RMSE	1069	556
	Rel. Bias	-0.10	0.05
WC Adj. (CR)	Est. Total	11796	13058
	95%-CI	(11388, 12203)	(12687, 13430)
	RMSE	1290	1399
	Rel. Bias	0.12	0.12
WC Adj. (QCR)	Est. Total	12284	13557
	95%-CI	(11924, 12643)	(13188, 13927)
	RMSE	1771	1894
	Rel. Bias	0.17	0.16

Estimator	Metric	2001	2003
WC Adj. (JR)	Est. Total 95%-CI RMSE Rel. Bias	9356 (8421, 10292) 1261 -0.11	10656 (10147,11165) 1049 -0.09
WC Adj. (JCR)	Est. Total 95%-CI RMSE Rel. Bias	11327 (10760, 11894) 854 0.08	13409 (12796, 14021) 1765 0.15
WC Adj. (JQCR)	Est. Total 95%-CI RMSE Rel. Bias	11479 (10999, 11959) 987 0.09	13742 (13076, 14408) 2098 0.18
WC Adj. (SUCC)	Est. Total 95%-CI RMSE Rel. Bias	11109 (10462, 11757) 673 0.06	13128 (12573, 13684) 1483 0.12
Double sampling for stratification (DSS)	Est. Total 95%-CI RMSE Rel. Bias	10523 (10299, 10747) 114 0.00	11672 (11339, 12005) 170 0.00
Unadjusted estimates (MCAR)	Est. Total 95%-CI RMSE Rel. Bias	14190 (13849, 14531) 3671 0.35	15446 (15056, 15835) 3779 0.32

The intermediate method incorporates a MAR model for response with a NMAR model for success. Therefore, the response model is the same as that used in the joint classification and instrumental variable approaches, but the success model includes the response indicator as a predictor. The intermediate approach is more effective than either of the two MAR approaches, with relative bias of 0.08 for 2001 and 0.16 for 2003. The RMSE is also considerably smaller for the intermediate approach compared to the estimates from the MAR approaches and the unadjusted estimates.

The R approach, which requires a nonrespondent subsample for every sample, underestimates the 2001 total elk harvest as compared to the double sample for stratification estimator, with relative bias of -0.10. However, the 2003 estimate overestimates the total by 5%. The CR approach, which incorporates the conditional success propensity score in the response propensity model rather than the outcome of interest, generates estimates of 2001 and 2003 total elk harvest with relative biases of 0.12. Similarly, the QCR approach, which incorporates quintiles of the conditional success propensity scores as predictors in the response propensity model, exhibits slightly more positive bias than the CR approach for both the 2001 and 2003 estimates of the total elk harvest with relative bias of 0.17 and 0.16, respectively.

NMAR joint classification approaches perform with mixed results. The JR approach, which forms adjustment cells on quintiles of the NMAR response propensity score and on quintiles of the conditional success propensity score, performs similarly to the R approach for 2001 and underestimates the harvest total by 11%. However, absolute relative bias for the 2003 data increases when NMAR joint classification is used, taking the relative bias from 0.05 for the R approach to -0.09 for the JR approach. The opposite pattern is observed for the JCR approach. NMAR joint classification improves the 2001 estimate of total elk harvest, decreasing the relative bias to 0.08 from 0.12 when joint classification is not used (CR). However, NMAR joint classification slightly increases the 2003 relative bias increases to 0.15 for the JCR approach compared to 0.12 from the CR approach. The JCQR approach performs similarly, with a decrease in the 2001 relative bias from 0.17 for the QCR approach to 0.09 with the JQCR approach and an increase in the 2003 relative bias from 0.16 for the QCR approach to 0.18 with the JQCR approach. The increases in bias from NMAR joint classification may result from applying the 2001 conditional success propensity model to the 2003 data.

Using the conditional success propensity score as the basis for adjustment cell formation (SUCC) performs well with relative bias of 0.06 for 2001 and 0.12 for 2003. This result illustrates that the conditional success propensity score are also effective as a balancing score. This approach generated the least biased estimate of the harvest total and the smallest RMSE across all of the weighting class adjustment estimates obtained for the 2001 data.

The only weighting class adjustment approach yielding confidence intervals that cover the design-based estimates of total elk harvest obtained from double sampling for stratification is the SUCC approach for the 2001 data. Otherwise, the lack of coverage indicates either poor confidence interval coverage by the weighting class adjustment or bias in the design-based estimates. Simulations will be conducted (Chapter 6) and these approaches for adjustment class formation will be assessed with known population totals to determine which perform best over a range of conditions.

4.5 CONCLUSIONS

Extensions of propensity score methodology to nonignorable missingness were proposed in Chapter 3. By accounting for the NMAR missingness mechanism in logistic regression models of the response indicator and/or binary outcome of interest, response and/or success propensity scores are used to form adjustment classes that satisfy the quasi-randomization assumption for the weighting class adjustment (Oh and Scheuren, 1983). These new approaches for forming adjustment cells under NMAR missingness are motivated by propensity score methodology (Rosenbaum and Rubin, 1983) and MAR nonresponse adjustment approaches, including joint classification (Little, 1986, Vartivarian and Little, 2003) and instrumental variable regression (David et al., 1983).

In this case study, MAR approaches to weighting class formation are not effective in reducing nonresponse bias as compared to the design-based estimates from double sample for stratification. NMAR approaches to forming weighting classes are more effective in reducing nonresponse bias. However, all but one (SUCC approach for the 2001 data) of the confidence intervals for the MAR and NMAR weighting class adjustment approaches did not cover the design-based estimates of elk harvest obtained from double sampling for stratification. Simulation will be used to determine if low coverage rates are due to poor coverage from the weighting class adjustment or bias in the design-based estimates obtained from double sampling for stratification.

The response propensity score quintiles are used in this chapter to form adjustment cells within which response rates are estimated by weighting class adjustments. The response propensity scores are more informative than simple adjustment cell classification variables. The response propensity scores and quintiles of the response propensities may be used as direct estimates of the response probability within each adjustment class. In Chapter 5, this additional model dependence is explored with the modified Horvitz-Thompson estimator of equation (3.2) as an improvement on the weighting class adjustment for NMAR data. Chapter 6 will include simulation results for testing and comparing the performance of all approaches under a range of conditions.

5: THE PROPENSITY SCORE ADJUSTMENT ESTIMATOR FOR A MISSING BINARY RESPONSE UNDER NONIGNORABLE NONRESPONSE

5.1 INTRODUCTION

When data are not-missing-at-random (NMAR), the distribution of the response indicator is dependent upon the outcome of interest. Therefore, the response propensity score and predicted mean of the outcome cannot be estimated independently. In Chapter 3, propensity score methodology was extended to the case of NMAR missingness. For a binary outcome, the conditional success propensity score is modeled from the response indicator when data are available from a subsample of nonrespondents. The conditional success propensity score or its quintiles are used as covariates in the response propensity model when the outcome of interest is missing. This mutual dependence between the response and the outcome of interest is consistent with the definition of NMAR missingness.

In Chapter 4, weighting class adjustments were used for different approaches to adjustment cell formation based on several ignorable and nonignorable models of missingness. When these techniques were applied to the case study data, the NMAR approaches for weighting class formation reduced nonresponse bias. However, confidence intervals did not cover the design-based estimators, indicating either poor coverage or bias in the estimates of the total obtained from double sampling for stratification.

When the response propensity model accounts for nonignorable nonresponse, is not just a covariate for NMAR weighting class formation; it is also an estimate of the response rate under the NMAR assumption. In this chapter, response propensity scores are estimated with different MAR and NMAR approaches and the response propensity scores or their quintiles are used directly as estimates of the response probability in a modified Horvitz-Thompson estimator for nonresponse (equation 3.2). This estimator, referred to as the "propensity score adjustment estimator" (PSAE), and its variance are developed and then applied to the case study data used in Chapter 4.

5.2 PROPENSITY SCORE ADJUSTMENT ESTIMATOR

In the MAR missingness case, Cassel et al. (1983) applies the predicted response propensities from logistic regression to directly adjust the modified Horvitz-Thompson estimator in equation (3.2). This approach, while unbiased for accurate estimation of response propensity scores, may produce inflated variance estimates when some of the estimated response propensities are very small (Little and Rubin, 2002). Several approaches to weight trimming have been proposed to avoid the problem of variance inflation (Deville and Särndal, 1992; Elliott and Little, 2000; Potter, 1988). Propensity score methodology from Chapter 3 is used to obtain unbiased estimates of the response propensity under NMAR missingness. Furthermore, quantiles of the response propensity scores are used to trim response propensity weights so that extreme weights do not inflate the variance of the estimator.

The propensity score adjustment estimator incorporates several different approaches for estimating the response propensity when nonresponse is nonignorable. Three main models for the response propensity score under nonignorable nonresponse were discussed in Chapter 3. The differences among the three models are distinguished by the information available for predicting success. The NMAR response propensity score (p_{ri}) is predicted directly from the outcome of interest when a nonrespondent subsample is available. The conditional response propensity score (p_{cri}) is modeled from estimates of the expectation of the outcome of interest from a response model, which are the estimated conditional success propensity scores when nonresponse is nonignorable. In Chapter 3, we found that the conditional response propensity score based on the conditional success propensity score $\left(p_{si}\right)$ is either 0 or 1. We attempt to avoid this trivial result by modeling response from the quintiles of the conditional success propensity score. Using quintiles of the conditional success propensity score may also reduce success model dependence. The conditional response propensity scores modeled from the quintiles of the estimated conditional success propensity scores are (p_{qcri}). Furthermore, we will compare the results to the Cassel et al. (1983) estimator which uses the estimated response propensities from a MAR logistic regression model of the response indicator (p_{rMARi}).

For each of the four response propensity models, the inverse of the estimated response propensity score may be used directly to weight each respondent outcome, or the weights may be coarsened into quintiles that are then used in the modified Horvitz-Thompson estimator from equation (3.2). These scenarios produce eight models for estimating the response propensity under nonignorable nonresponse. After a review of notation, the approach used to coarsen the response rate is discussed. Subsequently, the propensity score adjustment estimator and its variance estimator are derived and applied to the NMDGF case study data. The predictive performance of the conditional success propensity score and its quintiles will be assessed. The effects of coarsening the response propensity scores on the estimator of the total will also be examined.

5.2.1 Notation and assumptions

Assume that inference on the population total, τ, is made based on a sample of n units selected from a population of size N. The sampling design may be simple or complex, but inclusion probabilities are assumed known. Let $\mathbf{y} = (y_1, y_2,\ldots, y_N)$ be the complete data binary outcome of interest where, for $i = 1,\ldots N,$

$$y_i = \begin{cases} 1, & \text{unit } i \text{ is successful} \\ 0, & \text{otherwise.} \end{cases}$$

As in the weighting class adjustment, the outcome of interest is considered a fixed covariate while the sample and response indicators are treated as random variables. Let **D** = (D_1, D_2,…,D_N) be the vector of sample inclusion indicators where, for i = 1,…N,

$$D_i = \begin{cases} 1, & \text{unit } i \text{ is included in the sample} \\ 0, & \text{otherwise.} \end{cases}.$$

Assume that D_i is distributed as a Bernoulli random variable with mean π_i, where π_i represents the sample inclusion probability for unit i. Assume that π_i is a function only of design variables such that:

$$\pi_i = P\left(D_i = 1 \middle| \mathbf{y}, \mathbf{X}\right) = P\left(D_i = 1 \middle| \mathbf{X}\right) = \pi\left(\mathbf{X}_i, \mathbf{X}\right).$$

Similarly, assume that the joint inclusion probability for units i and j, where $i \neq j$, also depends only on design variables and not the outcome of interest, such that:

$$\pi_{ij} = P\left(D_i = 1, D_j = 1 \middle| \mathbf{y}, \mathbf{X}\right) = P\left(D_i = 1, D_j = 1 \middle| \mathbf{X}\right) = \pi^{(2)}\left(\mathbf{X}_i, \mathbf{X}_j, \mathbf{X}\right).$$

Let $\mathbf{R} = (R_1, R_2, \ldots, R_n)$ be the indicator of response for unit i given inclusion in the sample where, for $i = 1, \ldots n$,

$$R_i = \begin{cases} 1, & \text{unit } i \text{ responds} \\ 0, & \text{otherwise.} \end{cases}$$

Assume that $R_i \middle| D_i$ is distributed as a Bernoulli random variable with mean p_i, where p_i represents the response probability, or response propensity score, for unit i. The general form of the response propensity, p_i, is discussed initially and then the models for the response propensity discussed in Chapter 3 are applied. Let $m = \sum_{i=1}^{n} R_i$ denote the number of responding units where $m < n$ when nonresponse occurs. Define $\mathbf{y}^o$ as the observed outcome of interest such that $\mathbf{y}^o = \{y_i : D_i = 1, R_i = 1\}$ for $i = 1, \ldots n$. Assume further that the response probability for unit i does not depend on sample inclusion of any other unit other than itself, such that:

$$p_i = P\left(R_i = 1 \middle| \mathbf{y}, \mathbf{X}, \mathbf{D}, D_i = 1\right) = P\left(R_i = 1 \middle| \mathbf{y}, \mathbf{X}, D_i = 1\right) = p\left(y_i, \mathbf{X}_i, \mathbf{y}, \mathbf{X}\right).$$

Define the response probability weight as $\theta_i = \theta_i\left(y_i, \mathbf{X}_i, \mathbf{y}, \mathbf{X}\right) = p_i^{-1}$.

Similarly, assume that the joint response probability for units i and j, where $i \neq j$, is only a function of related covariates as well as the outcome of interest under nonignorable nonresponse given sample inclusion:

$$p_{ij} = P\left(R_i = 1, R_j = 1 \middle| \mathbf{y}, \mathbf{X}, \mathbf{D}, D_i = 1, D_j = 1\right)$$

$$= P\left(R_i = 1, R_j = 1 \middle| \mathbf{y}, \mathbf{X}, D_i = 1, D_j = 1\right)$$

$$= p^{(2)}\left(y_i, \mathbf{X}_i, y_j, \mathbf{X}_j, \mathbf{y}, \mathbf{X}\right).$$

Recall that the covariate matrix $\mathbf{X}_s$ contains all the columns of **X**, the response indicator vector, **R**, and any relevant interactions between **R** and columns of **X**. Similarly, the covariate matrix $\mathbf{X}_r$ contains all the columns of **X**, the outcome vector, **y**, and any relevant interactions between **y** and columns of **X**. Two variations on the design matrix used to model response are also used in response propensity models. $\mathbf{X}_{cr}$ does not include information on $\mathbf{y}$ but incorporates the conditional success propensity score, $\mathbf{p}_s$. Similarly, $\mathbf{X}_{qcr}$ does not include information on **y** but contains quintiles of the conditional success propensity score, $\mathbf{p}_s^*$.

Let $\xi = \xi\left(\beta_s, \beta_r\right)$ denote the superpopulation model for the success and response propensities. Let β_s and β_r represent the vectors of regression coefficients from logistic regression models of success and response propensities, respectively.

Assume that inference is based on a known model for $\boldsymbol{\beta}_s$ and $\boldsymbol{\beta}_r$, the estimated logistic regression coefficients used to estimate success and response propensities, respectively. For the logistic regression model $\text{logit}\left(p_{si}\right) = \mathbf{X}'_{si}\boldsymbol{\beta}_s$ where $\text{logit}\left(p_{si}\right) = \log\left(\frac{p_{si}}{1-p_{si}}\right)$, the success propensity score is calculated as $\hat{p}_{si} = \left[1+\exp\left(-\mathbf{X}'_{si}\boldsymbol{\beta}_s\right)\right]^{-1}$.

5.2.2 Coarsening the response rate

In MAR approaches to weighting adjustments for nonresponse, response rates may be estimated with inverse-probability weighting or by observed response rates within each adjustment class from values in the sample (Gelman and Carlin, 2002). Cassel et al. (1983) use response propensities estimated from logistic regression models incorporating as predictors variables related to the missingness mechanism. Weighting by the estimated response propensity weight obtained from logistic regression may inflate the variance unnecessarily for low estimated response propensities and places more reliance on correct logistic regression model specification (Little and Rubin, 2002). For this application, we examine this approach for NMAR missingness and attempt to alleviate the problems of variance inflation and model dependence by coarsening the response propensity score using the quintiles of the response propensity score.

In this application, let $H = 5$ as recommended by Cochran (1977). When coarsening is used, quintiles of the response propensity score are used to form five adjustment classes. The median of the response propensity scores within a given adjustment class is used to assign an overall response rate for units in each adjustment cell. Using the median has two advantages. First, the median is not influenced by very small propensities that cause variance inflation as are other measures of central tendency, such as the mean. Second, using the either the lower bound or the upper bound of the response propensity scores in each adjustment class cause the population total to be overestimated or underestimated, respectively, so using the median should provide a relatively unbiased estimate of the total.

Define θ_i^* as the quintile of the conditional response propensity weights corresponding to the i^{th} conditional response propensity weight, θ_i^*, for $\mathbf{q} = \left(q_1,...,q_5\right)$. In notation, this quantity is defined for each approach as:

$$\theta_i^* = \mathrm{Q}\left(\theta,\mathbf{q}\right)_i = \left\{Q\left(\theta,\frac{h-0.5}{H}\right): Q\left(\theta,q_{h-1}\right) < \theta_i \le Q\left(\theta,q_h\right)\right\}$$

as the quantile corresponding to the i^{th} conditional response weight. This value is used as the estimate of the response rate for nonresponse adjustment in approaches that employ coarsening of the response propensity scores. Because quintiles are used for coarsening, **q** = {0.1, 0.3, 0.5, 0.7, 0.9}.

Similarly, let p_{si}^{*} be defined similarly as the quintile corresponding to the i^{th} conditional success propensity score, p_{si}. In the QCR approach, these values are used as predictors in the conditional response propensity model. The conditional success propensity score is preferred over the conditional success propensity weight for estimating the response propensity because the success propensity score is bounded in (0,1). This boundedness performs better in predictive logistic regression response modeling than the success propensity weight which falls in the ranges (1, +∞). The response propensity scores are not used as predictors in the models examined in this chapter.

Maximum likelihood estimates (MLEs) of logistic regression coefficients are biased to order n^{-1} (Maiti and Pradhan, 2007). To remove this bias, Firth (1993) maximizes the penalized likelihood function with the invariant Jeffreys prior. These maximum penalized likelihood estimates (MPLEs) are asymptotically equivalent to MLEs as sample sizes increase. To ensure that the propensity scores are unbiasedly estimated, the Firth (1993) approach is employed. The first-order asymptotic equivalence of MLEs and MPLEs allows the assumption of asymptotic normality for MPLEs (Bull, Mak, and Greenwood, 2002). Therefore, the assumption of approximate unbiasedness for $\hat{p}_{si}^{*}$ may be made. Model selection follows the process outlined in Chapter 3.

Standard MLEs are used during the model selection process so that information criteria are appropriately implemented, but final estimates are calculated using the Firth (1993) method.

After the success and response models are selected, the quintiles of the conditional response propensity weights may be used to extrapolate outcomes from responding units to account for nonresponding units. Using the sample quantiles of the conditional response propensity weights, the Horvitz-Thompson estimator from equation (3.2) may be used to obtain an unbiased estimator of the population total. The propensity score adjustment estimator and its variance are derived and then applied to the case study from Chapter 4 to assess its performance when missingness is NMAR. Simulations are used in Chapter 6 to examine the estimator under a broader range of conditions.

5.2.3 The propensity score adjustment estimator

The goal of this work is to unbiasedly and precisely estimate of the population total, $\tau = \sum_{i=1}^{N} y_i$ for $i = 1, \ldots, N$. The modified Horvitz-Thompson estimator of the population total under nonresponse from equation (3.2) is adapted to incorporate the conditional response propensity scores under nonignorable missingness. The general form of the propensity score adjustment estimator for the population total, τ, is:

$$\hat{T} = \sum_{i=1}^{N} \frac{D_i R_i \hat{\theta}_i y_i}{\pi_i} = \sum_{i=1}^{m} \frac{\hat{\theta}_i y_i}{\pi_i}, \tag{5.1}$$

where $\hat{\theta}_i$ is the general form of the estimated response propensity weight.

To derive formulas for $E\left(\hat{T}\middle|\mathbf{y},\mathbf{X}\right)$ and $Var\left(\hat{T}\middle|\mathbf{y},\mathbf{X}\right)$, the following lemma is proposed:

Lemma. Suppose $S = \sum_{i=1}^{N} a_i B_i U_i$ where a_i is a constant, B_i is distributed as a Bernoulli random variable with mean p_i, U_i is a random variable jointly distributed with B_i, and $P\left(B_i = 1, B_j = 1\right) = p_{ij}$. Then

(a) $E\left(S\right) = \sum_{i=1}^{N} a_i p_i E\left(U_i \middle| B_i = 1\right)$,

(b) $E\left(S^2\right) = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j p_i p_j E\left(U_i U_j \middle| B_i = 1, B_j = 1\right)$,

(c) $Var\left(S\right) = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j \left[p_{ij} E\left(U_i U_j \middle| B_i = 1, B_j = 1\right) - p_i p_j E\left(U_i \middle| B_i = 1\right) E\left(U_j \middle| B_j = 1\right)\right]$

(d) $Var\left(S\right) = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j p_{ij} Cov\left(U_i, U_j \middle| B_i = 1, B_j = 1\right)$

$$+\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j p_{ij} E\left(U_i \middle| B_i = 1, B_j = 1\right) E\left(U_j \middle| B_i = 1, B_j = 1\right)$$

$$-\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j p_i p_j E\left(U_i \middle| B_i = 1\right) E\left(U_j \middle| B_j = 1\right)$$

Proof.

(a) $E(S) = E\left[\sum_{i=1}^{N} a_i B_i U_i\right] = \sum_{i=1}^{N} a_i E(B_i U_i)$

Because U_i is a function of the Bernoulli random variable, B_i, we can express the expectation of their product as:

$$E(B_i U_i) = E\left[E(B_i U_i)\middle|B_i\right] = E\left[B_i E\left(U_i \middle| B_i\right)\right]$$

$$= 1 \times E\left(U_i \middle| B_i = 1\right) P(B_i = 1) + 0 \times E\left(U_i \middle| B_i = 0\right) P(B_i = 0)$$

$$= E\left(U_i \middle| B_i = 1\right) p_i$$

Therefore,

$$E(S) = \sum_{i=1}^{N} a_i p_i E\left(U_i \middle| B_i = 1\right).$$

(b) $E\left(S^2\right) = E\left(\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j B_i B_j U_i U_j\right) = \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j E\left(B_i B_j U_i U_j\right)$

$$E\left(B_iB_jU_iU_j\right)=E\left[E\left(B_iB_jU_iU_j\,\middle|B_i,B_j\right)\right]=E\left[B_iB_jE\left(U_iU_j\,\middle|B_i,B_j\right)\right]$$

Because the expectation is non-zero only when $B_i = B_j = 1$, the expectation simplifies to:

$$E\left(B_iB_jU_iU_j\right)=1\times E\left(U_iU_j\,\middle|B_i=1,B_j=1\right)P\left(B_i=1,B_j=1\right)$$

$$=p_{ij}E\left(U_iU_j\,\middle|B_i=1,B_j=1\right)$$

Therefore,

$$E\left(S^2\right)=\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jE\left(B_iB_jU_iU_j\right)=\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jp_{ij}E\left(U_iU_j\,\middle|B_i=1,B_j=1\right).$$

(c) The result in (c) is obtained directly from the proofs for parts (a) and (b) as:

$$Var\left(S\right)=E\left(S^2\right)-\left[E\left(S\right)\right]^2$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_j\left[p_{ij}E\left(U_iU_j\,\middle|B_i=1,B_j=1\right)-p_ip_jE\left(U_i\,\middle|B_i=1\right)E\left(U_j\,\middle|B_j=1\right)\right].$$

(d) To obtain the proof for (d), use the result from part (c) and the fact that:

$$E\left(U_iU_j \middle| B_i=1,B_j=1\right)=Cov\left(U_i,U_j \middle| B_i=1,B_j=1\right)$$

$$+E\left(U_i \middle| B_i=1,B_j=1\right)E\left(U_j \middle| B_i=1,B_j=1\right).$$

Now we have that:

$$Var(S)=\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_j\left[p_{ij}E\left(U_iU_j \middle| B_i=1,B_j=1\right)-p_ip_jE\left(U_i \middle| B_i=1\right)E\left(U_j \middle| B_j=1\right)\right]$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jp_{ij}E\left(U_iU_j \middle| B_i=1,B_j=1\right)-\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jp_ip_jE\left(U_i \middle| B_i=1\right)E\left(U_j \middle| B_j=1\right)$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jp_{ij}Cov\left(U_i,U_j \middle| B_i=1,B_j=1\right)$$

$$+\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jp_{ij}E\left(U_i \middle| B_i=1,B_j=1\right)E\left(U_j \middle| B_i=1,B_j=1\right)$$

$$-\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jp_ip_jE\left(U_i \middle| B_i=1\right)E\left(U_j \middle| B_j=1\right)$$

□

The lemma is used to obtain forms of the bias and variance of the propensity score adjustment estimator. Given the form of the propensity score adjustment estimator, the bias of the estimator is defined as $Bias\left(\hat{T}\right) = E\left(\hat{T} \middle| \mathbf{y}, \mathbf{X}\right) - \tau$, where τ is the population total. Express the propensity score adjustment estimator as:

$$\hat{T} = \sum_{i=1}^{N} \frac{D_i R_i \hat{\theta}_i y_i}{\pi_i} = \sum_{i=1}^{N} \left(\pi_i^{-1} y_i\right)\left(D_i R_i\right)\hat{\theta}_i = \sum_{i=1}^{N} a_i B_i U_i \, ,$$

where $a_i = \pi_i^{-1} y_i$ is a constant, $B_i = D_i R_i$, and $U_i = \hat{\theta}_i$ is a random variable jointly distributed with D_i and R_i. The random variable, $B_i = D_i R_i$, has a Bernoulli distribution with mean $\pi_i p_i$ because

$$\begin{aligned} P\left(D_i R_i = 1 \middle| \mathbf{y}, \mathbf{X}\right) &= P\left(D_i = 1, R_i = 1 \middle| \mathbf{y}, \mathbf{X}\right) \\ &= P\left(D_i = 1 \middle| \mathbf{y}, \mathbf{X}\right) P\left(R_i = 1 \middle| \mathbf{y}, \mathbf{X}, D_i = 1\right) \\ &= \pi_i p_i . \end{aligned}$$

By part (a) of the lemma, the expectation of the propensity score adjustment estimator is:

$$E\left(\hat{T}\,|\mathbf{y},\mathbf{X}\right)=E\left[\sum_{i=1}^{N}\left(\pi_i^{-1}y_i\right)\left(D_iR_i\right)\hat{\theta}_i\,|\mathbf{y},\mathbf{X}\right]=\sum_{i=1}^{N}\left(\pi_i^{-1}y_i\right)E\left[\left(D_iR_i\right)\hat{\theta}_i\,|\mathbf{y},\mathbf{X}\right]$$

$$=\sum_{i=1}^{N}\left(\pi_i^{-1}y_i\right)\pi_i p_i E\left(\hat{\theta}_i\,|\mathbf{y},\mathbf{X},D_i=1,R_i=1\right)$$

$$=\sum_{i=1}^{N}y_i p_i E\left(\hat{\theta}_i\,|\mathbf{y},\mathbf{X},D_i=1,R_i=1\right).$$

Now the bias of the propensity score adjustment estimator may be expressed as:

$$Bias\left(\hat{T}\right)=E\left(\hat{T}\,|\mathbf{y},\mathbf{X}\right)-\tau$$

$$=\sum_{i=1}^{N}y_i p_i E\left(\hat{\theta}_i\,|\mathbf{y},\mathbf{X},D_i=1,R_i=1\right)-\sum_{i=1}^{N}y_i$$

$$=\sum_{i=1}^{N}y_i\left[p_i E\left(\hat{\theta}_i\,|\mathbf{y},\mathbf{X},D_i=1,R_i=1\right)-1\right].$$

To calculate the bias of the propensity score adjustment estimator, the value of the expectation $E\left(\hat{\theta}_i\,|\mathbf{y},\mathbf{X},D_i=1,R_i=1\right)$ must be calculated at least approximately. For nonignorable nonresponse, a nonrespondent subsample (either from the current year or a previous year if the underlying populations are similar) allows estimation of p_i and therefore $\hat{\theta}_i=\hat{p}_i^{-1}$. Note that, in this case, conditioning on **y** implies conditioning on the set of observed outcomes and outcomes obtained in the nonrespondent subsample, $\left(\mathbf{y}^{o}, \mathbf{y}^{(b)}\right)$.

Let $\mathbf{X}_{(i)}$ denote the covariate matrix $\mathbf{X}$ omitting the vector $\mathbf{X}_i$. Define $\mathbf{D}_{(i)}$ and $\mathbf{R}_{(i)}$ similarly. Consider the estimator $\hat{\theta}_i = \hat{\theta}\left(\mathbf{X}_i, D_i, R_i, \mathbf{y}^o, \mathbf{X}_{(i)}, \mathbf{D}_{(i)}, \mathbf{R}_{(i)}\right)$. Then

$$E\left(\hat{\theta}_i \left| \mathbf{y}, \mathbf{X}, D_i = 1, R_i = 1\right.\right) = E\left[\hat{\theta}\left(\mathbf{X}_i, D_i, R_i, \mathbf{y}^o, \mathbf{X}_{(i)}, \mathbf{D}_{(i)}, \mathbf{R}_{(i)}\right) \left| \mathbf{y}, \mathbf{X}, D_i = 1, R_i = 1\right.\right]$$

$$= E\left[\hat{\theta}\left(\mathbf{X}_i, D_i = 1, R_i = 1, \mathbf{y}^o, \mathbf{X}_{(i)}, \mathbf{D}_{(i)}, \mathbf{R}_{(i)}\right) \left| \mathbf{y}, \mathbf{X}, D_i = 1, R_i = 1\right.\right]$$

Applying the method of moments, the expectation is obtained as:

$$E\left(\hat{\theta}_i \left| \mathbf{y}, \mathbf{X}, D_i = 1, R_i = 1\right.\right) = \hat{\theta}\left(\mathbf{X}_i, D_i = 1, R_i = 1, \mathbf{y}^o, \mathbf{X}_{(i)}, \mathbf{D}_{(i)}, \mathbf{R}_{(i)}\right).$$

Note that this value is not the same as the observed $\hat{\theta}_i$ when $R_i = 0$.

For example, if p_i were estimated with logistic regression, then the nonignorable response propensity weight is obtained with the following model:

$$\hat{\theta}_{ri} = \hat{\theta}\left(\mathbf{X}_i, D_i, R_i, \mathbf{y}^o, \mathbf{X}_{(i)}, \mathbf{D}_{(i)}, \mathbf{R}_{(i)}\right) = \hat{p}_{ri}^{-1}$$

$$= \left[\hat{P}\left(R_i = 1 \middle| \mathbf{y}, \mathbf{X}, D_i = 1\right)\right]^{-1}$$

$$= 1 + \exp\left[-\left(a + by_i + \mathbf{c}'\mathbf{X}_i\right)\right],$$

where a, b, and $\mathbf{c}$ are calculated by including the nonrespondent subsample in the logistic regression data set. Alternatively, the expectation of y_i from the superpopulation model could be used as a predictor in the response propensity model. Let the notation $\left[\hat{E}\left(y_i \middle| \mathbf{y}^{\mathrm{o}}, \mathbf{X}, \mathbf{D}, \mathbf{R}\right)\right]\left(R_i, \mathbf{X}_i\right)$ indicate that the estimated conditional expectation of a binary outcome, y_i, is a function of R_i and $\mathbf{X}_i$, and define this quantity as:

$$\left[\hat{E}\left(y_i \middle| \mathbf{y}^{\mathrm{o}}, \mathbf{X}, \mathbf{D}, \mathbf{R}\right)\right]\left(R_i, \mathbf{X}_i\right) = \left\{1 + \exp\left[-\left(f + gR_i + \mathbf{h}'\mathbf{X}_i\right)\right]\right\}^{-1}.$$

$\left[\hat{E}\left(y_i \middle| \mathbf{y}^{\mathrm{o}}, \mathbf{X}, \mathbf{D}, \mathbf{R}\right)\right]\left(R_i, \mathbf{X}_i\right)$ estimates the conditional expectation of y_i, where f, g, and $\mathbf{h}$ are calculated from the nonrespondent subsample. When a nonrespondent subsample is not available for the current survey, the expectation may be estimated from a predictive mean model for the outcome of interest based on the previous sample and nonrespondent subsample. Then the conditional response propensity weight may be estimated as:

$$\hat{\theta}_{cri} = \hat{\theta}\left(\mathbf{X}_i, D_i, R_i, \mathbf{y}^o, \mathbf{X}_{(i)}, \mathbf{D}_{(i)}, \mathbf{R}_{(i)}\right)$$

$$= 1 + \exp\left[-\left(a + b\left[\hat{E}\left(y_i \left| \mathbf{y}^{\mathrm{o}}, \mathbf{X}, \mathbf{D}, \mathbf{R}\right.\right)\right]\left(R_i, \mathbf{X}_i\right) + \mathbf{c}'\mathbf{X}_i\right)\right].$$

Note that, in this case, the coefficients a, b, and **c** do not take the same values as those in the definition of the conditional response propensity weight, $\hat{\theta}_{ri}$.

The general estimator $\hat{\theta}_i = \hat{\theta}\left(\mathbf{X}_i, D_i, R_i, \mathbf{y}^o, \mathbf{X}_{(i)}, \mathbf{D}_{(i)}, \mathbf{R}_{(i)}\right)$ approximates $E\left(\hat{\theta}_i \left| \mathbf{y}, \mathbf{X}, D_i = 1, R_i = 1\right.\right)$, so the resulting bias estimate is also approximate:

$$Bias\left(\hat{T}\right) \approx \sum_{i=1}^{N} y_i \left[p_i E\left(\hat{\theta}_i \left| \mathbf{y}, \mathbf{X}, D_i = 1, R_i = 1\right.\right) - 1\right] = \sum_{i=1}^{N} y_i \left(p_i \hat{\theta}_i - 1\right),$$

where $\hat{\theta}_i$ in this case is the theoretical value assuming both sample inclusion and response and not necessarily the observed value of $\hat{\theta}_i$. Therefore, for each method of obtaining $\hat{\theta}_i$, the bias is a function of the ratio of $\hat{\theta}_i$ to the true response propensity score.

5.2.4 The variance of the propensity score adjustment estimator

The variance of the propensity score adjustment estimator is obtained with part (c) of the lemma such that:

$$Var\left(\hat{T}\middle|\mathbf{y},\mathbf{X}\right)=\sum_{i=1}^{N}\sum_{j=1}^{N}\pi_i^{-1}y_i\pi_j^{-1}y_j\pi_{ij}p_{ij}E\left(\hat{\theta}_i\hat{\theta}_j\middle|D_i=1,R_i=1,D_j=1,R_j=1\right)$$

$$-\sum_{i=1}^{N}\sum_{j=1}^{N}\pi_i^{-1}y_i\pi_j^{-1}y_j\pi_i\pi_jp_ip_jE\left(\hat{\theta}_i\middle|D_i=1,R_i=1\right)E\left(\hat{\theta}_j\middle|D_j=1,R_j=1\right)$$

Assume again that $\hat{\theta}_i=\hat{\theta}\left(\mathbf{X}_i,D_i,R_i,\mathbf{y}^o,\mathbf{X}_{(i)},\mathbf{D}_{(i)},\mathbf{R}_{(i)}\right)$ approximates $E\left(\hat{\theta}_i\middle|\mathbf{y},\mathbf{X},D_i=1,R_i=1\right)$ well. There is no clear approach for approximating the expectation $E\left(\hat{\theta}_i\hat{\theta}_j\middle|\mathbf{y},\mathbf{X},D_i=1,R_i=1,D_j=1,R_j=1\right)$, so the quantity is approximated in this application by $\hat{\theta}_i\hat{\theta}_j$. Now the form of the variance simplifies to:

$$Var\left(\hat{T}\middle|\mathbf{y},\mathbf{X}\right)=\sum_{i=1}^{N}\sum_{j=1}^{N}\pi_i^{-1}y_i\pi_j^{-1}y_j\left(\pi_{ij}p_{ij}\hat{\theta}_i\hat{\theta}_j-\pi_i\pi_jp_ip_j\hat{\theta}_i\hat{\theta}_j\right)$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N}\pi_i^{-1}y_i\pi_j^{-1}y_j\hat{\theta}_i\hat{\theta}_j\left(\pi_{ij}p_{ij}-\pi_i\pi_jp_ip_j\right)$$

$$=\sum_{i=1}^{N}\frac{\hat{\theta}_i^2p_iy_i}{\pi_i}\left(1-\pi_ip_i\right)+\sum_{i=1}^{N}\sum_{j\neq i}\frac{\hat{\theta}_i\hat{\theta}_jy_iy_j}{\pi_i\pi_j}\left(p_{ij}\pi_{ij}-p_ip_j\pi_i\pi_j\right) \qquad (5.2)$$

The response mechanism is further assumed to be independent among all units, *i.e.* $R_i \perp R_j$ for all $i \neq j$. This assumption implies that $p_{ij} = p_i p_j$. This further simplifies the variance of the propensity score adjustment estimator (equation 5.2) of the population total given in equation (5.1) to the following form:

$$Var\left(\hat{T}\right) = \sum_{i=1}^{N} \frac{\hat{\theta}_i^2 p_i y_i}{\pi_i}\left(1-\pi_i p_i\right) + \sum_{i=1}^{N}\sum_{j\neq i} \frac{p_i p_j \hat{\theta}_i \hat{\theta}_j y_i y_j}{\pi_i \pi_j}\left(\pi_{ij} - \pi_i \pi_j\right) \tag{5.3}$$

5.2.5 Estimator of $Var\left(\hat{T}\right)$

To obtain an unbiased estimate of the variance of the propensity score adjustment estimator of the total give in equation (5.1), the sample responses are weighted to account for sampling and nonresponse error. The following lemma provides an estimator for the variance of the propensity score adjustment estimator of the total.

Lemma: An unbiased estimator for the variance of the propensity score adjustment estimator is given by:

$$\hat{Var}\left(\hat{T}\right) = \sum_{i=1}^{m} \frac{\hat{\theta}_i^2 y_i}{\pi_i^2}\left(1-\pi_i p_i\right) + \sum_{i=1}^{m}\sum_{j\neq i} \hat{\theta}_i \hat{\theta}_j y_i y_j \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}}\right) \tag{5.4}$$

Proof: Stating the variance estimator in terms of the random vectors, **D** and **R**, we have:

$$\hat{Var}\left(\hat{T}\right)=\sum_{i=1}^{m}\frac{\hat{\theta}_i^2 y_i}{\pi_i^2}\left(1-\pi_i p_i\right)+\sum_{i=1}^{m}\sum_{j\neq i}\hat{\theta}_i\hat{\theta}_j y_i y_j\left(\frac{1}{\pi_i\pi_j}-\frac{1}{\pi_{ij}}\right)$$

$$=\sum_{i=1}^{N}\frac{D_i R_i\hat{\theta}_i^2 y_i}{\pi_i^2}\left(1-\pi_i p_i\right)+\sum_{i=1}^{N}\sum_{j\neq i}D_i D_j R_i R_j\hat{\theta}_i\hat{\theta}_j y_i y_j\left(\frac{1}{\pi_i\pi_j}-\frac{1}{\pi_{ij}}\right)$$

Assuming that the response mechanism is independent among units, *i.e.*

$P\left(R_i=1,R_j=1\right)=p_{ij}=p_i p_j$, the expectation of the variance estimator is computed as follows:

$$E\left[\hat{Var}\left(\hat{T}\right)\right]=E_D\left\{E_R\left[\hat{Var}\left(\hat{T}\right)\right]\right\}$$

$$=E_D\left\{E_R\left[\sum_{i=1}^{N}\frac{D_i R_i\hat{\theta}_i^2 y_i}{\pi_i^2}\left(1-\pi_i p_i\right)+\sum_{i=1}^{N}\sum_{j\neq i}D_i D_j R_i R_j\hat{\theta}_i\hat{\theta}_j y_i y_j\left(\frac{1}{\pi_i\pi_j}-\frac{1}{\pi_{ij}}\right)\right]\right\}$$

$$=E_D\left[\sum_{i=1}^{N}\frac{D_i E_R\left(R_i\hat{\theta}_i^2\right)y_i}{\pi_i^2}\left(1-\pi_i p_i\right)+\sum_{i=1}^{N}\sum_{j\neq i}D_i D_j E_R\left(R_i R_j\hat{\theta}_i\hat{\theta}_j\right)y_i y_j\left(\frac{1}{\pi_i\pi_j}-\frac{1}{\pi_{ij}}\right)\right]$$

$$=E_D\left[\sum_{i=1}^{N}\frac{D_i p_i\hat{\theta}_i^2 y_i}{\pi_i^2}\left(1-\pi_i p_i\right)+\sum_{i=1}^{N}\sum_{j\neq i}D_i D_j p_i p_j\hat{\theta}_i\hat{\theta}_j y_i y_j\left(\frac{1}{\pi_i\pi_j}-\frac{1}{\pi_{ij}}\right)\right]$$

$$= \sum_{i=1}^{N} \frac{\pi_i p_i \hat{\theta}_i^2 y_i}{\pi_i^2} (1 - \pi_i p_i) + \sum_{i=1}^{N} \sum_{j \neq i} \pi_{ij} p_i p_j \hat{\theta}_i \hat{\theta}_j y_i y_j \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}} \right)$$

$$= \sum_{i=1}^{N} \frac{p_i \hat{\theta}_i^2 y_i}{\pi_i} (1 - \pi_i p_i) + \sum_{i=1}^{N} \sum_{j \neq i} \frac{p_i p_j \hat{\theta}_i \hat{\theta}_j y_i y_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j)$$

which is equivalent to (5.4), proving that $\hat{Var}(\hat{T})$ is unbiased for the true variance of the propensity score adjustment estimator of the total given in equation (5.1). □

5.2.6 Response propensity approaches

In Chapter 3, propensity score methodology was extended to the case of nonignorable missingness. In Chapter 4, MAR and NMAR approaches to adjustment class formation for weighting class adjustment were developed based on the proposed NMAR propensity score methodology. In the current chapter, these approaches are used to obtain response propensities that directly adjust the respondent outcomes in a Horvitz-Thompson estimator that is adjusted for nonresponse (equation 3.2).

In Chapter 3, three estimators of the response propensity under nonignorable nonresponse were proposed: the NMAR response propensity score $(p_{ri} = p_r(Y_i, \boldsymbol{X}_i))$, the conditional response propensity score $(p_{cri} = p_{cr}(p_{si}, \boldsymbol{X}_i))$, and the conditional response propensity score from the quintiles of the conditional success propensity score$\left(p_{qcri} = p_{qcr}\left(p_{si}^{*}, \boldsymbol{X}_i\right)\right)$. The balancing properties of the three response propensity score estimators relative to the response mechanism were shown for a binary outcome of interest. These three response propensity models are used to estimate the response propensity score for the adjusted Horvitz-Thompson estimator (equation 3.2) in approaches referred to, respectively, as *NMAR response propensity modeling* (indexed by "R"), *conditional response propensity modeling* (indexed by "CR"), and *conditional response propensity modeling based on quintiles of the conditional success propensity score* (indexed by "QCR"). A MAR approach obtains response propensity scores from a logistic regression of the covariates, **X**, on the response indicator with no additional predictors for the outcome of interest. These four models (R, CR, QCR, and MAR) are used to obtain response propensity score for use in equation 3.2. Additionally for each model, quintiles of the response propensity scores are used to form adjustment classes, and the median score within each class is used as the estimate of the response rate for nonresponse adjustment in equation 3.2. This provides four additional models, denoted as R*, CR*, QCR*, and MAR* to emphasize that quintiles are used, for nonresponse adjustment with the propensity score adjustment estimator. Notation for the response propensity weights, design matrices, and regression coefficients, as well as some information on model assumptions, is provided in Table 5.1 for the eight response propensity models.

Several of the same response propensity modeling approaches from Chapter 4 are used in this chapter to estimate response propensity scores. Specifically, the CR and QCR approaches employ the exact models used in Chapter 3. However, the R and MAR approaches were based on a subset of covariates deemed related more closely to the response mechanism ($\mathbf{X}_r$) because joint classification with success propensity scores might cause collinearity issues (David et al., 1983). Since this is not an issue with the propensity score adjustment estimator, the complete set of covariates ($\mathbf{X}$) is used to model response in the R and MAR approaches described in this chapter.

Table 5.1: Models examined for the propensity score adjustment estimator

PSAE Approach	**Missingness mechanism**	**Response propensity scores, p**	**Response propensity weights, $\theta = p^{-1}$**	**Design matrix, X**	**Covariates for modeling response, R**	**p/θ a function of R?**	β	**Estimator of the total**
1	MAR	$\mathbf{p}_{rMAR}$	$\boldsymbol{\theta}_{rMAR}$	$\mathbf{X}$	$\mathbf{X}$	No	β_{rMAR}	$\hat{T}_{MAR}$
2	MAR	$\mathbf{p}^{*}_{rMAR}$	$\boldsymbol{\theta}^{*}_{rMAR}$	$\mathbf{X}$	$\mathbf{X}$	No	β_{rMAR}	$\hat{T}_{MAR*}$
3	NMAR	$\mathbf{p}_{r}$	$\boldsymbol{\theta}_{r}$	$\mathbf{X}_{r}$	$\mathbf{Y}, \mathbf{X}$	No	β_{r}	$\hat{T}_{r}$
4	NMAR	$\mathbf{p}^{*}_{r}$	$\boldsymbol{\theta}^{*}_{r}$	$\mathbf{X}_{r}$	$\mathbf{Y}, \mathbf{X}$	No	β_{r}	$\hat{T}_{r*}$
5	NMAR	$\mathbf{p}_{cr}$	$\boldsymbol{\theta}_{cr}$	$\mathbf{X}_{cr}$	$\mathbf{p}_{s}(\mathbf{R},\mathbf{X}), \mathbf{X}$	Yes	β_{cr}	$\hat{T}_{cr}$
6	NMAR	$\mathbf{p}^{*}_{cr}$	$\boldsymbol{\theta}^{*}_{cr}$	$\mathbf{X}_{cr}$	$\mathbf{p}_{s}(\mathbf{R},\mathbf{X}), \mathbf{X}$	Yes	β_{cr}	$\hat{T}_{cr*}$
7	NMAR	$\mathbf{p}_{qcr}$	$\boldsymbol{\theta}_{qcr}$	$\mathbf{X}_{qcr}$	$\mathbf{p}^{*}_{s}(\mathbf{R},\mathbf{X}), \mathbf{X}$	Yes	β_{qcr}	$\hat{T}_{qcr}$
8	NMAR	$\mathbf{p}^{*}_{qcr}$	$\boldsymbol{\theta}^{*}_{qcr}$	$\mathbf{X}_{qcr}$	$\mathbf{p}^{*}_{s}(\mathbf{R},\mathbf{X}), \mathbf{X}$	Yes	β_{qcr}	$\hat{T}_{qcr*}$

5.3 CASE STUDY

The propensity score adjustment estimator from equation (5.1) and variance estimator from equation (5.4) are applied to the NMDGF elk hunter survey introduced in Chapter 4. This approach incorporates propensity scores from various MAR and NMAR models to form adjustment classes. The estimates of elk harvest from the propensity score adjustment estimator are compared to the weighting class adjustment estimates obtained in Chapter 4. Covariates selected in the model selection procedure detailed in Chapter 4 are provided for each model in Appendix D.

Quintiles are obtained using L-estimators, which are weighted sums of order statistics used to estimate quantiles (Harrell and Davis, 1982; Kaigh and Lachenbruch, 1982; Sfakianakis and Verginis, 2008; Stromberg, 1997). An L-estimator of a vector $\boldsymbol{X}$ has the following form:

$$\mathrm{Q}_{\mathrm{p}}\left(\boldsymbol{X}\right)=\sum_{i=1}^{n} w_i X_{(i)},$$

where w_i is the weight and $X_{(i)}$ is the i^{th} order statistic. The response propensities exhibited considerable skewness. Harrell and Davis (1982) estimate quantiles with weighted sums of order statistics and a jackknife procedure to estimate the variance. Other L-estimators for quantiles have been developed (Kaigh and Lachenbruch 1982), but the Harrell-Davis estimator was found to be superior in performance (Stromberg 1997).

Recent work by Sfakianakis and Verginis (2008) provides three new quantile estimators with weights calculated from binomial probabilities. These estimators outperform the Harrell-Davis quantile estimator for asymmetric distributions and extreme quantiles. The "SV3" estimator has the least bias and smallest MSE for asymmetric distributions of their three proposed estimators. The SV3 estimator is biased (Sfakianakis and Verginis, 2008) but is the least biased of all L-estimators examined for two skewed distributions. Absolute bias of the SV3 estimator for lognormally-distributed random variables with a mean of 0 and a standard deviation of 1 are positively biased for upper quantiles, and lower quantiles are slightly underestimated (Verginis, personal communication). However, absolute relative bias for a range of quantiles falls below 0.2% when sample sizes are large (>750). The SV3 estimator is also examined for the exponential distribution with mean 1. For large samples, absolute relative bias of the 20th percentile is highest at 2.7%, with bias for all other quantile estimates falling below 1%. Therefore, the bias of the SV3 estimator is relatively small and is assumed to contribute minimally to the MSE. The SV3 L-estimator is used to obtain quintiles of the response and success propensity scores assuming that the bias in the quantile estimator is negligible for large samples.

5.3.1 Results

The results of the propensity score adjustment estimator analysis for the New Mexico Department of Game and Fish elk hunter survey and nonrespondent subsample are provided in Table 5.2. For survey years 2001 and 2003, the estimate of the total elk harvest, the 95%-confidence interval, RMSE, and relative bias are provided. Confidence interval coverage, root mean square error (RMSE), and relative bias are evaluated to assess the performance of the estimators. Relative bias is measured relative to the design-based estimator obtained from double-sampling for stratification. The unadjusted estimates calculated under the MCAR assumption are also compared to the estimates from each NMAR estimator. The predictive capability of the 2001 success model is assessed with the 2003 data to determine if the models derived from the nonrespondent subsamples can be applied to data not augmented by a nonrespondent subsample.

Table 5.2: 2001 and 2003 propensity score adjustment estimates (PSAE) of total harvest

Approach	Metric	2001	2003
R	Est. Total	4706	4269
	95%-CI	(4651, 4760)	(4216, 4322)
	RMSE	5818	7403
	Rel. Bias	-0.55	-0.63
R*	Est. Total	4681	4311
	95%-CI	(4631, 4731)	(4259, 4364)
	RMSE	5842	7360
	Rel. Bias	-0.56	-0.63
CR	Est. Total	11209	12692
	95%-CI	(10886, 11532)	(12264, 13120)
	RMSE	706	1043
	Rel. Bias	0.07	0.09
CR*	Est. Total	11333	13165
	95%-CI	(11022, 11644)	(12743, 13586)
	RMSE	825	1502
	Rel. Bias	0.08	0.13
QCR	Est. Total	11818	13343
	95%-CI	(11487, 12149)	(12911, 13776)
	RMSE	1306	1688
	Rel. Bias	0.12	0.14
QCR*	Est. Total	11772	13517
	95%-CI	(11449, 12096)	(13090, 13943)
	RMSE	1260	1862
	Rel. Bias	0.12	0.16
MAR	Est. Total	5035	4646
	95%-CI	(4960, 5110)	(4571, 4720)
	RMSE	5488	7027
	Rel. Bias	-0.52	-0.60
MAR*	Est. Total	4905	4575
	95%-CI	(4842, 4968)	(4508, 4643)
	RMSE	5618	7097
	Rel. Bias	-0.53	-0.61

Approach	Metric	2001	2003
DSS	Est. Total	10523	11672
	95%-CI	(10299, 10747)	(11339, 12005)
	RMSE	114	170
	Rel. Bias	0.00	0.00
Unadjusted	Est. Total	14190	15446
	95%-CI	(13849, 14531)	(15056, 15835)
	RMSE	3671	3779
	Rel. Bias	0.35	0.32

The R and R* approaches generate estimates of the total elk harvest that are very low, with relative bias values indicating that harvest is underestimated by at least 50% for both approaches and both years as compared to the double sampling for stratification estimates. This result is surprising given that the response propensity model accurately reflects the NMAR missingness mechanism by modeling the response indicator as a function of the outcome of interest and a large suite of covariates using the subsample of nonrespondents. The estimates of total harvest from the R and R* approaches are only slightly higher than the reported harvest (see Table 4.3), indicating that the estimated NMAR response propensity weights are very close to one for respondents.

The CR approach generates estimates of total elk harvest with much higher accuracy. The 2001 CR estimate of the total exhibits relative bias of 0.07, and the 2003 estimate demonstrates a higher positive relative bias of 0.09. This approach performs better than expected since it was shown both in Chapter 3 and Appendix B that the conditional response propensity score mathematically reduces to the response indicator. The quintiles of the conditional response propensity scores (CR*) did not change these results substantially, with relative biases of 0.08 for 2001 and 0.13 for 2003.

The approaches that employ the quintiles of the conditional success propensity scores as predictors in the response model (QCR and QCR*) are just slightly more biased than the approaches that use the uncoarsened conditional success propensity scores as predictors (CR and CR*). The QCR approach provides estimates of total elk harvest with relative bias of 0.12 for 2001 and 0.14 for 2003. The QCR* approach yields similar estimates with relative bias of 0.12 for 2001 and 0.16 for 2003. Using quintiles of the conditional success propensity scores as predictors in the response model increases the RMSE as compared to the CR and CR* approaches. The RMSE for the CR estimates are 706 in 2001 and 1043 in 2003. When the quintiles of the conditional success propensity score are used (QCR approach), the RMSE increases to 1306 in 2001 and 1688 in 2003. Similar results are found when comparing the approaches with coarsened response propensity scores. The RMSE values from the CR* approach are 825 in 2001 and 1502 in 2003. Coarsening the conditional success propensity score (QCR*) increases the RMSE to 1260 in 2001 and 1862 in 2003. This increase in RMSE results from increases in both variance and bias.

The MAR approaches generate estimates of total elk harvest that are very similar to those obtained from the R approach. This approach underestimates the harvest total by more than half despite incorporating information from the nonrespondent subsample. Confidence interval width is relatively narrow for these estimates, indicating that the inflated RMSE values are due to the large negative bias of the estimates.

Coarsening was used in the propensity score adjustment estimator in two ways. First, estimates of the response rate were obtained by using either the estimated response propensities or the values obtained from quintiles of the response propensity scores. This level of coarsening was used to alleviate the variance-inflating effects of small response propensities. Second, in the QCR and QCR* approaches, quintiles of the success propensity score were used as predictors in the response propensity model. This level of coarsening was used to reduce dependence on the success propensity model. With respect to the first level of coarsening, the use of quintiles of response propensity scores rather than the response propensity scores did not greatly affect inference. Coarsening the response propensity scores increases bias slightly in this example and exhibit no obvious benefit in this case. To evaluate the second level of coarsening, the CR approach is compared to the QCR approach, and the CR* approach is compared to the QCR* approach. In each case, modeling response with coarsened success propensity scores increased bias and RMSE. This result suggests that coarsening the success predictor may not be prudent. These two levels of coarsening will be examined further for a range of scenarios in the simulations described in Chapter 6.

Overall, the R, R*, MAR, and MAR* approaches perform poorly by greatly underestimating the elk harvest total as compared to the design-based estimates obtained from double sampling for stratification. An examination of the estimated response propensity scores from each approach revealed that the estimated response propensity scores are generally much closer to one for the R and MAR approaches compared to the CR and QCR approaches (Figure 5.1). The methods also generate extremely low propensities, generally for nonrespondents within certain combinations of weapon type, hunt size, and landowner type. This tendency of the estimated response propensity scores toward 0 or 1 is referred to as *separation* and can be a problem in discrete-data regression, especially when binary predictors that exhibit collinearity are used (Gelman, Jakulin, Pittau, and Su, 2001). In this case, linear combinations of certain model variables more perfectly predict the response indicator.

Weakly informative priors may help this problem, but the Jeffrey's prior (Firth, 1993) incorporated in this approach did not resolve the issue. The simplest correction to this problem may be the use of the CR or QCR approaches which employ a real-valued estimate of the success propensity rather than the binary outcome of interest to obtain more stable models and unbiased estimates. The CR approach was expected to exhibit separation as evidenced by the theoretical simplification of $p_{cri} = p_{cr}\left(p_{si}\left(R_i = r\right), \mathbf{X}_i\right) = r$. However, the CR approach worked well in practice. This favorable performance may be a result of large samples and conditional response models consisting of many predictors and interactions with the conditional success propensity score so that simplification to the response indicator is not possible. These considerations will discussed further for the simulations reviewed in Chapter 6.

Overall, the CR approach for the propensity score adjustment estimator yielded estimates with the least bias and RMSE. However, none of the confidence intervals for the propensity score adjustment estimator approaches covered the design-based estimates. Furthermore, the degree of bias is slightly higher for the 2003 estimates than for the 2001 estimates. As discussed in Chapter 4, these issues may result from unanticipated bias in the design-based estimates from double sampling for stratification or poor performance of the 2001 success propensity model for the 2003 data. We will examine these issues further in the simulations discussed in Chapter 6.

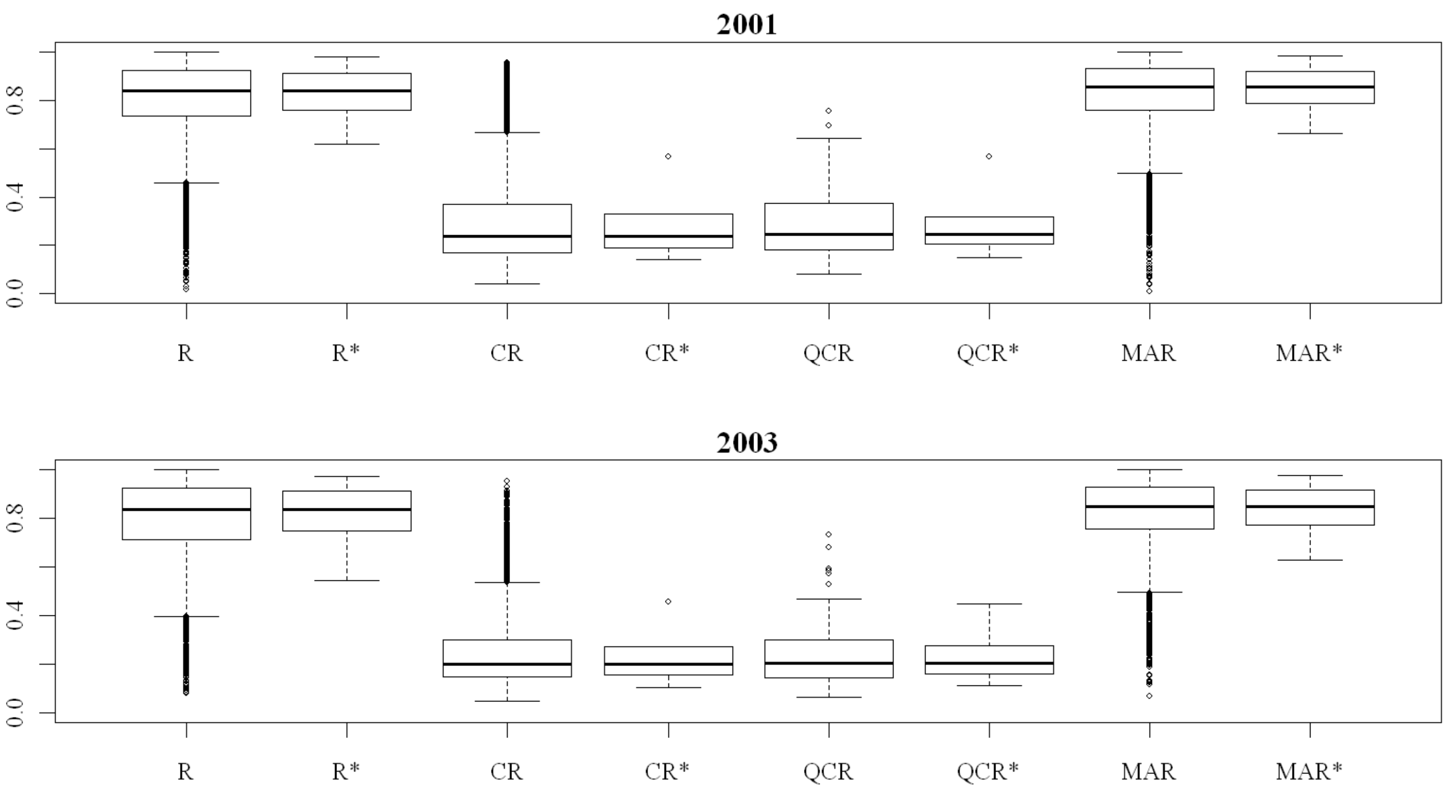

Figure 5.1: Boxplots of estimated response propensities for the eight propensity score adjustment estimator approaches

5.3.2 Comparison to weighting class adjustment

The results from the weighting class adjustment analysis in Chapter 4 (Table 4.8) are provided again in Table 5.3 for ease of discussion. In the previous section, the poor performance of the R and R* approaches for the propensity score adjustment estimator was discussed. The NMAR response propensity score perform better as the basis for adjustment cells in a weighting class adjustment even when subject to separation (Table 5.3). The weighting class adjustment estimates of the total for the R approach exhibited relative bias of -0.10 for 2001 and 0.05 for 2003, and when joint classification was used with the conditional success propensity score (JR approach), the relative bias was -0.11 and -0.09 for 2001 and 2003, respectively. The R and JR approaches to the weighting class adjustment exhibited the least absolute relative bias for the 2003 pilot data as compared to all other approaches. This result indicates that, when separation is a problem, the NMAR response propensity scores are not accurate enough for direct nonresponse adjustment but are more effective for grouping units with similar response rates. However, the relative bias for the R weighting class adjustment approach is -0.10 for 2001, and the relative bias for the R* weighting class adjustment approach is -0.11 for 2001 and -0.09 for 2003. These are the only weighting class adjustment approaches that underestimate the total relative to the double sampling for stratification estimator, indicating that the NMAR propensity scores may also perform poorly for weighting class adjustment when separation occurs.

Because the propensity score adjustment estimator greatly underestimates the population total when the R, R*, MAR, and MAR* approaches are used, the CR and QCR approaches are compared between the weighting class adjustment and the propensity score adjustment estimators. In most cases, the propensity score adjustment estimates exhibit smaller RMSE than the weighting class adjustments based on the same propensity scores. For example, the 2001 relative bias for the CR and CR* propensity score adjustment estimator approaches is 706 and 825, respectively. For the 2001 weighting class adjustment estimates, the RMSE is higher for both the CR approach (1290) and the JCR approach (854). Similarly, the 2003 relative bias for the CR and CR* propensity score adjustment estimator approaches is 1043 and 1502, respectively. For the 2003 weighting class adjustment estimates, the CR approach (1399) and the JCR approach (1894) both produce higher RMSE values. An exception occurs with the JQCR estimator for 2001 from the weighting class adjustment. In this case, the RMSE of 987 is lower than the RMSE from the propensity score adjustment estimator for the QCR approach (1306) and the QCR* approach (1260). Across both estimators, the CR approaches are less biased than the QCR approaches, indicating that the coarsening of the success propensity score as a predictor in the response propensity model may not be helpful.

Table 5.3: 2001 and 2003 weighting class adjustment estimates

Estimator	Metric	2001	2003
WC Adj. (Joint classification)	Est. Total	13888	15404
	95%-CI	(13618, 14158)	(15058, 15751)
	RMSE	3368	3736
	Rel. Bias	0.32	0.32
WC Adj. (Instrumental variable regression)	Est. Total	13805	15593
	95%-CI	(13537, 14074)	(15247, 15939)
	RMSE	3285	3925
	Rel. Bias	0.31	0.34
WC Adj. (Intermediate)	Est. Total	11409	13525
	95%-CI	(11010, 11809)	(12944, 14107)
	RMSE	909	1877
	Rel. Bias	0.08	0.16
WC Adj. (R)	Est. Total	9460	12207
	95%-CI	(9234, 9686)	(11912, 12502)
	RMSE	1069	556
	Rel. Bias	-0.10	0.05
WC Adj. (CR)	Est. Total	11796	13058
	95%-CI	(11388, 12203)	(12687, 13430)
	RMSE	1290	1399
	Rel. Bias	0.12	0.12
WC Adj. (QCR)	Est. Total	12284	13557
	95%-CI	(11924, 12643)	(13188, 13927)
	RMSE	1771	1894
	Rel. Bias	0.17	0.16
WC Adj. (JR)	Est. Total	9356	10656
	95%-CI	(8421, 10292)	(10147,11165)
	RMSE	1261	1049
	Rel. Bias	-0.11	-0.09
WC Adj. (JCR)	Est. Total	11327	13409
	95%-CI	(10760, 11894)	(12796, 14021)
	RMSE	854	1765
	Rel. Bias	0.08	0.15
WC Adj. (JQCR)	Est. Total	11479	13742
	95%-CI	(10999, 11959)	(13076, 14408)
	RMSE	987	2098
	Rel. Bias	0.09	0.18
WC Adj. (SUCC)	Est. Total	11109	13128
	95%-CI	(10462, 11757)	(12573, 13684)
	RMSE	673	1483
	Rel. Bias	0.06	0.12

Estimator	Metric	2001	2003
Double sampling for stratification	Est. Total 95%-CI RMSE Rel. Bias	10523 (10299, 10747) 114 0.00	11672 (11339, 12005) 170 0.00
Unadjusted estimates	Est. Total 95%-CI RMSE Rel. Bias	14190 (13849, 14531) 3671 0.35	15446 (15056, 15835) 3779 0.32

5.3.3 Comparison to additional NMDGF hunt information

Additional information may help in assessing the performance of the propensity score adjustment estimator and the NMAR weighting class adjustment estimators in the case study. In 2006, NMDGF discontinued the censusing of licensees and changed to a mandatory survey return program. Licensees that do not respond via telephone or internet are now fined and considered ineligible for any big game hunts in the state during the following hunt year. Success rates are calculated as the estimated harvest divided by the number of licensees for each year. NMDGF success rates for 2001, 2003, and 2006 through 2008 are provided for comparison in Table 5.4.

Table 5.4: Estimated success rates from all approaches and sources

Estimate	2001	2003	2006	2007	2008
Response rate	0.29	0.24	0.83	0.86	0.86
WC Adj. (JC)	0.36	0.38			
WC Adj. (IVR)	0.36	0.38			
WC Adj. (INT)	0.30	0.33			
WC Adj. (R)	0.25	0.30			
WC Adj. (CR)	0.31	0.32			
WC Adj. (QCR)	0.32	0.33			
WC Adj. (JR)	0.24	0.26			
WC Adj. (JCR)	0.30	0.33			
WC Adj. (JQCR)	0.30	0.34			
WC Adj. (SUCC)	0.29	0.32			
PSAE (R)	0.12	0.11			
PSAE (R*)	0.12	0.11			
PSAE (CR)	0.29	0.31			
PSAE (CR*)	0.30	0.33			
PSAE (QCR)	0.31	0.33			
PSAE (QCR*)	0.31	0.33			
PSAE (MAR)	0.13	0.11			
PSAE (MAR*)	0.13	0.11			
MCAR	0.37	0.38	0.28*	0.32*	0.33
DSS	0.28	0.29			
Min. success rate	0.11	0.09	0.23	0.28	0.29
Max. success rate	0.81	0.85	0.40	0.42	0.43

* Estimates from a subpopulation of hunts in core occupied elk range

Response rates in 2006, 2007, and 2008 increased to 83%, 86%, and 86%, respectively, from the low response rates observed in 2001 and 2003 before the mandatory survey return program began. Propensity score adjustment, weighting class adjustment, and MCAR estimates of the success rate are provided for comparison of the 2001 and 2003 estimates to those obtained from a similar survey with a higher response rate. Furthermore, the minimum and maximum success rates are provided as "realistic intervals" on the true success rate. These bounds are calculated assuming that either all nonrespondents are unsuccessful or all nonrespondents are successful and provide a realistic interval of the success rate (Molenberghs, Kenward, Verbeke, Beunckens, and Sotto, 2009). It is important to note that the estimates for 2006 and 2007 represent only hunts in core occupied elk range rather than all hunts statewide. MCAR success rate estimates based only on the survey returns were lower in 2006 through 2008 (0.28, 0.32, and 0.33, respectively) than in 2001 and 2003 (0.37 and 0.38, respectively). This result may indicate either a decrease in overall success during that period of time or sampling that better captures the subpopulation of licensees that would not respond to the original survey without deleterious consequences. Information from the 2001 and 2003 nonrespondent subsamples suggest that the latter is plausible.

To compare the propensity score adjustment and weight class adjustment estimates directly to the MCAR estimates, one must be willing to make several assumptions. First, one must assume that the annual success rates have not changed significantly over the period from 2001 to 2008. Also, the success rate in core occupied elk range must not be substantially different from the statewide success rate to make comparisons. Furthermore, one must assume that the remaining 14% to 17% nonresponse in the mandatory return survey is missing completely at random with response to the binary outcome for success. Under these assumptions, we can compare the propensity score adjustment and weight class adjustment estimates to the MCAR estimates, which are now taken as "true" values of the success rate.

If these assumptions are valid, then the success rate is generally about 0.28 to 0.33 and ranges from 0.23 to 0.43 based on the realistic intervals from 2006 through 2008. The success rate estimates from double sampling for stratification of 0.28 in 2001 and 0.29 in 2003 fall within this interval. The R, R*, MAR, and MAR* approaches of the propensity score adjustment estimator greatly underestimate the success rate with respect to the mandatory return estimates. The CR, CR*, QCR, and QCR* approaches for the propensity score adjustment estimator yield success rates that fall within the range of mean success for the 2006, 2007, and 2008 estimates. For the weighting class adjustment, the INT, CR, QCR, JCR, JQCR, and SUCC approaches provide estimates of the success rate that fall between 0.28 and 0.33, the range of the estimated success rates from 2006 though 2008. However, all of the estimates of the success rate from the weighting class adjustment fall within the range of realistic interval endpoints (0.23 to 0.43) observed between 2006 and 2008.

If the previous assumptions are not valid, then the more conservative approaches may be more reasonable for wildlife management. Conservative approaches include those that tend to slightly overestimate rather than underestimate total elk success, such as the CR and CR* approaches of the propensity score adjustment estimator and the INT, JCR, and SUCC approaches for the weighting class adjustment. Underestimating harvest could prompt NMDGF to increase license numbers, resulting in increased hunting pressure that may negatively impact the elk population. Therefore, a moderate bias adjustment would be more prudent that an overcorrection that resulted in underestimation of the total harvest and an unsustainable increase in elk licenses.

5.4 DISCUSSION

In this chapter, the propensity score adjustment estimator is introduced for nonignorable nonresponse. This estimator uses NMAR extensions of propensity score methodology to estimate response propensities used in a modified Horvitz-Thompson estimator for nonresponse (equation 3.2). These techniques require obtaining a nonrespondent subsample. Four modeling approaches were examined, three of which are appropriate for NMAR missingness. For each response propensity modeling approach, either the estimated response propensity scores or quintiles of the estimated response propensity scores were used to estimate the response rate in equation 3.2.

In the case study example, the propensity score adjustment estimator was applied to two years of data from an elk hunter survey. For each year, a nonrespondent subsample was obtained and design-based estimates of total harvest were calculated. Modeling the response propensity scores from the binary indicator of success performed surprisingly poorly due to collinearity issues from a large suite of binary predictors. Because response propensity models for the CR, CR*, QCR, and QCR* approaches incorporate real-valued estimated success propensity scores or quintiles of estimated success propensity scores, estimated response propensity scores were less susceptible to bias from separation (Gelman et al., 2001). Overall, the CR approach was found to provide the least bias and smallest RMSE across all propensity score adjustment estimate approaches, indicating that coarsening of the response propensity scores or success propensity scores provides no benefit for these data sets.

Estimation with the propensity score adjustment estimator is dependent on large samples, data sets with rich covariate information, and the ability to subsample nonrespondents. Therefore, this approach may be most useful to state and federal resource management agencies with large list frames containing numerous explanatory variables. Accurate and complete databases are needed to appropriately identify nonrespondents and a substantial subsample of nonrespondents must be obtained to accurately model the success outcome. The cost of the NMDGF nonrespondent subsample exceeded $10,000 each year, but the additional data collection and analysis allows the severity of the bias to be assessed and adjusted if necessary. If factors influencing success do not change greatly over time, then the success model from one nonrespondent subsample could be used for multiple years for a savings in survey costs and improved accuracy in harvest estimates. Bias may be reduced for only the cost of data analysis.

5.5 CONCLUSIONS

The propensity score adjustment estimator, a novel extension of the estimator proposed by Cassel et al. (1983), may be a useful tool for nonignorable nonresponse adjustment. Accounting for the dependence between the response indicator and the outcome of interest in the response propensity model allows propensity score adjustment to be used when data are NMAR. When a nonrespondent subsample is available, methods incorporating the conditional success propensity score rather than the outcome of interest may be more robust and unbiased. Coarsening of either the response propensity score for nonresponse adjustment or the success propensity score for response propensity prediction did not demonstrate any benefits in the analysis of the pilot data sets. Simulations are employed to determine how the propensity score adjustment estimator approaches perform under various conditions, including a range of odds ratios, success rates, and response rates. Changes in success and response rates between the population used for modeling and the population to which the success model is applied are generated to determine how approaches perform and if coarsening is effective.

6: SIMULATION STUDY

6.1 INTRODUCTION

When data are NMAR, a subsample of nonrespondents may be used to develop propensity score methodology to deal with nonignorable nonresponse. In Chapter 4, several success and response propensity score models were used to form adjustment classes for estimation with the weighting class adjustment (Oh and Scheuren, 1983). These models reflected a range of assumptions of missingness. For example, MAR approaches to propensity score estimation were obtained from Vartivarian and Little (2003) and David et al. (1983). An intermediate approach between MAR and NMAR missingness was developed by incorporating the success propensity score as a predictor of response but modeling response independently of success. Several new NMAR approaches to adjustment class formation were also proposed.

NMAR models that account for the dependence of the response on success were also developed from the nonrespondent subsample information. NMAR response propensity classification (indexed by "R") is based on a response propensity score that includes the outcome of interest for both respondents and nonrespondents. A nonrespondent subsample is necessary for this estimator. Conditional response propensity classification (indexed by "CR") is based on a response propensity model which includes estimates of the predicted mean of the outcome of interest; when the outcome of interest is a binary response, the predicted mean may also be called the success propensity score. When the model for success accurately predicts success for respondents and nonrespondents, the predictions are used to model the propensity to respond under nonignorable missingness when a nonrespondent subsample is not available. A similar NMAR modeling approach, the conditional response propensity classification, is based on quintiles of the conditional success propensity score (indexed by "QCR"). In this approach, quintiles of the conditional success propensity scores are used as predictors of response propensity rather than the conditional success propensity scores. In conditional success propensity classification (SUCC), the conditional success propensity score is used alone to form adjustment cells and represents predictive mean stratification (Little, 1986) for NMAR missingness. Additionally, classes formed from each of the response propensity scores are cross-classified with those formed by the conditional success propensity score in the spirit of joint classification discussed by Vartivarian and Little (2003) and are referred to as joint NMAR response propensity classification (JR), joint conditional response propensity classification (JCR), and joint conditional response propensity classification based on quintiles of the conditional success propensity score (JQCR).

The adjusted Horvitz-Thompson estimator of the population total under nonresponse (equation 3.2) incorporates response propensity estimates from each of the MAR and NMAR response models. In Chapter 5, this estimator is called the *propensity score* *adjustment estimator*, or PSAE, and is given in equation (5.1). Either the estimated response propensity scores or quintiles of the response propensity scores were used as the estimate of the response rate for each responding sampling unit.

A simulation is conducted to examine the performance of the NMAR propensity score methodology for weighting class adjustments and the propensity score adjustment estimator discussed in Chapters 4 and 5, respectively. A variety of success and response propensities, degrees of model specification, and odds ratios were considered to determine how well each estimator performs under a range of conditions. The role of each factor is discussed in the subsequent sections. Performance is measured with relative bias, confidence interval coverage, and root mean square error (RMSE).

6.2 SIMULATION METHODS

6.2.1 Description of the simulation populations

First, the success and response covariates are obtained. For both the outcome of interest and the response indicator, independent sets of covariates including a four-level factor and a Bernoulli random variable are generated as predictors. These variables reflect the factors and binary covariates used for modeling success and response in the NMDGF case study. A multinomial random variable is generated with four levels occurring with probabilities 0.4, 0.3, 0.2, and 0.1. These probabilities were chosen so that they sum to 1 and roughly reflect the frequencies observed among weapon types for the NMDGF pilot data. Therefore, in this example, roughly 40% of the licensees in the simulation are licensed for the combined hunt category consisting of center-fire, muzzle-loader, or bow hunts (referred to as "rifle hunts" because rifles are most commonly used), 30% are licensed for bow-only hunts, 20% are licensed for muzzle-loader only hunts, and 10% are licensed in hunts for impaired hunters. Three binary random variables are obtained from the first three levels of the multinomial random variable by creating indicator variables for the first three levels. The fourth level is treated as a reference factor. In other words, let X_{s0i} represent the multinomial random variable influencing success such that X_{s0i} is distributed as a $\text{Multinomial}(0.4, 0.3, 0.2, 0.1)$ random variable for all *i*. Three binary random variables are defined as:

$$X_{s1i} = \begin{cases} 1, \text{ unit } i \text{ in rifle hunt} \\ 0, \text{ otherwise} \end{cases}$$

$$X_{s2i} = \begin{cases} 1, \text{ unit } i \text{ in bow-only hunt} \\ 0, \text{ otherwise} \end{cases}$$

$$X_{s3i} = \begin{cases} 1, \text{ unit } i \text{ in muzzle-loader only hunt} \\ 0, \text{ otherwise} \end{cases}$$

A separate Bernoulli random variable, X_{s4i}, with an arbitrarily-chosen mean of 0.5 was independently generated.as a predictor of the binary outcome of interest. This dichotomous covariate mirrors predictors from the NMDGF pilot data such as landowner type or residency. Therefore, the set of success model predictors for unit *i* includes X_{s1i}, X_{s2i}, X_{s3i}, and X_{s4i}. Response covariates X_{r1i}, X_{r2i}, X_{r3i}, and X_{r4i} were generated similarly as the success covariates.

The true success propensity is obtained from a logistic regression model, $\text{logit}(\mathbf{p}_s) = \mathbf{X}_s \boldsymbol{\beta}_s$, where $\mathbf{X}_s = (\mathbf{X}_{s1}, \mathbf{X}_{s2}, \mathbf{X}_{s3}, \mathbf{X}_{s4})$. The values of $\boldsymbol{\beta}_s$ were selected to generate mean success propensities of 0.3, 0.6, and 0.9 so that a wide range of success propensities could be examined (Table 6.1). The same approach was also used to compute response propensities, and success and response covariates were obtained from independent processes. Recall from section 4.4.2 that the success rate for the NMDGF pilot data was 0.28 in 2001 and 0.29 in 2003. Recall that the NMDGF response rates for 2001 and 2003 were 0.29 and 0.24, respectively. Therefore, the range of mean success rates $(\bar{p}_s)$ and mean response rates $(\bar{p}_r)$ examined in the simulation treats the observed success and response rates from the pilot data as lower limits.

Table 6.1: Values of regression coefficients for each mean success/response propensity

$\bar{p}_s$ or $\bar{p}_r$	β_s **or** β_r
0.3	1, -1.5, -1.5, -1.5
0.6	1, 1, -1, -1
0.9	1, 1, -1, 1

Given the success and response propensities, the response mechanism (**R**) and Bernoulli outcomes (**Y**) were generated from a multivariate Bernoulli distribution. The correlation between **R** and **Y** was calculated as a function of the odds ratio (OR). The odds ratio is defined as follows:

$$OR = \frac{P\left(R=1 \middle| Y=1\right)P\left(R=0 \middle| Y=0\right)}{P\left(R=1 \middle| Y=0\right)P\left(R=0 \middle| Y=1\right)}$$

Therefore, the odds ratio represents the ratio of the odds of response for successful units to the odds of response for unsuccessful units. When the odds ratio is greater than one, successful units are more likely to respond than unsuccessful units. When the odds ratio is less than one, unsuccessful units are more likely to respond. When the odds ratio is 1, the nonresponse is ignorable (Nandram and Choi, 2000). When the odds ratio is substantially different from one, then the nonresponse is nonignorable and appropriate adjustment techniques must be used. Odds ratios of 0.33, 0.5, 1, 2, and 3 were considered in this exercise to examine a range of bias severity. These odds ratios were chosen to exemplify a plausible range of odds ratios based on results from the case study.

Four types of populations were generated. The first population has correct specification for success and response models. The second population exhibits a 30% decline in the mean response rate. The third population is generated with a 30% decline in the mean success rate. The fourth population demonstrates 30% declines in both the mean success rate and mean response rate. Declines of 30% were used because this decline is large enough to affect inference but small enough to avoid generating very small values of the mean response rate. Declines are generated by changing the regression coefficients in the success and/or response propensity models to produce propensity scores with the desired mean. NMAR approaches in which a nonrespondent subsample is only available for Population 1 and not for Populations 2 through 4 include the CR, CR*, QCR, and QCR* approaches of the propensity score adjustment estimator and the CR, QCR, JCR, and JQCR approaches of the weighting class adjustment. For these approaches, the success model based on Population 1 is used to predict success for samples from Populations 2 through 4. The bias and precision of the estimators indicate how approaches perform when the models are not correctly specified.

To summarize, the factors listed in Table 6.2 are used to generate a variety of populations in the simulation. A total of 180 scenarios are created from all possible combinations of the simulation factors which include 5 levels of the odds ratio, 3 levels of the mean success rate, 3 levels of the mean response rate, 2 levels of change in the mean success rate, and 2 levels of change in the mean response rate. This range of conditions will be used to examine the performance of the estimators from Chapters 4 and 5.

Table 6.2: Simulation factors and levels

Factor	**Levels**
Odds ratio	0.33, 0.5, 1, 2, 3
Mean success rate, $\overline{p}_s$	0.3, 0.6, 0.9
Mean response rate, $\overline{p}_r$	0.3, 0.6, 0.9
Decrease in mean success rate, $\overline{p}_s$	None, 30%
Decrease in mean response rate, $\overline{p}_r$	None, 30%

6.2.2 Simulation steps

For each of the 180 scenarios enumerated by Table 6.2, a total of 100 populations were generated in the simulation. For each population generated as described in section 6.2.1, three simple random samples of size 300 were drawn from each population. We assume that complete outcomes for all nonrespondents in the sample are obtained in a nonrespondent subsample. For each sample, each of the 21 estimators is calculated. Means and variances of the total, relative bias, confidence interval coverage, and root mean square error (RMSE) are summarized across all samples and iterations. Relative bias for each estimator was calculated as follows:

$$\text{Relative bias} = \frac{\text{Estimate of total}}{\text{True total}} - 1 .$$

The true total is known for each simulated population, so relative bias is known rather than estimated. This summary statistic is negative if the estimate underestimates the true total and positive if overestimation occurs.

6.2.3 Estimation methods

The weighting class adjustment estimator of the total (equation 4.1) and the variance estimator (equation 4.6) proposed in Chapter 4 were applied in the simulation. In Chapter 5, three NMAR response propensity models and one MAR response propensity model provide estimates of the response propensities that are used either directly or are used to obtain quintiles of the response propensity weights in the propensity score adjustment estimator (equation 5.1).

A total of 21 estimators are calculated in this simulation study. In addition to the ten weighting class adjustment approaches examined in Chapter 4 and eight propensity score adjustment estimator approaches from Chapter 5, the double-sampling for stratification estimator will be calculated assuming that a nonrespondent subsample is available for each population (equations 4.7 and 4.8). Furthermore, two MCAR estimators are examined. The approach denoted as *MCAR* applies the MCAR assumption that the responding units are a random sample of the original sample (Little and Rubin, 2002). The nonresponse is accounted for by reducing the sample size to that observed (m) and adjusting inclusion probabilities as if a sample of m respondents had been actually drawn. The MCAR estimator using the Horvitz-Thompson estimator for a reduced sample size of m respondents is:

$$\hat{T}_{MCAR} = \sum_{i=1}^{m} \pi_i^{-1} p^{-1} y_i \,,$$

where $p = \frac{m}{n}$ adjusts the inclusion probabilities for the reduced sample size. When the sampling design is a simple random sample, the adjusted inclusion probability simplifies to $\pi_i' = \pi_i p = \frac{n}{N} \times \frac{m}{n} = \frac{m}{N}$, which would be the inclusion probability for a simple random sample of size m from a population of size N. The variance estimator for the Horvitz-Thompson estimate of the total is the standard estimator but incorporates the adjusted weights for the reduced sample of m respondents. This estimator is given by:

$$\hat{Var}\left(\hat{T}_{MCAR}\right)=\sum_{i=1}^{m}\frac{\left(1-\pi_i'\right)}{\left(\pi'\right)^2}y_i^2+\sum_{i=1}^{m}\sum_{j\neq i}\frac{\pi_{ij}'-\pi_i'\pi_j'}{\pi_{ij}'\pi_i'\pi_j'}y_iy_j\ ,$$

where y_i is the outcome of interest for unit *i* and $\pi_{ij}' = p^2\pi_{ij}$ is the adjusted joint inclusion probability for units *i* and *j*.

A second MCAR estimator, *MCAR2*, is used assuming that no adjustments for nonresponse are made. In this approach, the nonresponse is completely ignored and inclusion probabilities are not adjusted for nonresponse. This estimator of the total, referred to as the *MCAR2* estimator, is given by the standard Horvitz-Thompson estimator:

$$\hat{T}_{MCAR2}=\sum_{i=1}^{m}\pi_i^{-1}y_i$$

with the estimator of the variance given by:

$$\hat{Var}\left(\hat{T}_{MCAR2}\right)=\sum_{i=1}^{m}\frac{\left(1-\pi_i\right)}{\pi_i^2}y_i^2+\sum_{i=1}^{m}\sum_{j\neq i}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}y_iy_j\ .$$

6.3 SIMULATION RESULTS

The performance of the 21 estimators is evaluated by examining the mean relative bias, mean confidence interval coverage, and mean root mean square error (RMSE) summarized across simulation iterations (Appendix E). These summary statistics are used to assess the accuracy and precision of the estimators under a range of conditions. General plots are provided subsequent to the discussion of each summary statistic, and more specific discussion may be verified by referring to Appendix E.

6.3.1 Relative bias

Boxplots of the estimates of relative bias for each of the 21 estimators are provided across odds ratios (Figures 6.1 to 6.5). Each box plot represents the range of estimates for each approach across the levels of the odds ratio, mean success rate, and mean response rate. Most approaches tend to underestimate the true population total when the odds ratio is less than one and tend to overestimate the total when the odds ratio is greater than one. As the odds ratio increases or decreased from one, the absolute relative bias increases. When the odds ratio is one, the missingness mechanism is MAR, so bias is generally less severe in this case. Relative bias for the MCAR and MCAR2 estimators is almost identical, indicating that a simple adjustment for nonresponse is inadequate in for bias reduction. The design-based estimator for double-sampling for stratification (DSS) exhibits near-unbiasedness across all scenarios. For most estimators, the relative bias is very low when the mean success rate and the mean response rate are high ($\overline{p}_s = \overline{p}_r = 0.9$).

For Population 1, the MCAR estimators (MCAR and MCAR2) and the MAR approaches for weighting class adjustment (JC and IVR) are most biased (Figure 6.1). However, some of the results from the QCR and QCR* approaches of the propensity score adjustment estimator and the QCR approach of the weighting class adjustment produce estimates are as or more biased than the estimates from the MCAR and MAR approaches. The weighting class adjustments using the INT, R, JR, JCR, and JCQR approaches to adjustment class formation reduce bias as compared to the MCAR approaches over the range of scenarios. The intermediate (INT) approach and the three NMAR joint classification approaches (JR, JCR, and JCQR) all incorporate joint classification on response and success propensities. The favorable performance of the intermediate (INT) approach is surprising considering that the response propensity is based on a MAR model for missingness. When the odds ratio is less than 1, the SUCC approach also provides nearly-unbiased estimation of the population total. The MAR propensity score adjustment estimators (MAR and MAR*) are as or more biased as the estimates from the MCAR approaches for all scenarios except when the mean success rate is high ($\overline{p}_s = 0.9$). The increase in bias from using the MAR and MAR* approaches for the propensity score adjustment estimator is most extreme when the mean response rate is low ($\overline{p}_r = 0.3$) and the mean success rate is low to moderate

($\overline{p}_s \leq 0.6$). The NMAR propensity score adjustment estimators (R and R* approaches) are more negatively biased when the odds ratio is less than 1, the mean response rate is low ($\overline{p}_r = 0.3$), and the mean success rate is low to moderate ($\overline{p}_s \leq 0.6$).

For Population 2, the mean response rate, $\overline{p}_r$, declines by 30% overall producing a mean response rate of about 0.21. For a mean success rate of 0.9, the relative bias decreases to a range of -0.04 to 0.04 for the MCAR approaches (Figure 6.2). The propensity score adjustment estimates tend to underestimate the population total for all odds ratios examined and increase the bias most severely when the odds ratio is less than one or the mean response rate is 0.3. This result suggests that, in these cases, the estimated response propensities are overestimated by the logistic regression model. The weighting class adjustment estimates for the JR adjustment class approach are nearly unbiased except when the odds ratio is 0.33; in this case, the JR approach can underestimate the population total up to 16%. When the odds ratio is less than 1, the SUCC approach of the weighting class adjustment performs well. The INT, JCR, and JQCR approaches also reduce bias due to nonresponse and often perform as well as the JR and SUCC approaches. The estimates from the QCR weighting class adjustment approach were most consistently biased across the scenarios for Population 2 and, in some cases, inflate the bias more severely than either of the MCAR approaches. The MAR propensity score adjustment estimators (MAR and MAR* approaches) increased the severity of the bias as compared to the two MCAR estimators, with relative bias nearly doubling for odds ratios greater than one and underestimating the population total by nearly a half even when the odds ratio is one and missingness is truly MAR.

Population 3 exhibits a 30% decline in mean success rates compared to the population on which the model was obtained. Absolute relative bias is uniformly less than 0.10 when the mean response rate is high ($\overline{p}_r = 0.9$). Relative bias of the estimates from the R and R* approaches of the propensity score adjustment estimator is not affected by the decline in the mean success rate because these estimators incorporate a nonrespondent subsample of Population 3 rather than the conditional success model that is based on Population 1 (Figure 6.3). These estimators are least-biased for most scenarios except when the mean response rate and the mean success rate are low ($\overline{p}_r = \overline{p}_s = 0.3$). In these cases, the INT, JCR, and JQCR approaches of the weighting class adjustment provide less-biased estimates. The propensity score adjustment estimator approaches incorporating the conditional success model rather than the actual outcome of interest (CR, CR*, QCR, QCR*) exhibit relative bias ranging from -0.39 to 0.22 for moderate to high mean success rates ($\overline{p}_s \geq 0.6$) and low to moderate mean response rates ($\overline{p}_r \leq 0.6$). For odds ratios less than one, these propensity score adjustment estimators generate estimates that exhibit more severe bias than those obtained from the two MCAR estimators.

Population 4 experiences 30% declines in both the mean response rate and the mean success rate. In this case, all of the propensity score adjustment estimates demonstrate extreme negative bias that is more severe than that observed in the estimates from the two MCAR approaches (Figure 6.4). When the mean response rate is low ($\overline{p}_r$ = 0.3) and the odds ratio is less than one, the NMAR propensity score adjustment estimator approaches (R, R*, CR, CR*, QCR, and QCR*) underestimate the population total and exhibit relative bias ranging from -0.57 to -0.19. The NMAR weighting class adjustments (INT, R, CR, QCR, JR, JCR, JQCR, and SUCC) perform better than the NMAR propensity score adjustment estimators when the mean response rate is low ($\overline{p}_r$ = 0.3), with the INT, JCR, and JQCR approaches providing the least biased estimates of the population total. Note that these three approaches employ NMAR joint classification which appears to be effective in reducing bias. However, when the mean success rate is high ($\overline{p}_s$ = 0.9), the relative bias for these three approaches falls as low as -0.20 for an odds ratio of 0.33 and is as high as 0.19 when the odds ratio is 3 which is more extreme than the relative bias of MCAR estimates which ranges from -0.18 to 0.15 when the mean success rate is high ($\overline{p}_s$ = 0.9). Relative bias is consistently poor across scenarios for the QCR approach of the weighting class adjustment.

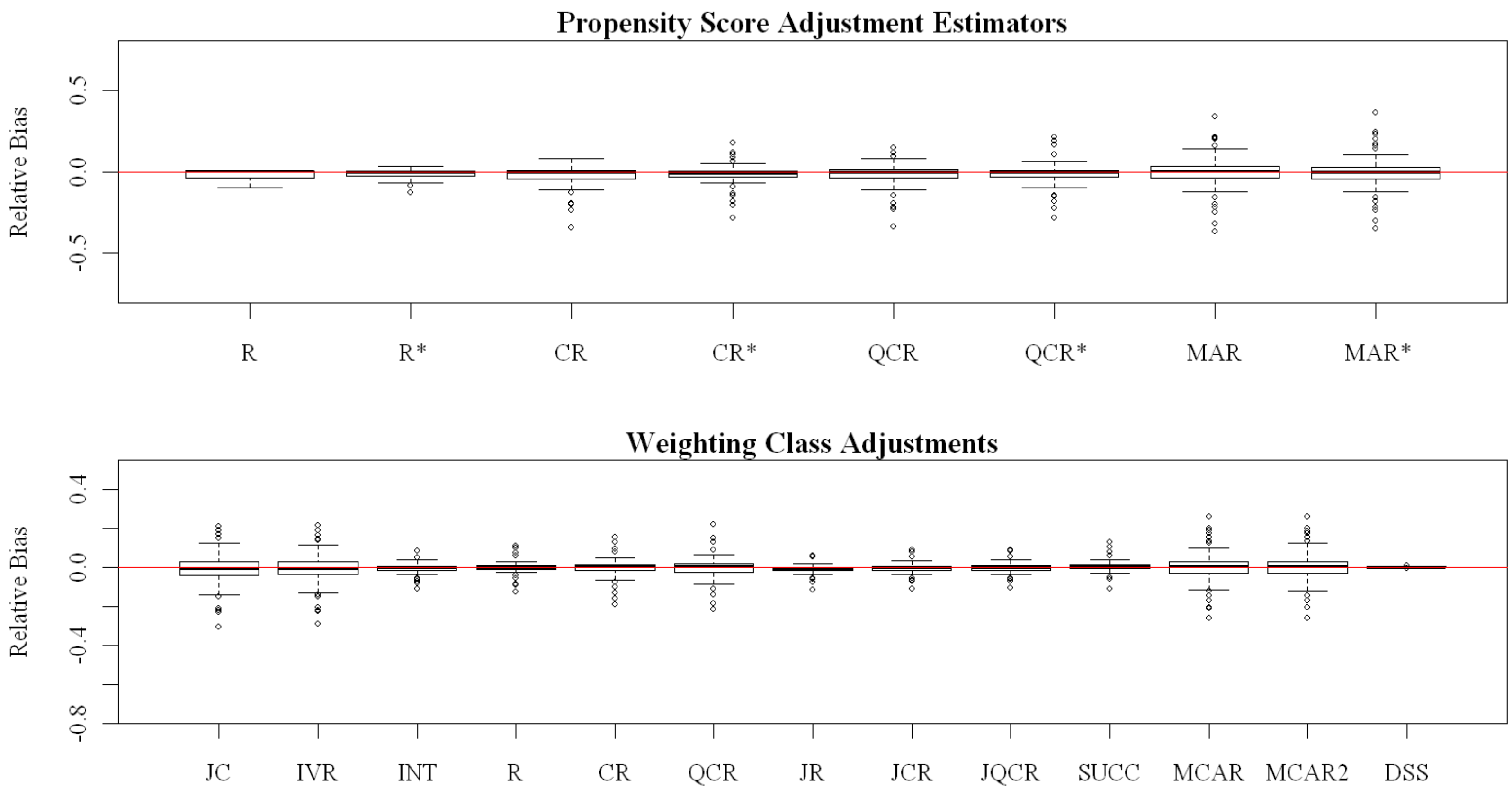

Figure 6.1: Relative bias across all scenarios and odds ratios for the population with no change (Population 1 – correct specification for success and response propensity models)

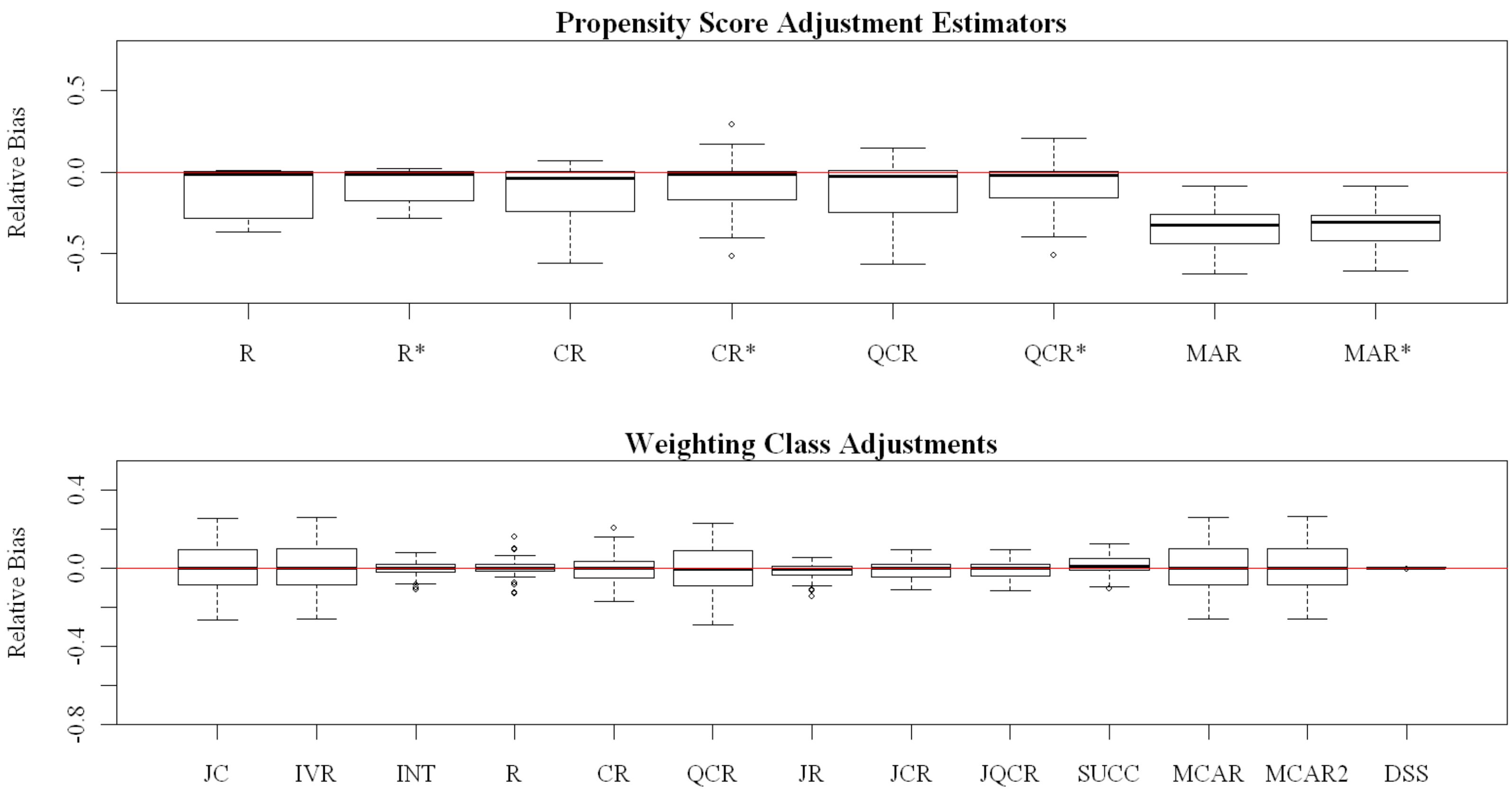

Figure 6.2: Relative bias across all scenarios and odds ratios for the population with a 30% decline in the mean response rate (Population 2 – incorrect specification for response propensity model)

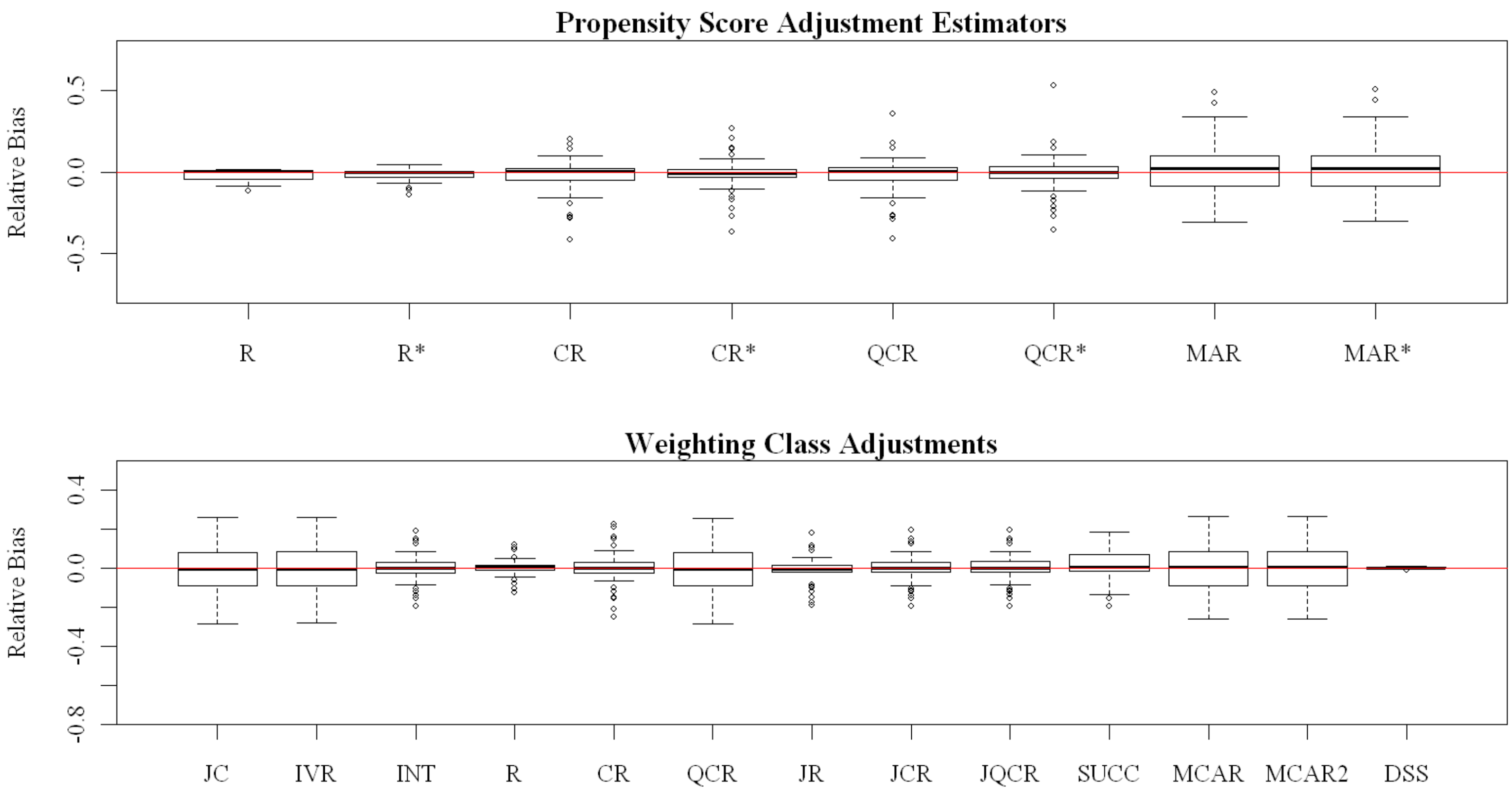

Figure 6.3: Relative bias across all scenarios and odds ratios for the population with a 30% decline in the mean success rate (Population 3 – incorrect specification for success propensity model)

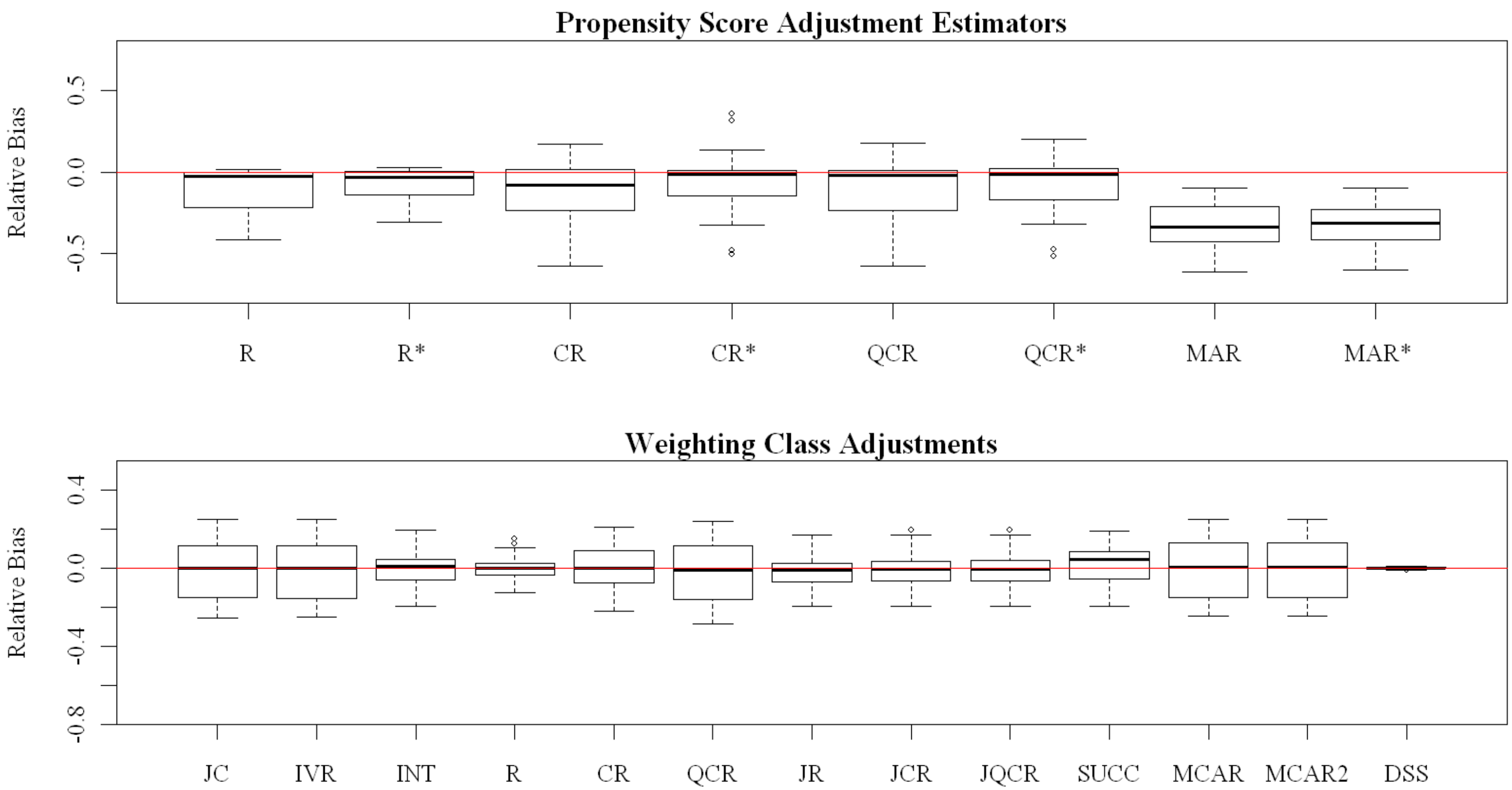

Figure 6.4: Relative bias across all scenarios and odds ratios for the population with 30% declines in the mean success and mean response rates (Population 4 – incorrect specification for success and response propensity models)

6.3.2 Confidence interval coverage

Confidence interval coverage is compared among the 21 estimators. Box plots comparing results are provided in Figures 6.5 through 6.8. Confidence intervals are calculated for a Type I error level of 0.05. Confidence interval coverage is considered nominal for rates of 0.93 to 0.97.

For Population 1, confidence interval coverage achieves nominal coverage rates when the propensity score adjustment estimator is used for odds ratios greater than 1 and the NMAR weighting class adjustments are used for odds ratios less than 1. The poorest confidence interval coverage across odds ratios and levels of mean response and success rates is obtained when the MAR or MCAR methods are used (Figure 6.5). When the mean success rate is high ($\overline{p}_s = 0.9$), confidence intervals from the MAR propensity score adjustment approaches cover the population total 98% to 100% of the time, indicating large variance estimates. The MCAR2 approach provides poor coverage even when the odds ratio is one and missingness is MAR. The CR, CR*, QCR, and QCR* approaches of the propensity score adjustment estimator provide poor coverage when odds ratios are less than one, the mean success rate is moderate ($\overline{p}_s = 0.6$), and the mean response rate is low to moderate ($\overline{p}_r \leq 0.6$). Under these circumstances, the confidence interval coverage ranges from 0.37 to 0.82. Confidence interval coverage for the INT, JCR, JQCR, and SUCC weighting class adjustment approaches demonstrates nominal levels for moderate odds ratios (0.5, 1, and 2).

For Population 2, confidence interval coverage for the population total is uniformly low for all propensity score adjustment estimates that employ the conditional success propensity model (CR, CR*, QCR, QCR*) when the mean response rate is low ($\overline{p}_r$ = 0.3). For this population, the mean response rate decreases 30% from 0.3 to 0.21, and coverage falls below 0.64 for all scenarios. When the mean response rate is moderate to high ($\overline{p}_r \geq 0.6$), the R weighting class adjustment approach more consistently provides produces nominal confidence interval coverage than any other NMAR propensity score adjustment or weighting class adjustment estimator. An exception occurs when the mean response rate is moderate to high ($\overline{p}_r \geq 0.6$), the mean success rate is moderate ($\overline{p}_s = 0.6$), and the odds ratio is 0.33 or 3; in these cases, the confidence interval coverage is below the nominal rate. For $\overline{p}_r = 0.6$, $\overline{p}_s = 0.3$, and an odds ratio greater than 0.33, the CR and CR* propensity score adjustment estimators also achieve nominal coverage. The MAR propensity score adjustment estimates yield extremely low coverage rates overall with less than 10% coverage when odds ratios are less than one. The MCAR weighting class adjustment approach exhibits nominal confidence interval coverage only when the odds ratio is one or when the mean success rate is high, the mean response rate is low, and the odds ratio is less than one. When the mean response rate declines by 30%, nominal coverage rates are most consistently obtained through design-based estimation from double-sampling for stratification.

Estimates from Population 3, which experiences a 30% decline in the mean success rate ($\overline{p}_s$), obtain the most nominal coverage rates for low mean response rates of ($\overline{p}_r$ = 0.3) with the R and R* approaches of the propensity score adjustment estimator. These approaches of the propensity score adjustment estimator are maintain better coverage rates than any of the weighting class adjustment approaches, which exhibit poor coverage as the mean success rate ($\overline{p}_s$) increases from 0.3. However, coverage can fall below 0.90 when the odds ratio is less than one and the mean response rates is low ($\overline{p}_r$ = 0.3). With the exception of the double sampling for stratification estimator, the R approach for the weighting class adjustment provides the best confidence interval coverage when the mean response rate is moderate to high ($\overline{p}_r \geq 0.6$). Given that these approaches all require a nonrespondent subsample for every sample, the double sampling for stratification design-based estimator is more appropriate because it is design unbiased under all scenarios. It is unsurprising that approaches that incorporate the success propensity model do not perform well when the mean success rate declines by 30%.

For Population 4, 30% declines in both the mean response rate and the mean success rate impact the confidence interval coverage most severely. As with Populations 2 and 3, the confidence interval coverage was best overall for the R approach to the weighting class adjustment (Figure 6.8). When the mean response rate is low ($\overline{p}_r$ = 0.3) and the odds ratio is less than 1, the JR approach for the weighting class adjustment gives the better, but less than nominal, confidence interval coverage. Confidence interval coverage for the MAR propensity score adjustment estimators (MAR and MAR*) is uniformly low, and coverage rates for the MAR weighting class adjustment approaches (JC and IVR) are low overall except when the missingness is truly MAR (for an odds ratio of one). When both the mean response rate and the mean success rate change in a population whose surveys exhibit NMAR missingness, confidence interval coverage is best achieved by obtaining a nonrespondent subsample and using design-based estimators from the double sample for stratification.

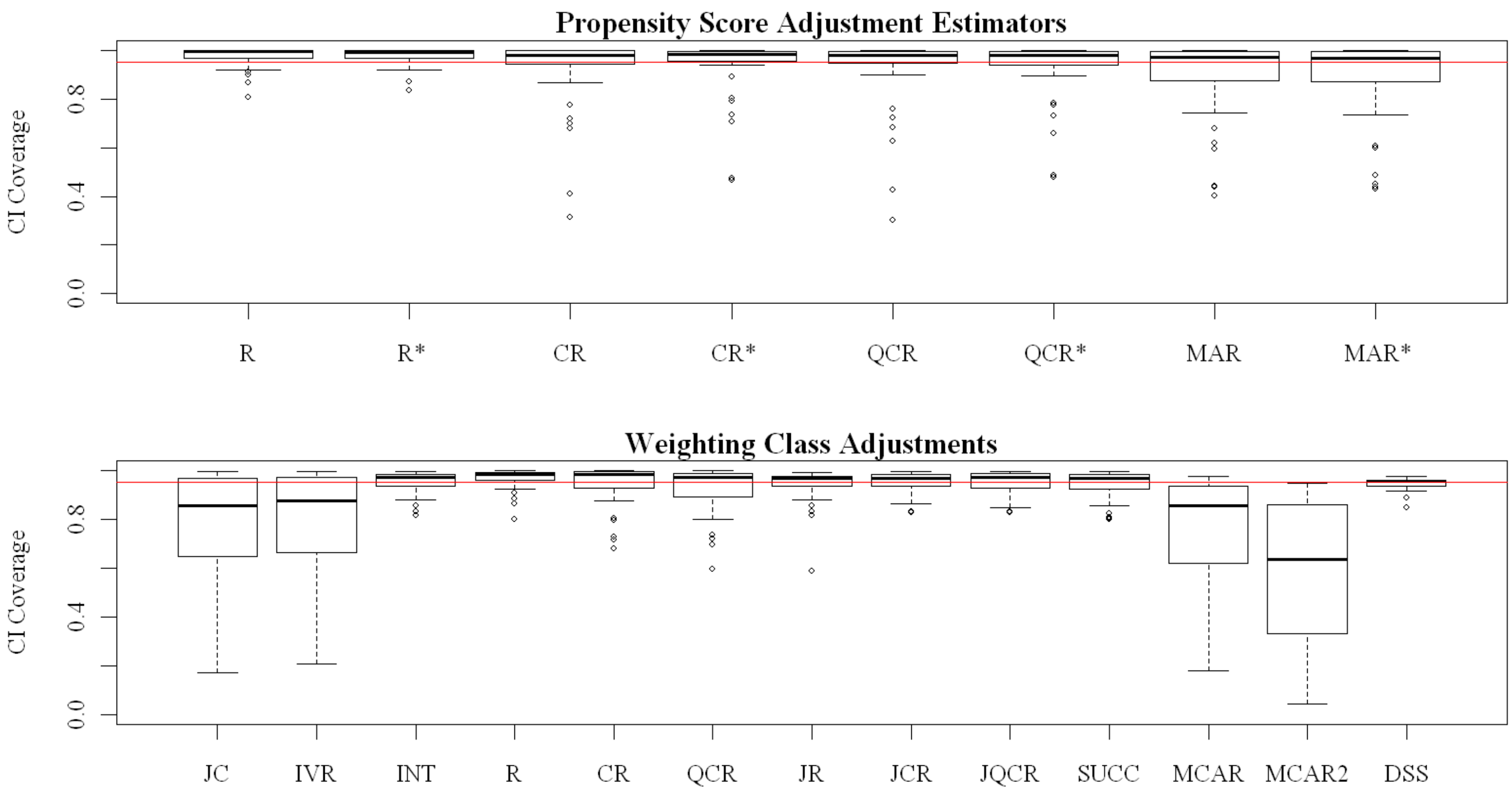

Figure 6.5: Confidence interval coverage across all scenarios and odds ratios for the population with no change (Population 1 – correct specification for success and response propensity models)

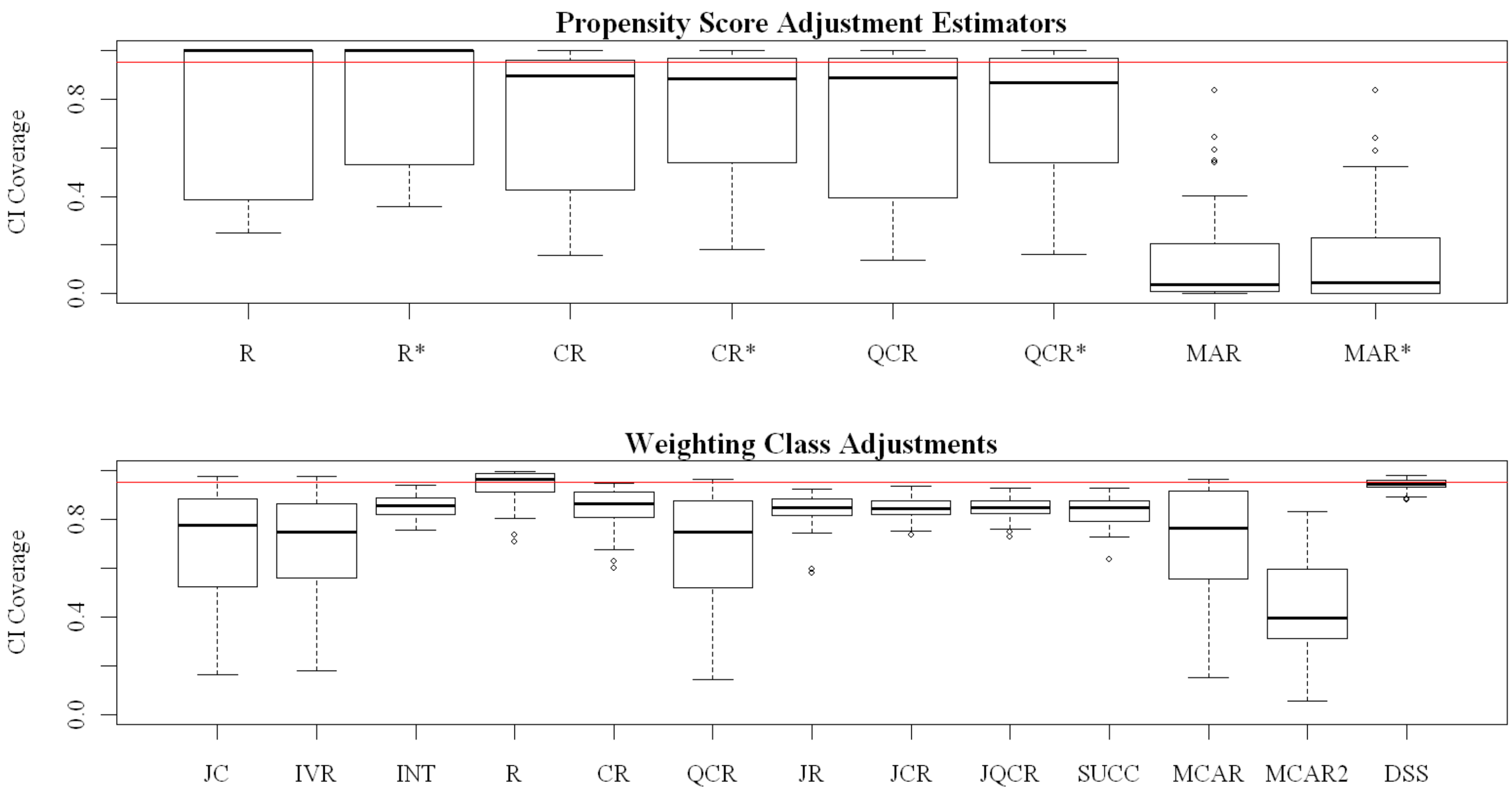

Figure 6.6: Confidence interval coverage across all scenarios and odds ratios for the population with a 30% decline in the mean response rate (Population 2 – incorrect specification for response propensity model)

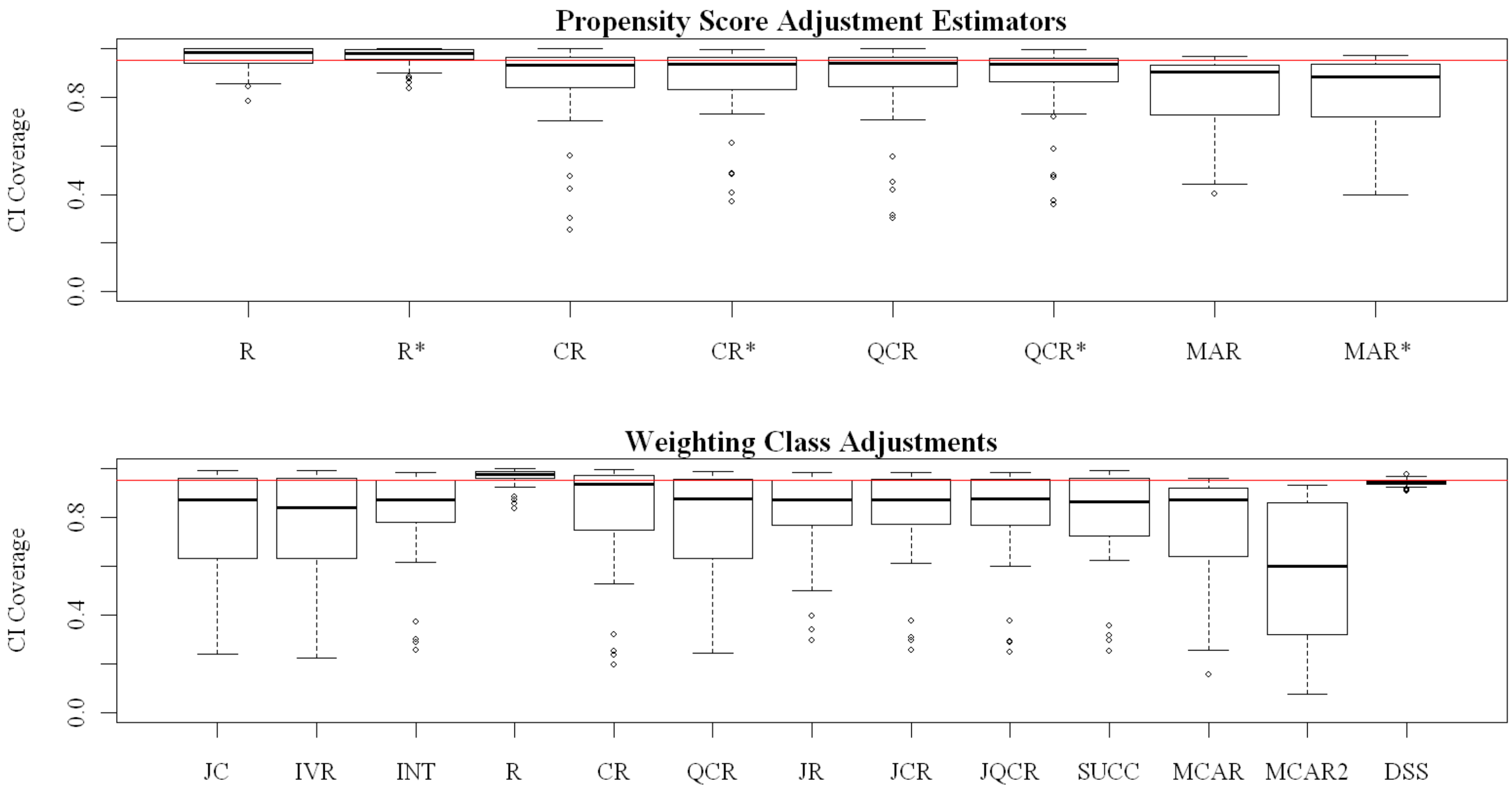

Figure 6.7: Confidence interval coverage across all scenarios and odds ratios for the population with a 30% decline in the mean success rate (Population 3 – incorrect specification for success propensity model)

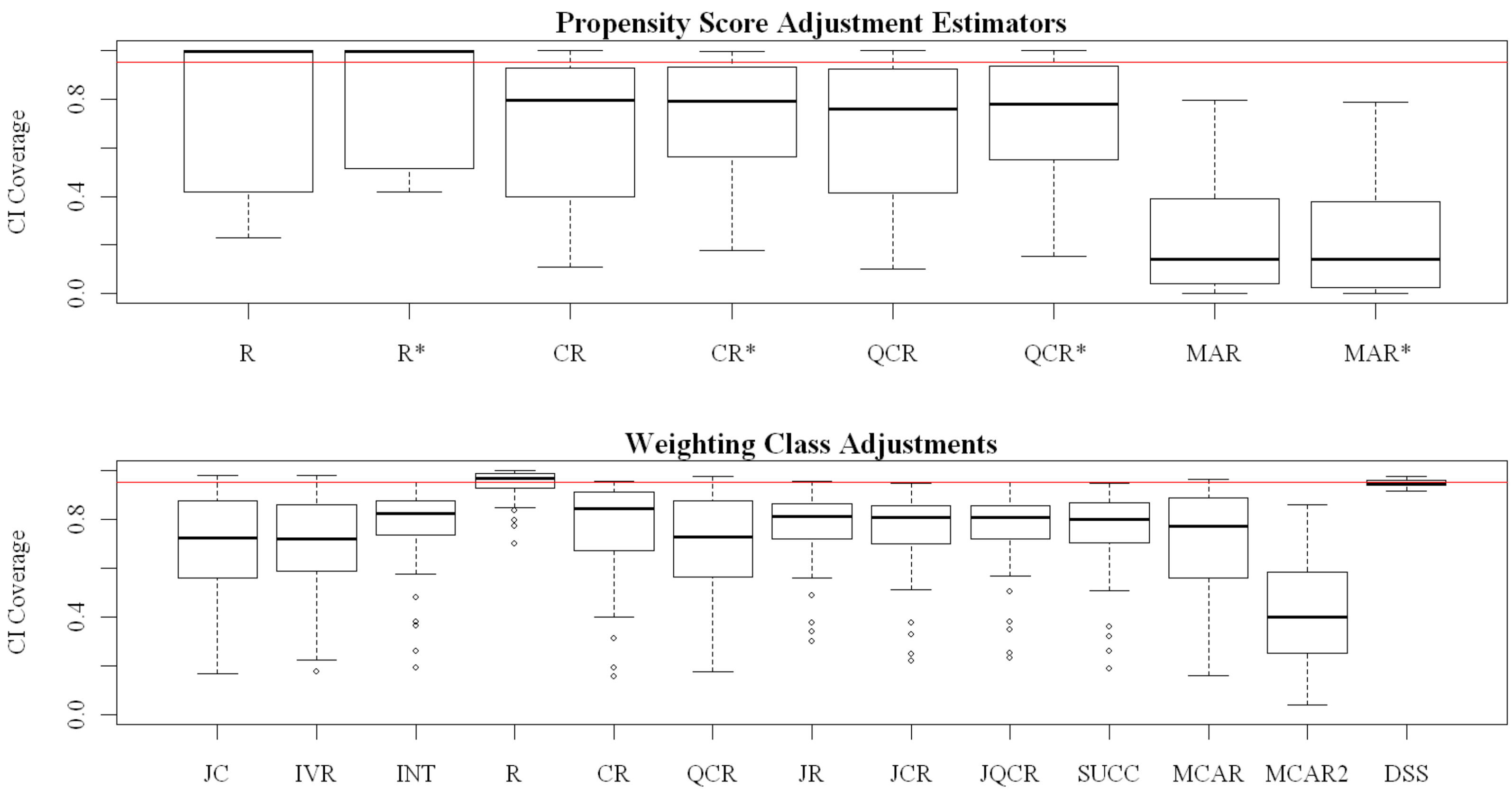

Figure 6.8: Confidence interval coverage across all scenarios and odds ratios for the population with 30% declines in the mean success rate and the mean response rate (Population 4 – incorrect specification for success and response propensity models)

6.3.3 Root mean square error

The effects of bias and variance on inference are mitigated when the root mean square error is minimized. However, low values of RMSE for estimators with poor confidence interval coverage may result from variance underestimation. Therefore, an acceptable estimator will minimize RMSE while maintaining nominal coverage rates.

For Population 1, the JR approach of the weighting class adjustment minimizes bias and variance to obtain the smallest measures of RMSE overall (Figure 6.9). However, several NMAR weighting class adjustments (INT, R, JCR, and JQCR) achieve similarly low values of the RMSE when the mean response rate is low to moderate ($\bar{p}_r \leq 0.6$) and the odds ratio is greater than one for Population 1. For odds ratios less than 1 and low to moderate mean response rates ($\bar{p}_r \leq 0.6$), the SUCC model also provides low RMSE. When the mean response rate is moderate ($\bar{p}_r = 0.6$) and the mean success rate is low to moderate ($\bar{p}_s \leq 0.6$), the R approach for the weighting class adjustment provides slightly smaller RMSE. When the odds ratio is 1, the MCAR2 approach obtains RMSE values similar to that obtain from the double sampling for stratification estimator. When the mean response rate ($\bar{p}_r$) is 0.9, RMSE values of the MCAR2 approach are less than those from the preferred NMAR weighting class adjustment estimators. Given the low confidence interval coverage of the MCAR2 approach discussed in the previous section, the low RMSE values for the MCAR2 method do not necessarily indicate a favorable performance.

For Population 2, RMSE nearly doubles for the propensity score adjustment estimates as compared to Population 1 (Figure 6.10). RMSE values are comparable between Populations 1 and 2 for the INT, JR, JCR, JQCR, and SUCC approaches to the weighting class adjustment. Typically, the SUCC approach performs better when the odds ratio is less than one, and the INT, JCR, and JQR approaches provide lowest RMSE when odds ratios are greater than one. The INT, JCR, JQCR, and SUCC approaches of the weighting class adjustment provide comparable RMSE values to those of the JR approach, indicating that the conditional success propensity score may perform as well as the outcome of interest when the success propensity model is correct. These weighting class adjustments also provide lower RMSE values than either of the MCAR estimators.

When the mean success rate declines by 30% relative to that on which the success propensity model is based (Population 3), the JR weighting class adjustment approach generates the lowest values of RMSE when the mean success rate is low ($\overline{p}_s = 0.3$), and the R weighting class adjustment approach performs best when the mean success rate is moderate to high ($\overline{p}_s \geq 0.6$) (Figure 6.11). In most cases, the propensity score adjustment estimators exhibit higher RMSE values compared to those obtained from the weighting class adjustment. When the mean response rate is 0.9, the MCAR2 approach generates RMSE values less than those from the preferred NMAR weighting class adjustment estimators. The smaller RMSE values for the weighting class adjustments and the MCAR2 approaches are a result of the variance underestimation which causes the poor confidence interval coverage discussed in the previous section.

When both the mean response rate and the mean success rate decline by 30% relative to the rates for the population used to obtain the success model (Population 4), the RMSE values from the weighting class adjustment are uniformly low compared to those from the propensity score adjustment estimator (Figure 6.12). However, this comparison is confounded by the poor confidence interval coverage of the weighting class adjustment for some scenarios and the increased bias for NMAR propensity score adjustment estimator approaches (R and R*). When the mean response rate is low to moderate ($\overline{p}_r \leq 0.6$), the JR weighting class adjustment approach generally provides the lowest values of RMSE. When the mean response rate is high ($\overline{p}_r = 0.9$) or for moderate response and moderate to high mean success rate ($\overline{p}_r = 0.6$ and $\overline{p}_s \geq 0.6$), the R approach of the weighting class adjustment provides the lowest RMSE across estimators and scenarios. As mentioned previously, the poor confidence interval coverage discussed in the previous section indicates variance underestimation. Therefore, minimizing RMSE may not be appropriate when choosing among these estimators of the population total.

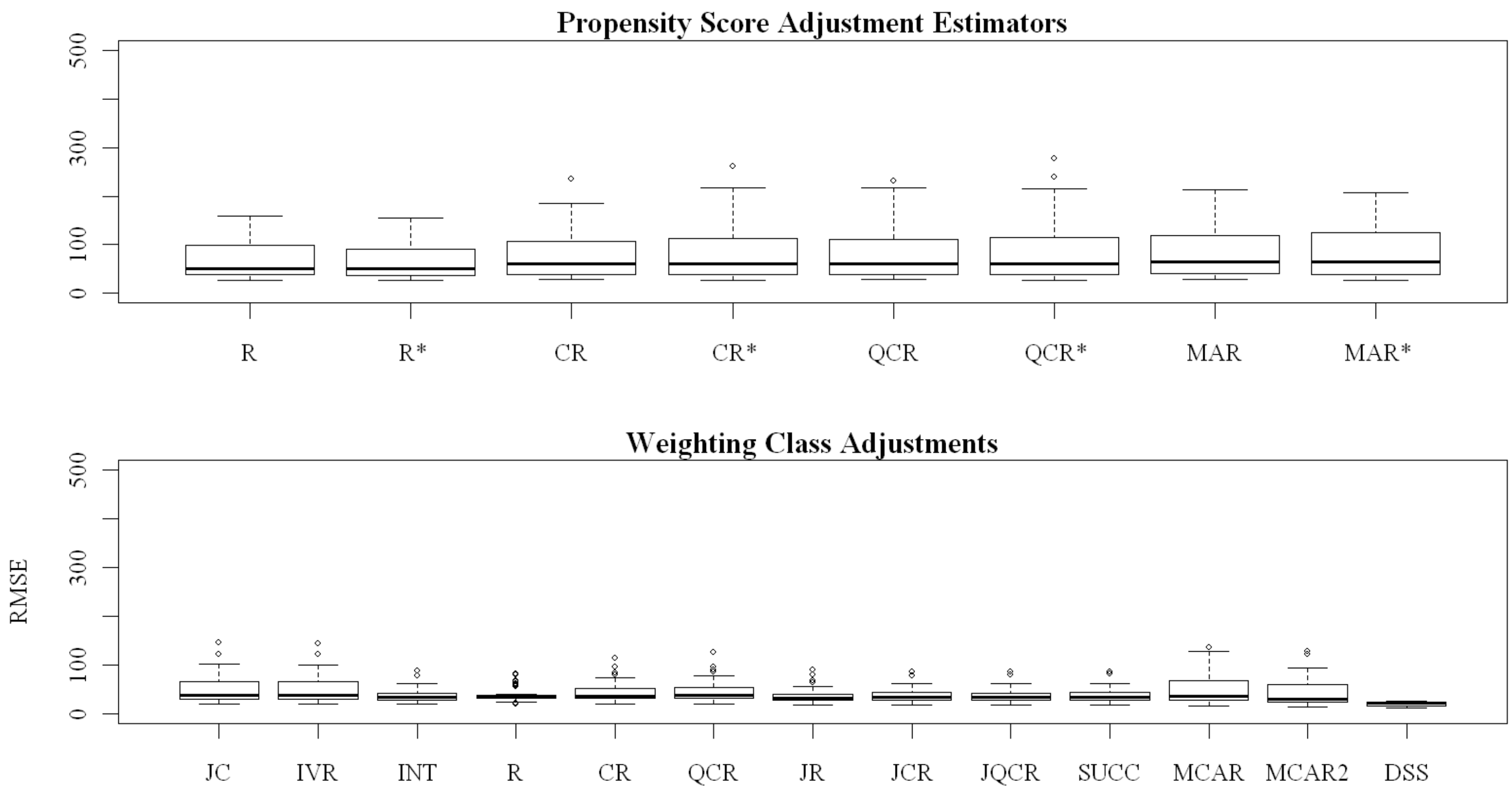

Figure 6.9: RMSE across all scenarios and odds ratios for the population with no change (Population 1 – correct specification for success and response propensity models)

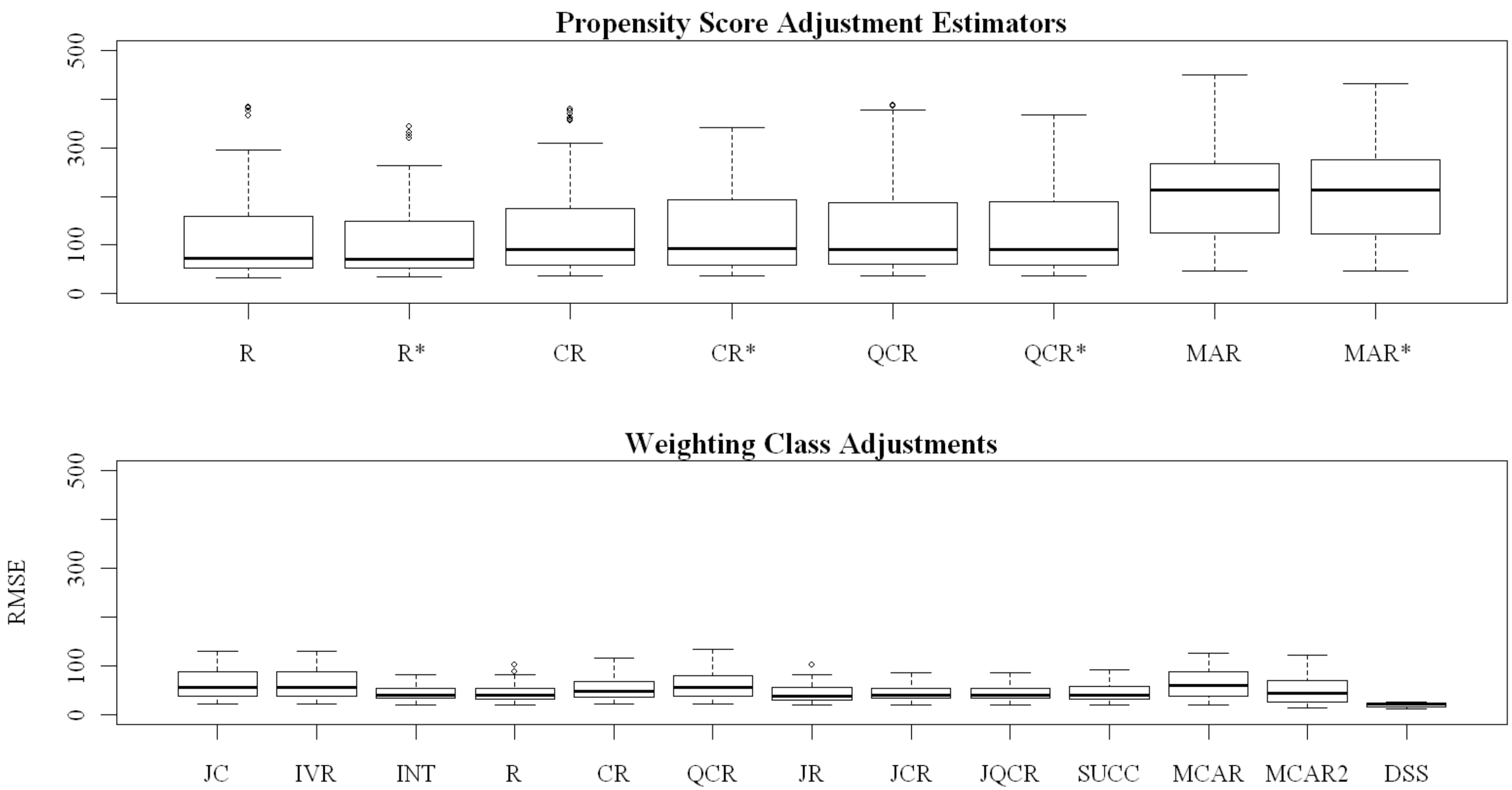

Figure 6.10: RMSE across all scenarios and odds ratios for the population with a 30% decline in the mean response rate (Population 2 – incorrect specification for response propensity model)

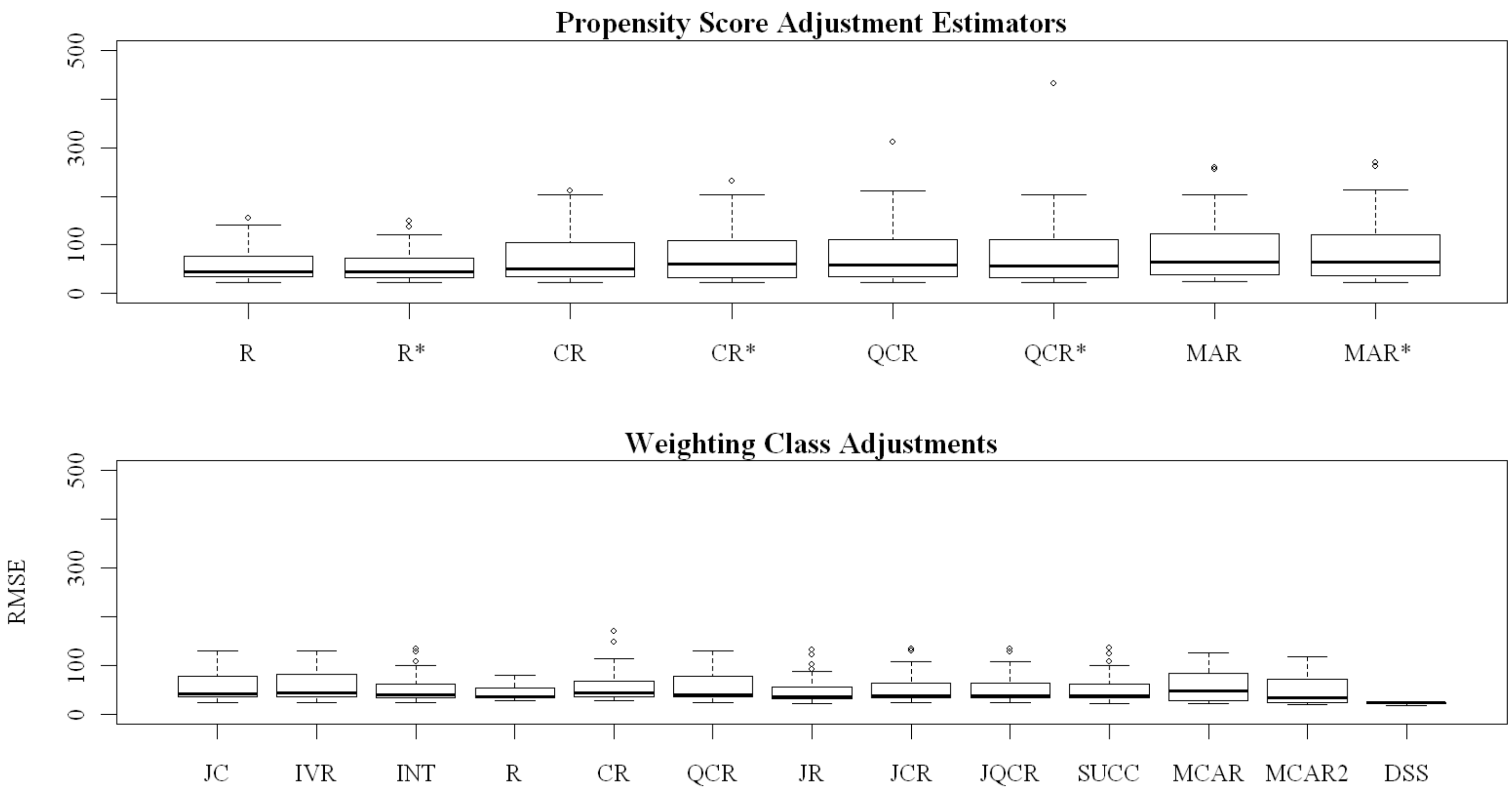

Figure 6.11: RMSE across all scenarios and odds ratios for the population with a 30% decline in the mean success rate (Population 3 – incorrect specification for success propensity model)

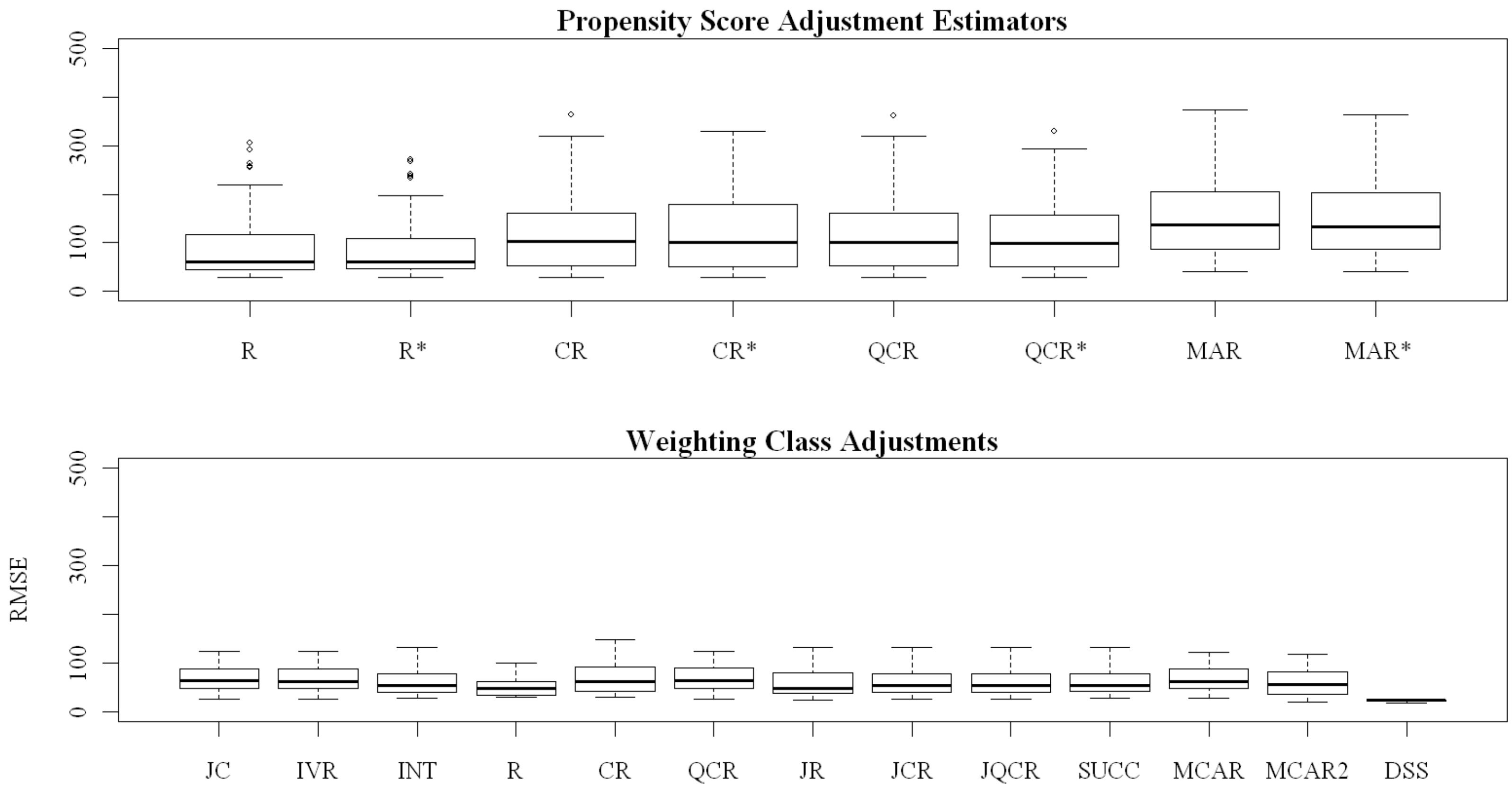

Figure 6.12: RMSE across all scenarios and odds ratios for the population with 30% declines in the mean success rate and the mean response rate (Population 4 – incorrect specification for success and response propensity models)

6.3.4 Discussion

6.3.4.1 General results

When the response propensity model and the success propensity model are correctly specified (Population 1), the most accurate and precise estimates of the total are obtained with INT, JR, JCR, JQCR, and SUCC approaches for the weighting class adjustment. The INT, JR, JCR, or JQCR approaches perform best when the odds ratio is greater than 1, and the SUCC approach is preferred when the odds ratio is less than 1. The INT, JR, JCR, JQCR, and SUCC approaches generate RMSE values only slightly larger than those observed for the JR weighting class adjustment approach, indicating that the conditional success model performs well in adjustment class formation when the conditional success and conditional response propensity models are accurately specified. The weighting class approaches provide confidence interval coverage slightly below the nominal rate for an odds ratio of 3 for low mean response and success rates, and the CR propensity score adjustment estimator provide better accuracy and confidence interval coverage in this scenario.

When the mean response rate declines by 30% (Population 2), the estimates exhibiting the least bias and RMSE are obtained with the JR, JCR, JQCR, and SUCC weighting class adjustments. However, none of these approaches provide nominal coverage rates when the mean response rate is low ($\overline{p}_r$ = 0.3) because all of these approaches incorporate the conditional success propensity score. The R weighting class adjustment approach, which does not rely on the conditional success propensity model, provides nominal coverage rates for most scenarios in which the mean response rate is moderate to high ($\overline{p}_r \geq 0.6$). The CR and CR* approaches of the propensity score adjustment estimator provide relatively unbiased estimates with nominal coverage rates when the response rate is moderate to high ($\overline{p}_r \geq 0.6$) and the mean success rate is low ($\overline{p}_s$ = 0.3). For Population 2, the mean response rate has decreased by 30%, so the relationship between the outcome of interest and the response indicator is different than that described by the conditional success propensity model. When the mean response rate is small ($\overline{p}_r$ = 0.3), a 30% decline yields a mean response rate of 0.21. Such low response may be prohibitive for attaining nominal coverage rates in this population with either the weighting class adjustment estimator or the propensity score adjustment estimator.

When the mean success rate exhibits a 30% decline (Population 3), the INT, R, JR, JCR, and JQCR approaches for weighting class adjustments provide estimates with the least bias and the lowest RMSE overall. The INT, JCR, and JQCR approaches are nearly as unbiased as the JR approach but do not require a nonrespondent subsample for each sample. However, nominal confidence interval coverage is only obtained for the R approach when either the mean response and success rates are low ($\overline{p}_s = \overline{p}_r$ = 0.3) or the mean response rate is moderate to high ($\overline{p}_r \geq 0.6$). When the mean success rate and the mean response rate are low ($\overline{p}_s = \overline{p}_r = 0.3$), the CR weighting class adjustment provides unbiased estimation and nearly nominal coverage of at least 0.87. The R and R* propensity score adjustment approaches provide nominal or nearly-nominal coverage when $\overline{p}_r = 0.3$ but the coverage rate tends to be too high when the mean response rate is moderate to high ($\overline{p}_r \geq 0.6$). An exception occur when the response rate is high ($\overline{p}_r = 0.9$) and the success rate is low ($\overline{p}_s = 0.3$); in this case, the R and R* propensity score adjustment approaches achieve nominal rates for most odds ratios.

When both the mean response rate and the mean success rate exhibit a 30% decline (Population 4), the INT, JCR, and JQCR weighting class adjustment approaches perform well overall in reducing bias and RMSE but demonstrate poor coverage rates. When the mean success rate is low ($\overline{p}_s = 0.3$) and the mean response rate is moderate to high ($\overline{p}_r \geq 0.6$), the CR weighting class adjustment and the CR, CR*, QCR, and QCR* approaches of the propensity score adjustment estimator reduce bias and RMSE as well as obtain nominal or nearly nominal (≥ 0.88) confidence interval coverage. For a moderate to high mean success rate ($\overline{p}_s \geq 0.6$), bias is more severe and/or confidence interval coverage is poor for all of the estimators. The double sampling for stratification estimator is the only estimator that consistently reduces bias and RMSE while providing confidence interval coverage when the mean success rate ($\overline{p}_s$) exceeds 0.3 for Population 3.

In many scenarios, the approaches that yield relatively unbiased and precise estimates for low mean response rates require a nonrespondent subsample. If a nonrespondent subsample is feasible for each sample, the least biased and most precise estimator of the population total is obtained from the design-based estimator for double sampling for stratification. This estimator was least biased across all scenarios with relative bias ranging from -0.01 to 0.01. Confidence interval coverage ranged from 0.85 to 0.99 for the double sampling for stratification estimator, with the poorest coverage occurring for high mean success rate ($\overline{p}_s = 0.9$) and an odds ratio greater than one. The RMSE did not exceed 25 for the double sampling for stratification estimator. This estimator minimizes bias and variance which achieving nominal coverage rates, indicating that the most optimal use of the nonrespondent subsample incorporates the survey design.

The R and R* approaches performed well overall in the simulation exercise, exhibiting very little bias and attaining nominal confidence interval coverage. Recall that these estimators performed poorly in the case study example in Chapter 5. Estimated response propensities for certain combinations of correlated binary predictors tended toward 0 or 1, a phenomenon called separation (Gelman et al., 2001). The binary predictors used in the simulation were generated from independent random processes, so separation did not affect the simulation estimates.

6.3.4.2 Applying the conditional success propensity model

The effectiveness of the methods that incorporate information from a previous nonrespondent subsample is examined. When the mean response rate and mean success rate are stable between surveys (Population 1), the conditional success propensity model obtained from a subsample of nonrespondents is an effective tool for nonresponse adjustment. The INT, JR, JCR, JQCR, and SUCC weighting class adjustment approaches provide low relative bias and nominal confidence interval coverage. All of these estimators use the conditional success propensity score, indicating that predictions from this model are effective in weighting class formation. When the mean response rate demonstrates a 30% decline (Population 2), the conditional success propensity score provides unbiased inference with nominal coverage rates for a low mean success rate ($\bar{p}_s$ = 0.3), moderate to high mean response rates ($\bar{p}_r \geq 0.6$), and odds ratios of at least one with the CR and CR* approaches of the propensity score adjustment estimator. Similarly, when the mean success rate declines by 30% (Populations 3 and 4) and the mean success rate is low ($\bar{p}_s$ = 0.3), the CR, CR*, QCR, and QCR* approaches of the propensity score adjustment estimator provide estimates of the harvest total that are nearly unbiased and exhibit nominal or nearly nominal coverage rates. When both the mean response rate and the mean success rate decline by 30% (Population 4), the weighting class adjustment approaches that incorporate the conditional success propensity score (INT, CR, QCR, JCR, JQCR, and SUCC) exhibit lower coverage rates when odds ratios are 0.33 or 3. These specific scenarios emphasize the cases in which a modeling approach may be used with nonrespondent subsample data to apply to surveys not augmented by additional sampling. These approaches often perform best in the simulation when the mean success rate is low ($\bar{p}_s$ = 0.3). In this case, a 30% decline in a mean success rate of 0.3 (which decreases by 0.09 to 0.21) represents a smaller absolute decline as would be observed in a mean success rate of 0.6 (which decreases by 0.18 to 0.42) or 0.9 (which decreases by 0.27 to 0.63). Therefore, predictions from the conditional success model are more accurate when the absolute change in the mean success rate ($\overline{p}_s$) is small.

6.3.4.3 Effectiveness of joint classification

Joint classification for NMAR methods is applied in the weighting class adjustment for the INT, JR, JCR, and JQCR approaches. In most cases, NMAR joint classification reduces bias and RMSE and improves confidence interval coverage as compared to methods that form adjustment cells from a single propensity score (CR, QCR, and SUCC). Benefits of NMAR joint classification are greater when the mean response rate is low ($\overline{p}_r$ = 0.3) or the odds ratio is less than one, and benefits are negligible when the mean success rate is high ($\overline{p}_s$ = 0.9). Note that joint classification approaches require the use of 25 adjustment classes rather than the five classes used by single classification approaches because adjustment classes are formed from quintiles of a response propensity score as well as quintiles of the conditional success propensity score. Using more adjustment classes may allow more precise estimation of response rates for weighting class adjustment.

6.3.4.4 Effectiveness of coarsening

Two levels of coarsening were examined in this simulation: coarsening of the response propensity scores in the propensity score adjustment estimator to reduce variance inflation due to small response propensities and coarsening of the success propensity scores used as predictors in the response propensity model. The first level of coarsening is evaluated by comparing the results of the four response propensity models to their coarsened counterparts, i.e. comparing the R approach to the R* approach, the CR approach to the CR* approach, etc. Compared to the approaches that do not use coarsening, coarsening reduces bias and improves confidence interval coverage when the mean response rate declines by 30% for all propensity score adjustment estimator approaches when the mean response rate is low ($\overline{p}_r$ = 0.3). However, under these circumstances, the confidence interval coverage is uniformly below the nominal rate and the design-based estimator from double-sampling for stratification is preferred.

The second level of coarsening is examined by comparing the approaches that do and do not incorporate the quintiles of the success propensity score as predictors in the response propensity model. Comparing the CR approach of the propensity score adjustment estimator with the QCR approach, bias is reduced when the mean response rate is low ($\overline{p}_r = 0.3$), the mean success rate is low to moderate ($\overline{p}_s \leq 0.6$), and the odds ratio is at least one. However, this improvement is not observed when comparing the CR* approach to the QCR* approach. Comparing the CR and QCR approaches of the weighting class adjustment, the only benefit to coarsening the success propensity scores is found when the mean response rate is low to moderate ($\overline{p}_r \leq 0.6$), the mean success rate is high ($\overline{p}_s = 0.9$), and the mean success rate experiences a 30% decline (Populations 3 and 4). In these cases, coarsening reduces bias and improves confidence interval coverage.

6.4 REVISITING THE CASE STUDY

For the New Mexico Department of Game and Fish elk hunter questionnaire, the response rates were 30% and 24% in 2001 and 2003, respectively. The estimates of the success rates from double sampling for stratification were 0.28 for 2001 and 0.29 for 2003 (Table 5.4). Both the response rate and the success rate exhibit declines but not of the magnitude used in the simulation study. Given the separation problems exhibited by the R approach to the propensity score adjustment estimator, the design-unbiased estimator of the total from double sampling for stratification is assumed to be the least biased estimate of the total for assessing the performance of the model-assisted approaches. Assuming that the estimates under double sampling for stratification are accurate, the response rate declined 20% between 2001 and 2003, and the success rate increased by 4% between the two years. The ratio of the odds of response for successful hunters to the odds of response for unsuccessful hunters was estimated to be 2.12 in 2001 and 2.35 in 2003.

Using these measures of the response rate, success rate, and odds ratio, the simulation results are examined to determine which approaches best suit the case study data. The 2001 estimates are compared to those from Population 1 from the simulation because the response and success models are obtained from the 2001 pilot data set and are assumed correct. With the exception of the R and JR approaches of the weighting class adjustment, the pilot data exhibit slightly higher but similar relative bias as compared to estimates from Population 1.

The 2003 pilot data are examined to determine which simulation population is most appropriate for comparison. The results of a two-sided test assuming large-sample asymptotic properties of and independence between the success rates from 2001 and 2003 provides suggestive but inconclusive evidence of a difference between years ($p = 0.0521$). For the comparison of the 2003 pilot data results to the simulation results, the mean success rate will be assumed stable. The scenario for the 2003 data shows similarities with both Population 1, which exhibited both stable success and response rates, and Population 2, which exhibited a stable success rate but a declining response rate. The 20% decline in response rate is not as extensive as the response rate decline examined in the simulations (30%). Overall, the relative bias estimates from the NMDGF pilot data are more similar to those observed for Population 1, indicating that either the observed declines in the success rate and response rate were not substantially large or the double sampling for stratification estimates of harvest total were slightly biased in the case study.

Several factors may explain the discrepancies observed between the pilot data results and the simulation results. First, the simulations were conducted assuming that a complete sample was obtained through nonrespondent subsampling. In the pilot study, only about 10% of the nonrespondents were contacted which may explain additional bias and variation. The double sampling for stratification estimator demonstrated near-perfect accuracy in the simulation but might be slightly more biased in simulations in which missingness is not completely resolved.

Table 6.3: Pilot data and simulation relative bias results for an odds ratio of 2, a mean success rate of 0.3, and a mean response rate of 0.3

Estimator	**NMDGF 2001**	**NMDGF 2003**	**Pop. 1**	**Pop. 2**	**Pop. 3**	**Pop. 4**
WC Adj. (INT)	0.08	0.16	0.04	0.02	0.00	-0.03
WC Adj. (R)	-0.10	0.05	0.07	0.16	0.10	-0.26
WC Adj. (CR)	0.12	0.12	0.10	0.14	0.04	0.10
WC Adj. (QCR)	0.17	0.16	0.14	0.10	0.09	0.08
WC Adj. (JR)	-0.11	-0.09	0.03	0.00	-0.01	-0.18
WC Adj. (JCR)	0.08	0.15	0.04	0.01	0.00	-0.04
WC Adj. (JQCR)	0.09	0.18	0.04	0.01	0.00	-0.04
WC Adj. (SUCC)	0.06	0.12	0.09	0.07	0.05	0.04
PSAE (CR)	0.07	0.09	0.02	-0.23	-0.05	-0.25
PSAE (CR*)	0.08	0.13	0.05	-0.18	-0.01	-0.19
PSAE (QCR)	0.12	0.14	0.07	-0.16	0.01	-0.22
PSAE (QCR*)	0.12	0.16	0.10	-0.11	0.04	-0.15
DSS	0.00†	0.00†	0.00	0.00	0.00	0.01

† Unbiased by assumption

Secondly, the simulations were based on a few covariates, and the models were correctly specified. The pilot data consist of a large number of variables, and important variables may not have been collected. Third, the declines in the mean success rate and/or the mean response rate were generated in the simulations by changing the regression coefficients in the success and/or response propensity regression models. Relationships between success and response may be more complex in the NMDGF elk hunter survey.

Final inference requires a consideration of the effects of the nonresponse bias. If total harvest is overestimated, elk licenses sales may be reduced and elk populations may grow beyond levels considered sustainable by the New Mexico Department of Game and Fish. Underestimation of harvest might encourage an increase in elk license sales, putting the population at risk. The conservative approach is to tend toward overestimation of the elk harvest rather than underestimation.

6.5 CONCLUSIONS

A simulation study conducted over a range of odds ratios, mean success and response rates, and population changes to reflect changes in the success and/or response rate was used to assess the precision and accuracy of the NMAR propensity score methodology introduced in Chapter 3 and implemented in Chapters 4 and 5. A nonrespondent subsample forms the basis of a predictive model for success and can be applied to data from surveys not augmented by a nonrespondent subsample if the mean response and mean success rates do not differ between the populations. When the mean response rate and/or mean success rate differ, weighting class adjustment and propensity score adjustment estimators may provide biased estimates and/or poor confidence interval coverage. In these cases, a nonrespondent subsample and design-based estimation with double sampling for stratification provide the best basis for unbiased inference with nominal coverage rates.

7: CONCLUSIONS

The results of this thesis research may be summarized in several main points. First, the propensity score methodology extensions to nonignorable missingness proposed in this thesis are effective in reducing bias in modified Horvitz-Thompson approaches for nonresponse adjustment when propensity score models are correct. Second, when propensity score models are incorrect due to population-level changes in the response and/or success rate, NMAR adjustments that employ the conditional success propensity score may increase bias and RMSE and produce confidence interval with less than nominal coverage rates. Third, using MAR adjustments when data are NMAR may increase the bias of the estimate of the total compared to simply ignoring the nonresponse.

The methods proposed in this thesis are relatively easy to implement. Weighting class adjustments and the propensity score adjustment estimator require the use of logistic regression, quantile calculation, and weighting. The methods in this thesis provide straightforward methodology that might encourage agencies to examine their survey nonresponse more closely and evaluate the direction and degree of nonresponse bias. An application and improvements of the methods proposed in this thesis are discussed further in this chapter.

7.1 APPLICATIONS

The approaches discussed in this thesis may be applied to surveys from other agencies that conduct surveys for which unit nonresponse occurs and for which the outcome of interest is a binary response. These agencies might include federal, state, and private organizations that survey individuals regarding hunting, fishing, or use of a natural resource. These approaches have more general applications to surveys obtaining any dichotomous outcome that is NMAR, such as a yes/no question on a sensitive topic or presence/absence of a species in an occupancy survey.

Occupancy estimation under imperfect detection is an area for possible application of these methods. Occupancy estimation typically involves estimating the proportion of area that is occupied by a species (MacKenzie, Nichols, Royle, Pollock, Bailey, and Hines, 2006). The outcome of interest is a binary indicator of presence or absence of the species. Under imperfect detection, an additional level of error is introduced when an entity is present but not detected during the survey. Logistic regression models incorporate covariates related to occupancy and detection rates, and maximum likelihood is used to obtain estimates of occupancy and detection rates. These occupancy estimation methods are model based. Complex survey designs are not recommended for model-based occupancy estimation because the survey design probabilities currently cannot be incorporated into this model-based approach, potentially causing invalid inference. Model-assisted occupancy estimation for complex survey design is an interesting area for further research.

When complex survey designs must be used, the propensity score techniques proposed in this thesis may provide model-assisted approaches for occupancy estimation. In this setting, occupancy rates are represented by the success rates explored in this thesis, and detection rates correspond to response rates. Therefore, nondetection is analogous to nonresponse. Results for the NMAR weighting class adjustment and propensity score adjustment estimators are expected to apply for the levels of the occupancy and detection rates investigated in this thesis. A nonrespondent subsample is necessary to use the NMAR propensity score methods discussed in this thesis. This subsample establishes the conditional success propensity score model which is the occupancy model in this setting. It is important to note that occupancy inference relative to the odds ratio requires the implicit assumption that the probability of a false-positive detection is non-zero, i.e. $P\left(R_i = 1 \middle| Y_i = 0\right) > 0$, where in the occupancy setting R_i is 1 when the species of interest is detected in unit *i* and Y_i is 1 when the species of interest occupies unit *i*. Because the odds of a false positive detection are likely to be small relative to the odds of a true detection, occupancy analysis will more often involve odds ratios greater than one.

7.2 FUTURE WORK

First, an additional area of development for the NMAR propensity score methodology includes an examination of how these methods perform for estimates of cross-classes, which cut across adjustment cells (Little, 1986). Because adjustment cells are based on propensity scores which reduce the dimension of the covariate data set to a single variable, subpopulations of interest are more likely to cut across adjustment classes, especially if the factors defining the subpopulation exhibit interactions with other factors. For example, NMDGF obtains elk harvest estimates for subpopulations such as each weapon type which exhibits many significant interactions with other variables in the pilot data and for which each level occurs in several adjustment classes.

Secondly, in the simulation, the nonrespondent subsample was assumed to census all nonrespondents from the original sample so that a complete sample was ultimately obtained. Examining what proportion of a nonrespondent subsample is necessary for unbiased inference under a range of conditions would provide information on a minimum proportion or number of nonrespondents that are needed for unbiased and precise inference with nominal coverage rates. Factors that are likely to influence these results are the odds ratio of response for successful versus unsuccessful units, the second-phase response rate, and the survey mode.

Third, propensity score methodology for nonignorable nonresponse could be extended to non-binary outcomes by replacing the logistic regression model for the binary outcome of interest with the appropriate predictive mean model. These predictions could be used as covariates in the response propensity model in an approach analogous to the CR approaches of the propensity score adjustment estimator or the weighting class adjustment.

Finally, NMAR propensity score methodology may be extended to approaches that do not use weighting, such as imputation techniques. Stochastic regression imputation could be applied in a multiple imputation approach using the conditional response propensity score model to impute missing outcomes. Methods that accommodate complex survey design should also be investigated.

7.3 CONCLUSIONS

For declining success rates, some of the scenarios examined in the simulation produce biased estimates with very poor confidence interval coverage. Paradoxically, declines in success rates are difficult to detect due to the nature of nonignorable nonresponse.

Means from respondent outcomes might not be similar to nonrespondent means. Overall, the presence and rate of declines in the response rate and/or the success rate affect the choice of a nonresponse adjustment approach. When the response rate fell below 0.3 in the simulation, the design-based estimate from double sampling for stratification was the only method that provides unbiased estimation with nominal coverage rates. If the cost of biased inference (for example, overhunting an elk population) outweighs the cost of additional sampling, then the reduction of bias due to nonresponse is paramount. Changes in the survey design or mode might be needed to find a method less subject to nonresponse bias. Nonrespondent subsampling with double sampling for stratification could be included as a regular feature of the survey design. The initial sampling effort may be scaled back to accommodate the additional survey costs of the nonrespondent subsample.

Ultimately, NMDGF applied the best approach for reducing nonresponse bias by minimizing the nonresponse rate (Lohr, 1999). Incentives and dual-language survey instruments were ineffective in increasing response rates from the original questionnaire. The mandatory survey return program, discussed in Chapter 6, roughly tripled the return rate compared to the original survey design. Harvest data are collected automatically in the mandatory program with telephone and internet survey modes. The penalties obtained from nonrespondents cover the costs of the mandatory return program, and the resulting estimates are less susceptible to error from nonresponse.

Despite response rates of 83% to 86%, the NMDGF estimates may still be subject to bias and error from nonresponse. The simulation results for an odds ratio of 2, a success rate of 0.3, and a response rate of 0.9 indicated a relative bias of 0.12 for the MCAR2 approach. If this level of bias is unacceptable for management, then nonrespondent subsampling would provide additional information to determine if the remaining missingness is MAR or NMAR. This information would aid in selecting the appropriate adjustment approach. The methodology proposed in this thesis provides an approach that could provide NMDGF with a more accurate and complete set of annual harvest information by adjusting harvest data sets from the original questionnaire using the information from the 2001 and 2003 nonrespondent subsamples. Resolving the problem of nonignorable nonresponse bias is not trivial and requires additional survey effort at the design and analysis stages. Applying the appropriate survey methodology may improve estimates subject to bias and error from nonignorable nonresponse so that management decisions are made on the best possible information.

8: BIBLIOGRAPHY

2007 Behavioral Risk Factor Surveillance System Summary Data Quality Report (2008). Washington, D.C.: National Center for Chronic Disease Prevention and Health Promotion.

Bailey, L. (1986). *Nonresponse Adjustment for The National Sample of the National Crime Survey.* Washington, D.C.: U.S. Census Bureau. (Statistical Research Report Number CENSUS/SRD/RR-86/08)

Bailey, L., A.E. Jansto, and C. Smith (1991). *An Evaluation of Imputed Data for Non-Respondents To The 1987 Economic Censuses-Single-Unit Establishments.* Washington, D.C.: U.S. Census Bureau. (Statistical Research Report Number CENSUS/SRD/RR-90/08)

Bailey, L. (2005). *Examining Weighting Alternatives to Compensate for Longitudinal Nonresponse in the Survey of Income and Program Participation*. Washington, D.C.: U.S. Census Bureau.

Bartholomew, D.J. (1961). A Method of Allowing for `Not-at-Home' Bias in Sample Surveys. *Applied Statistics* 10(1), 52-59.

Bethlehem, J.G. (2002). Weighting nonresponse adjustments based on auxiliary information. In R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (Eds.), *Survey Nonresponse*. (pp. 275-288). New York: John Wiley and Sons, Inc.

Billingsley, P. (1995). Probability and Measure. (3rd ed.). New York: Wiley.

Birkes, D. (2009). Personal interview. November, 2009.

Birmingham, J and G.M. Fitzmaurice (2002). A pattern-mixture model for longitudinal binary responses with nonignorable nonresponse. *Biometrics* 58(4), 989-996.

Buckland, S.T., D.R. Anderson, K.P. Burnham, and J.L. Laake (1993). *Distance Sampling*. New York: Chapman and Hall.

Bull, S.B., C. Mak, and C.M.T. Greenwood (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis* 39,57-74.

Cassel, C.M., C-E, Sarndal, and J.H. Wretman (1977). *Foundations of inference in survey sampling*. New York: John Wiley and Sons, Inc.

Cassel, C.M., C-E, Sarndal, and J.H. Wretman (1983). Some uses of statistical models in connection with the nonresponse problem. In W.G. Madow, I. Olkin, and D.B. Rubin (Eds.), *Incomplete data and in sample surveys*. (Vol. 3. pp. 143-160). New York: Academic Press.

Cochran, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 24, 295-313.

Cochran, W.G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley and Sons, Inc.

Copas, A.J. and V.T. Farewell (1998). Dealing with non-ignorable non-response by using an 'enthusiasm-to-respond' variable. *Journal of the Royal Statistical Society*, Series A 161(3), 385-396.

Curtin, R., S. Presser, and E. Singer (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly* 69(1), 87-98.

David, M., R. Little, M. Samuhel, and R. Triest (1983). Nonrandom nonresponse models based on the propensity to respond. *Proceedings of the Business Economics Statistics Section of the American Statistical Association*, 168-174.

David, M., R.J.A. Little, M.E. Samuhel, and R.K. Triest (1986). Alterntaive methods for CPS income imputation. *Journal of the American Statistical Association* 81, 29-41.

Dawid, A.P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B*, 41(1), 1-31.

Dehejia, R.H. and S. Wahba (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics,* 84(1), 151–161.

Dempster, A. P., N. M. Laird, D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*. Series B 39(1), 1-38.

Detlefsen, R. (2007). *Issues Related to Achieving Target Response Rates for Economic Surveys at the U.S. Census Bureau*. Washington, D.C: Federal Economic Statistics Advisory Committee, Bureau of Labor Statistics and U.S. Census Bureau.

Deville, J.-C., and Sarndal, C.E. (1992). Calibration Estimation in Survey Sampling. *Journal of the American Statistical Association* 87, 376-382.

Dillman, D.A. (2000). *Mail and Telephone Surveys: The Tailored Design Method*. New York: John Wiley and Sons, Inc.

Dillman, D.A., J.L. Eltinge, R.M. Groves, and R.J.A. Little (2002). Survey nonresponse in design, data collection, and analysis. In R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (Eds.), *Survey Nonresponse*. (pp. 3-26). New York: John Wiley and Sons, Inc.

Dillman, D.A., G. Phelps, R. Tortora, K. Swift, J. Kohrell, J. Berck, and B.L. Messer. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research* 38, 1- 18.

Elliott, M.R. and R.J.A. Little (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics* 16, 191-209.

Elliott, M.R., R.J.A. Little, and S. Lewitsky (2000). Subsampling callbacks to improve survey efficiency. *Journal of the American Statistical Association* 95, 730-738.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* 80(1), 27-38.

Foster, E.M. (2003). Propensity score matching: An illustrative analysis of dose response. *Medical Care* 41(10), 1183–1192.

Gelman, A. and J.B. Carlin (2002). Poststratification and weighting adjustments. In R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (Eds.), *Survey Nonresponse*. (pp. 289-302). New York: John Wiley and Sons, Inc.

Gelman, A., A. Jakulin, M.G.Pittau, and Y.-S. Su., (2001). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2(4), 1360-1383.

Hansen, M.H. and W.N. Hurwitz (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association* 41(236), 517- 529.

Harrell, F.E. and C.E. Davis (1982). A new distribution-free quantile estimator. *Biometrika* 69(3), 635-640.

Harrel, F.E., Jr, K.L. Lee, and D.B. Mark (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 15(4), 361-87.

Harrod, L.A. (2007). *2003-04 Elk Harvest Survey Nonresponse Bias Ana*lysis. Santa Fe, New Mexico: New Mexico Department of Game and Fish. Contract 01-516-60. 43 pp.

Helsel, D.R. (2005). *Nondetects and Data Analysis.* New Jersey: John Wiley and Sons, Inc.

Holt, D. and T.M.F. Smith (1979). Post stratification. *Journal of the Royal Statistical Society A* 142, 33-46.

Horvitz, D.G. and D.J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47(260), 663-685.

Hosmer, D.W. and S. Lemeshow (2000). *Applied Logistic Regression* (2nd ed.). New York: John Wiley and Sons, Inc.

Ibrahim, J.G., M-H Chen, and S.R. Lipsitz (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika* 88(2), 551-564.

Imbens, G.W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* 87(3), 706-710.

Kaigh, W.D. and P.A. Lachenbruch (1982). A generalized quantile estimator. *Communications in Statistics, Part A, Theory and Methods* 11, 2217-2238.

Kalton, G. and D. Kasprzyk (1986). The treatment of missing survey data. *Survey Methodology* 12, 1-16.

Kauff, J., R. Olsen, and T. Fraker (2002). Nonrespondents and Nonresponse Bias: Evidence from a Survey of Former Welfare Recipients in Iowa. Washington, D.C.: Mathematica Policy Research, Inc. (MPR Reference No.: 8217-909 and 8703-106)

Keeter, S., C. Kennedy, M. Dimock, J. Best, and P. Craighill (2006). Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. *Public Opinion Quarterly* 70(5), 759-779.

Kennickell, A.B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson (Eds.), *Record Linkage Techniques*. (pp. 248–267). Washington, D.C.: National Academy Press.

Kish, L. (1978). Populations for survey sampling. *Survey Statistician* 1, 14-15.

Kott, P. (2005). *"No" Is the Easiest Answer: Using Calibration to Assess Nonignorable Nonresponse in the 2002 Census of Agriculture*. Minneapolis, Minnesota: Joint Statistical Meetings.

Lesser, V.M., D.A. Dillman, J. Carlson, F. Lorenz, R. Mason, and F. Willits (2001). Quantifying the influence of incentives on mail survey response rates and their effects on nonresponse error. *Proceedings of the Annual Meeting of the American Statistical Association*. Atlanta, Georgia.

Lessler, J.T. and W.D. Kalsbeek (1992). *Nonsampling Error in Surveys*. New York: John Wiley and Sons, Inc.

Little, J.A.R. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association* 77(37), 237 -250.

Little, J.A.R. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review* 54(2), 139 -157.

Little, J.A.R. and D.B. Rubin (2002). *Statistical Analysis with Missing Data* (2nd ed.). New Jersey : John Wiley and Sons, Inc.

Little, J.A.R. and Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine* 22, 1589-1599.

Lohr, S.L. (1999). *Sampling: Design and Analysis.* California: Brooks/Cole Publishing Company.

Lu, B. (2005). Propensity score matching with time-dependent covariates. *Biometrics* 61, 721-728.

MacDonald, D. and E.G. Dillman (1968). Techniques for estimating non-statistical bias in big game harvest surveys. *Journal of Wildlife Management* 32(1), 119-129.

MacKenzie, D.I., J.D. Nichols, J.A. Royle, K.H. Pollock, L.L. Bailey, and J.E. Hines (2006). *Occupancy Estimation and Modeling*. Academic Press: San Francisco.

Maiti, T. and V. Pradhan (2007). A comparative study of the bias corrected estimates in logistic regression. *Statistical Methods in Medical Research* 17, 621–634.

Mallows, C.L. (1973). Some comments on Cp. *Technometrics* 15, 661-676.

Ming, K. and P.R. Rosenbaum (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* 56, 118-124.

Molenberghs, G. M.G. Kenward, G. Verbeke, C. Beunckens, and C. Sotto. (2009). *Every missing not at random model for incomplete data has a missing at random counterpart with equal fit*. Presentation at the 2009 Joint Statistical Meetings. (Activity number 136). Washington, D.C.

Nandram, B. and J.W. Choi (2000). A hierarchical Bayesian nonresponse model for binary data with uncertainty about ignorability. *American Statistical Association Proceedings of the Joint Statistical Meetings, Survey Research Section*. (pp. 574-579). Alexandria, VA: American Statistical Association.

Oh, H.L. and F.J. Scheuren (1983). Weighting adjustment for unit nonresponse. In W.G. Madow, I. Olkin, and D.B. Rubin (Eds.), *Incomplete data and in sample surveys*. (Vol. 2. pp. 143-184). New York: Academic Press.

Pendleton, G.W. (1992). Nonresponse patterns in the federal waterfowl hunter questionnaire survey. *Journal of Wildlife Management* 56(2), 344-348.

Pickreign, J.D. and J.R. Gabel (2005). Health Benefits Offer Rates: Is There a Nonresponse Bias? *Health Services Research* 40(2), 401-412.

Politz, A. and W. Simmons (1949). An attempt to get the "not at homes" into the sample without callbacks. *Journal of the American Statistical Association* 44, :9-31.

Potter, F. (1988). Survey of Procedures to Control Extreme Sampling Weights. *ASA Proceedings of the Section on Survey Research*, Alexandria, VA: American Statistical Association.

Qin, J., D. Leung, and J. Shao (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association* 97(457), 193-200.

Rao, P.S.R.S. (1983). Callbacks, follow-ups, and repeated telephone calls. In W.G. Madow, I. Olkin, and D.B. Rubin (Eds.), *Incomplete data and in sample surveys*. (Vol. 2. pp. 33-44). New York: Academic Press.

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

Rotnitzky, A. and J.M. Robins (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* 82, 805-820.

Rubin, D.B. (1983). Conceptual issues in the presence of nonresponse. In W.G. Madow, I. Olkin, and D.B. Rubin (Eds.), *Incomplete data and in sample surveys*. (Vol. 2. pp. 123-420). New York: Academic Press.

Rubin, D.B. and N. Thomas (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association* 95(450), 573-585.

Salvucci, S., E. Walter, V. Conley, S. Fink, M. Saba, and S. Kaufman (1997). *NCES measurement error programs.* Washington, D.C.: U.S. Department of Education. (NCES 97-464)

Särndal, C-E, B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer.

SAS/STAT (2002) software, Version 9 of the SAS System for Windows (2002). SAS Institute Inc., Cary, NC, USA.

Scharfstein, D.O., A. Rotnitzky, and J.M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 94(448), 1096-1120.

Sfakianakis, M.E. and D.G. Verginis (2008). A new family of nonparametric quantile estimators. *Communications in Statistics - Simulation and Computation* 37, 337-345.

Shtatland, E. S., Kleinman K., and Cain E. M. (2004). A new strategy of model building in proc logistic with automatic variable selection, validation, shrinkage and model averaging. *SUGI 29 Proceeding*. (Paper 191-29). Cary, NC: SAS Institute, Inc.

Shtatland, E. S., Moore, S., Dashevsky I., Miroshnik, I., Cain, E. M. and Barton, M. B. (2000). How to be Bayesian in SAS: model selection uncertainty in PROC LOGISTIC and PROC GENMOD. *Proceedings of the 13th Annual NorthEast SAS® Users Group Conference.* (pp. 724 -732). Pittsburgh, PA: NorthEast SAS Users Group, Inc.

Singh, B. and J. Sedransk (1983). Bayesian survey design for nonresponse. In W.G. Madow, I. Olkin, and D.B. Rubin (Eds.), *Incomplete data and in sample surveys*. (Vol. 3. pp. 227-247). New York: Academic Press.

Standards and Guidelines for Statistical Surveys (2006). Washington, DC.: Office of Management and Budget.

Stasny, E.A. (1991). Hierarchical models for the probabilities of a survey classification and nonresponse: An example from the National Crime Survey. *Journal of the American Statistical Association* 86, 296-303.

Steinhorst, R.K. and M.D. Samuel (1989). Sightability adjustment methods for aerial surveys of wildlife populations. *Biometrics* 45, 415-425.

Steyerberg, E.W., M.J. Eijkeman, F.E. Harrell, Jr., and J.D. Habbema (2000). Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine* 19, 1059-1079.

Stromberg, A.J. (1997). Robust covariance estimates based on resampling. *Journal of Statistical Planning and Inference* 57, 321-334.

Tang, G, R.J.A. Little, and T.E. Raghunathan (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* 90(4), 747- 764.

Thompson, S.K. (1992). *Sampling*. New York: John Wiley and Sons, Inc.

US Census Monitoring Board Presidential Members (2001). US Census Monitoring Board Presidential Members Final Report to Congress, Washington.

Vartivarian, S. and R.J.A. Little (2003). On the formation of weighting adjustment cells for unit nonresponse. Ann Arbor, Michigan: The University of Michigan Department of Biostatistics Working Paper Series. (Working Paper 10)

Venables, W.N. and B.D. Ripley (2002). *Modern Applied Statistics with S*. New York: Springer.

Verginis, D. (2009). Email interview. Mar. 13 and 19, 2009.

Wang, R., J. Sedransk, and J. H. Jinn (1992). Secondary data analysis when there are missing observations. *Journal of the American Statistical Association*, 87(420), 952- 961.

Wolter, K. (2007). *Introduction to Variance Estimation* (2nd ed.). New York: Springer.

Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association* 66, 411-414.

Zhang, P. (2003). Multiple imputation: Theory and method. *International Statistical Review* 71(3), 581–592

APPENDICES

APPENDIX A: Notation definition and glossary of terms

A.1 Notation definition

- β_{cr} : Vector of regression coefficients from the logistic regression model of conditional response propensity scores.
- β_{qcr} : Vector of regression coefficients from the logistic regression model of conditional response propensity scores calculated from the quintiles of the conditional success propensity score.
- β_r : Vector of regression coefficients from the logistic regression model of NMAR response propensity scores.
- β_{rMAR} : Vector of regression coefficients from the logistic regression model of MAR response propensity scores.
- β_s : Vector of regression coefficients from the logistic regression model of success propensities.
- $\theta = (\theta_1, ..., \theta_n)$: Vector of response propensity weights for the sample where θ_i is the response propensity weight of unit *i* and $\theta_i = p_i^{-1}$ for general response propensity score p_i.

- θ_i^* : The response propensity weight quintile associated with the *i*th unit where $\theta_i^* = \mathrm{Q}(\boldsymbol{\theta}, \mathbf{q})_i$. We will assume that $\theta_{ri}^* \approx p_{ri}^{-1}$ for the approaches propensity score adjustment estimator that employ coarsening.
- π_i : The inclusion probability for unit *i,* where $\pi_i = P(D_i = 1)$.
- π_{ij} : The joint inclusion probability for units *i* and *j*

  where $\pi_{ij} = P(D_i = 1, D_j = 1)$.
- *τ*: the population total parameter such that $\tau = \sum_{i=1}^{N} y_i$ for values y_i of variables

  Y_i, $i = 1, \ldots, N$.
- φ :all unknown parameters related to response
- $\boldsymbol{\psi}$: Unknown parameters in the independent distributions of $\mathbf{X}_1, \ldots, \mathbf{X}_N$.
- **D** = (D_1, D_2,…,D_N)**:** the vector of sample inclusion indicators where, for *i* = 1,…*N,*

  $$D_i = \begin{cases} 1, & \text{unit } i \text{ is included in the sample} \\ 0, & \text{otherwise.} \end{cases},$$

  $E(D_i) = \pi_i$, and $Var(D_i) = \pi_i(1 - \pi_i)$.
- $e(\mathbf{X}_i)$: Response propensity score in the MAR setting of Rosenbaum and Rubin (1983) such that $e(\mathbf{X}_i) = P(\mathrm{R}_i = 1 | \mathbf{X}_i)$.

- $f\left(R_i \middle| Y_i, \mathbf{X}_i\right)$: The conditional distribution of R_i, which is used to describe missing data mechanisms.
- $f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\psi})$: The distribution of the independent covariate matrices, **X**.
- m: The number of respondents in the sample such that $m = \sum_{i=1}^{n} R_i$ and $m \le n$.
- N: The population size.
- n: the number of elements of s, i.e. the sample size.
- n_b: The nonrespondent subsample size.
- $\mathbf{p}_{cr}$: The vector of conditional response propensity scores where $p_{cri} = P\left(R_i = 1 \middle| p_{si}, \mathbf{X}_i\right) = p_{cr}\left(p_{si}, \mathbf{X}_i; \boldsymbol{\beta}_{cr}\right)$ and $\boldsymbol{\beta}_{cr}$ is a vector of logistic regression coefficients.
- $\mathbf{p}_{cr}^{*}$: The vector of quintiles of the conditional response propensity scores such that $p_{cri}^{*} = \mathrm{Q}\left(\mathbf{p}_{cr}, \mathbf{q}\right)_i = \left\{ Q\left(\mathbf{p}_{cr}, \frac{h-0.5}{H}\right) : Q\left(\mathbf{p}_{cr}, q_{h-1}\right) < p_{cri} \le Q\left(\mathbf{p}_{cr}, q_h\right) \right\}$ for h = 1, …, H where H = 5.
- $\mathbf{p}_{qcr}$: The vector of conditional response propensity scores based on quintiles of the conditional success propensity score where $p_{qcri} = P\left(R_i = 1 \middle| p_{si}^{*}, \mathbf{X}_i\right) = p_{qcr}\left(p_{si}^{*}, \mathbf{X}_i; \boldsymbol{\beta}_{qcr}\right)$ and $\boldsymbol{\beta}_{qcr}$ is a vector of logistic regression coefficients.

- $\mathbf{p}_{qcr}^{*}$: The vector of quintiles of $\mathbf{p}_{qcr}$ such that

  $p_{qcri}^{*} = \mathrm{Q}\left(\mathbf{p}_{qcr}, \mathbf{q}\right)_i = \left\{ Q\left(\mathbf{p}_{qcr}, \frac{h-0.5}{H}\right) : Q\left(\mathbf{p}_{qcr}, q_{h-1}\right) < p_{qcri} \le Q\left(\mathbf{p}_{qcr}, q_h\right) \right\}$ for h = 1, …, H where H = 5.

- $\mathbf{p}_r$: The vector of NMAR response propensity scores such that

  $\mathbf{p}_r = \left(p_{r1}, ..., p_{rn}\right)$. For i = 1,…n, p_{ri} is the probability of response for unit i such that $p_{ri} = P\left(R_i = 1 \middle| D_i = 1\right)$. In the modeling context, p_{ri} is characterized as $p_{ri} = P\left(R_i = 1 \middle| \mathrm{Y}_i, \mathbf{X}_i\right) = p_r\left(Y_i, \mathbf{X}_i; \beta_r\right)$, where β_r is a vector of logistic regression coefficients.

- $\mathbf{p}_r^{*}$: The vector of quintiles of the NMAR response propensity scores such that

  $p_{ri}^{*} = \mathrm{Q}\left(\mathbf{p}_r, \mathbf{q}\right)_i = \left\{ Q\left(\mathbf{p}_r, \frac{h-0.5}{H}\right) : Q\left(\mathbf{p}_r, q_{h-1}\right) < p_{ri} \le Q\left(\mathbf{p}_r, q_h\right) \right\}$ for h = 1, …, H where H = 5.

- $\overline{p}_r$: Mean response rate used in the simulations taking the value of 0.3, 0.6, or 0.9.

- $\mathbf{p}_{rMAR}$: The vector of estimated MAR response propensity scores. For the weighting class adjustment $\mathbf{p}_{rMAR}$ is modeled only from variables related to response to avoid collinearity problems. For the propensity score adjustment estimator, $\mathbf{p}_{rMAR}$ is modeled from all variables related to response or success.

- $\mathbf{p}^*_{rMAR}$: The vector of quintiles of the MAR response propensity scores such that p^*_{rMARi} =

  $$\mathrm{Q}\left(\mathbf{p}_{rMAR},\mathbf{q}\right)_i = \left\{ Q\left(\mathbf{p}_{rMAR},\frac{h-0.5}{H}\right) : Q\left(\mathbf{p}_{rMAR},q_{h-1}\right) < p_{rMARi} \le Q\left(\mathbf{p}_{rMAR},q_h\right) \right\}$$

  for $h = 1, \ldots, H$ where $H = 5$.

- p_{rij} : The joint probability of response for units i and j where

  $p_{rij} = P\left(R_i = 1, R_j = 1 \middle| D_i = 1, D_j = 1\right)$.

- $\mathbf{p}_s$: The vector of true conditional success propensity scores, where

  $\mathbf{p}_s = \left(p_{s1},\ldots,p_{sn}\right)$, $p_{si} = p_s\left(\mathrm{R}_i,\mathrm{X}_i\right) = P\left(\mathrm{Y}_i = 1 \middle| \mathrm{R}_i,\mathrm{X}_i\right)$, and

  $\operatorname{logit}\left(p_{si}\right) = \log\left(\frac{p_{si}}{1-p_{si}}\right) = \mathbf{X}'_{si}\boldsymbol{\beta}_s$.

- $\mathbf{p}^*_s$: The vector of conditional success propensity score quintiles

  $p^*_{si} = \mathrm{Q}\left(\mathbf{p}_s,\mathbf{q}\right)_i = \left\{ Q\left(\mathbf{p}_s,\frac{h-0.5}{H}\right) : Q\left(\mathbf{p}_s,q_{h-1}\right) < p_{si} \le Q\left(\mathbf{p}_s,q_h\right) \right\}$ for $h = 1, \ldots,$ H where $H = 5$.

- $\overline{p}_s$: Mean success rate used in the simulations taking the value of 0.3, 0.6, or 0.9.

- $\mathbf{p}_{sIVR}$: The vector of estimated MAR success propensity scores for instrumental variable regression (David et al., 1983) modeled from variables related to success and the MAR response propensity score such that

$\mathbf{p}_{sIVR} = P\left(\mathbf{Y} = 1 \middle| \mathbf{p}_{rMAR}, \mathbf{X}\right)$.

- $\mathbf{p}^*_{sIVR}$: The vector of success propensity score quintiles for instrumental variable regression used for adjustment cell formation in weighting class adjustment calculated such that

  $p^*_{sIVRi} = \mathrm{Q}\left(\mathbf{p}_{sIVR}, \mathbf{q}\right)_i = \left\{Q\left(\mathbf{p}_{sIVR}, \dfrac{h-0.5}{H}\right) : Q\left(\mathbf{p}_{sIVR}, q_{h-1}\right) < p_{sIVRi} \leq Q\left(\mathbf{p}_{sIVR}, q_h\right)\right\}$

  for $h = 1, \ldots, H$ where $H = 5$.

- $\mathbf{p}_{sMAR}$: The vector of estimated MAR success propensity scores modeled from variables related to success.

- $\mathbf{p}^*_{sMAR}$: The vector of MAR success propensity score quintiles such that

  $p^*_{sMARi} =$

  $\mathrm{Q}\left(\mathbf{p}_{sMAR}, \mathbf{q}\right)_i = \left\{Q\left(\mathbf{p}_{sMAR}, \dfrac{h-0.5}{H}\right) : Q\left(\mathbf{p}_{sMAR}, q_{h-1}\right) < p_{sMARi} \leq Q\left(\mathbf{p}_{sMAR}, q_h\right)\right\}$

  for $h = 1, \ldots, H$ where $H = 5$.

- $Q\left(\mathbf{W}, q\right)$: The quintile function for a vector, **W**, evaluated at the probability q such that $Q\left(\mathbf{W}, q\right) = \left\{w : P\left(W \leq w\right) = q\right\}$. When **q** is a vector, $Q\left(\mathbf{W}, \mathbf{q}\right)$ is a vector of quintiles evaluated at those proportions.

- $Q(.)_i$: Quintile (with probability notation suppressed) corresponding to the i^{th} unit.
- **q**: Vector of probabilities for which the quintiles of propensity scores are calculated.
- **R =** $(R_1, R_2,\ldots,R_n)$**:** the indicator vector of response given inclusion in the survey where, for $i = 1,\ldots n$,

  $$R_i = \begin{cases} 1, & \text{unit } i \text{ responds} \\ 0, & \text{otherwise.} \end{cases}$$

  $E(R_i) = p_{ri}$, and $Var(R_i) = p_{ri}(1 - p_{ri})$
- S_h : The set of units in the sample falling in weighting class h, where h = 1, …, 5.
- S: nonempty set such that $S \subseteq \mathcal{U}$ represents an unordered sample
- $\hat{T}$: The general form for the estimated total from the propensity score adjustment estimator
- $\hat{T}_d$: The estimated total from the design-based estimator for double sampling for stratification.
- $\hat{T}_{MCAR}$: Estimator of the population total for the MCAR method. See MCAR definition (3).

- $\hat{T}_{MCAR2}$: Estimator of the population total for the MCAR method. See MCAR2 definition.
- $\mathcal{T}$: the set of all sets s.
- $\mathcal{U}$: finite population consisting of the set of all units for which inference is to be made
- u_i: the i^{th} unit in the finite population set
- w_i : the sample inclusion weight such that $w_i = \pi_i^{-1}$
- $\mathbf{X}$: The complete set of available covariates such that $\mathbf{X} = \{\mathbf{X}_1, ..., \mathbf{X}_N\}$.
- $\mathbf{X}^{(b)}$: The covariate matrix for outcomes observed in the nonrespondent subsample.
- $\mathbf{X}_r$: The covariate matrix used to model response containing variables related to success, the outcome of interest (**Y**), and their interactions.
- $\mathbf{X}_{cr}$: The covariate matrix used to model response containing variables related to success, the estimated conditional success propensity score ($\hat{p}_{si}$), and their interactions
- $\mathbf{X}_{qcr}$: The covariate matrix used to model response containing variables related to success, the estimated conditional success propensity score quintile ($\hat{p}_{si}^*$), and their interactions.

- $\mathbf{X}_s$: The covariate matrix used to model success containing the response indicator (**R**) and its interactions.
- $\mathbf{x}_i$: Realized covariate matrix for unit *i*.
- **Y**: The vector of binary outcomes in the population such that **Y** = (Y_1, $Y_2,\ldots,Y_N$), where, for $i = 1,\ldots N$

$$Y_i = \begin{cases} 1, & \text{unit } i \text{ is successful} \\ 0, & \text{otherwise.} \end{cases}.$$

- $\mathbf{Y}^m$: the vector of missing Y_i such that $\mathrm{Y}^m = \{\mathrm{Y}_i : D_i = 1, R_i = 0\}$ for $i = 1,\ldots n$.
- $\mathbf{Y}^o$: the vector of observed Y_i such that $\mathrm{Y}^o = \{\mathrm{Y}_i : D_i = 1, R_i = 1\}$ for $i = 1,\ldots n$.
- $\mathbf{Y}^{(b)}$: the outcomes obtained from a nonrespondent subsample
- y_i : observed value of the random variable, Y_i
- $\mathrm{y}^{(b)}$: The vector of observed values of the random vector, **Y**, from responding units in the nonrespondent subsample.
- y^o : The vector of observed values of the random vector, **Y**, from responding units.

A.2 Glossary of terms

Conditional classification: Term introduced here to indicate that the adjustment classes are constructed under the assumption of nonignorable nonresponse.

Conditional propensity score: General reference to propensity scores computed from models that account for nonignorable missingness. This is a general term for the conditional response propensity score or the conditional success propensity score.

Conditional response propensity score (p_{ri}): The response propensity score from a logistic or probit regression model that that accounts for nonignorable nonresponse. In the methods proposed here, this requires modeling the response weight from the conditional success propensity score quintile.

Conditional response propensity weight ($\theta_i = p_{ri}^{-1}$): The inverse of the response propensity score from a regression model that accounts for nonignorable nonresponse. In the methods proposed here, this requires modeling the response weight from the conditional success propensity score quintile.

Conditional response propensity weight quintile (θ_i^*): The quintile of the conditional response propensity weights associated with unit *i*.

Conditional success propensity score (p_{si}): The probability of success from a logistic or probit regression model that accounts for nonignorable nonresponse such that $p_{si} = P\left(\mathrm{Y}_i = 1 \middle| \mathrm{R}_i, \mathrm{X}_i\right) = p_s\left(\mathrm{R}_i, \mathrm{X}_i\right)$.

Conditional success propensity score quintile (p_{si}^{*}): The quintile of the conditional success propensity scores associated with unit *i*.

C*R:* (1) NMAR classification approach of the weighting class adjustment in which adjustment classes are formed from quintiles of the conditional response propensity score, $\boldsymbol{p}_{cr}$. (2) Propensity score adjustment estimator approach in which the response rate for a modified Horvitz-Thompson estimator (equation 3.2) is estimated by the conditional response propensity scores, $\boldsymbol{p}_{cr}$.

CR:* Propensity score adjustment estimator approach in which the response rate for a modified Horvitz-Thompson estimator (equation 3.2) is estimated by the quintiles of the conditional response propensity scores, $\boldsymbol{p}_{cr}^{*}$.

Double-sampling for stratification (DSS): Design-based approach that employs a two-phase sample to account for nonresponse. Using the initial sample as the first phase, a sample or census of the nonrespondents is contacted in a second-phase.

INT: NMAR joint classification approach of the weighting class adjustment in which adjustment classes are formed from quintiles of the MAR response propensity score, $\boldsymbol{p}_{rMAR}$, and quintiles of the conditional success propensity score, $\boldsymbol{p_s}$.

JR: NMAR joint response propensity classification. A NMAR joint classification approach of the weighting class adjustment in which adjustment classes are formed from quintiles of the NMAR response propensity score, $\boldsymbol{p}_r$, and quintiles of the conditional success propensity score, $\boldsymbol{p_s}$.

JCR: Joint conditional response propensity classification. NMAR joint classification approach of the weighting class adjustment in which adjustment classes are formed from quintiles of the conditional response propensity score, $\boldsymbol{p}_{cr}$, and quintiles of the conditional success propensity score, $\boldsymbol{p_s}$.

JQCR: Joint conditional response propensity classification based on quintiles of the conditional success propensity score. A NMAR joint classification approach of the weighting class adjustment in which adjustment classes are formed from quintiles of the conditional response propensity classification based on quintiles of the conditional success propensity score, $\boldsymbol{p}_{qcr}$, and quintiles of the conditional success propensity score, $\boldsymbol{p}_{s}$.

MAR: (1) Missingness mechanism for which the response indicator is related to the outcome of interest through covariates. The distribution of the response indicator conditional on the observed outcomes and related covariates is independent of the missing outcomes and reduces to $f\left(R\middle|Y,X,\varphi\right)=f\left(R\middle|X,\varphi\right)$ for all Y and φ. (2) MAR classification approach of the weighting class adjustment in which adjustment classes are formed from quintiles of the MAR response propensity score, $\boldsymbol{p}_{rMAR}$. (3) Propensity score adjustment estimator approach in which the response rate for a modified Horvitz-Thompson estimator (equation 3.2) is estimated by the MAR response propensity scores, $\boldsymbol{p}_{rMAR}$.

MAR:* Propensity score adjustment estimator approach in which the response rate for a modified Horvitz-Thompson estimator (equation 3.2) is estimated by the quintiles of the MAR response propensity scores, $\boldsymbol{p}^{*}_{rMAR}$.

MAR response propensity score $\left(\boldsymbol{p}_{rMAR}\right)$: Response propensity scores under MAR missingness such that $p_{rMARi} = P\left(R_i = 1 \middle| \underset{\sim}{\mathbf{X}}_i\right) = p_r\left(\mathbf{X}_i; \boldsymbol{\beta}_{rMAR}\right)$, where $\boldsymbol{\beta}_{rMAR}$ is a vector of logistic regression coefficients.

MCAR: (1) Missingness mechanism for which the missing data is not related to the outcome of interest, any covariates, or any unknown parameters. In this case, the conditional distribution of **R** reduces to $f\left(R \middle| \boldsymbol{Y}, X, \varphi\right) = f\left(R \middle| \varphi\right)$ for all Y and φ. (2) References a design-based estimator used in the simulation to demonstrate the effect of assuming MCAR missingness under a range of circumstances. For this estimator, the standard Horvitz-Thompson estimator is used but inclusion probabilities are adjusted for nonresponse by assuming MCAR missingness, i.e. $\pi_i = \dfrac{m}{N}$.

MCAR2: References a design-based estimator used in the simulation to demonstrate the effect of assuming MCAR missingness under a range of circumstances. For this estimator, the standard Horvitz-Thompson estimator is used and inclusion probabilities are not adjusted for nonresponse, i.e. $\pi_i = \dfrac{n}{N}$.

NMAR: Not missing at random missingness mechanism. This missingness mechanism is related to the outcome of interest, and the form of the conditional distribution of the response indicator, $f\left(R \middle| Y, X, \varphi\right)$, cannot be simplified.

NMAR joint classification: Extension of joint classification (Vartivarian and Little, 2002) for MAR missingness in which adjustment classes for weighting class adjustments are formed from quintiles of the conditional success propensity score and quintiles of either the MAR response propensity score (INT approach), NMAR response propensity score (JR approach), conditional response propensity score (JCR approach), or the conditional response propensity score from the quintiles of the conditional success propensity score (JQCR approach).

NMAR response propensity score $\left(\boldsymbol{p}_r\right)$: Response propensity scores under NMAR missingness such that $p_{ri} = P\left(R_i = 1 \middle| \mathrm{Y}_i, \underset{\sim}{\mathbf{X}}_i\right) = p_r\left(Y_i, \mathbf{X}_i; \boldsymbol{\beta}_r\right)$, where $\boldsymbol{\beta}_r$ is a vector of logistic regression coefficients.

Odds ratio (OR): The ratio of the odds of response for successful units to the odds of response for unsuccessful units, such that:

$$OR = \frac{P\left(R=1\middle|Y=1\right)P\left(R=0\middle|Y=0\right)}{P\left(R=1\middle|Y=0\right)P\left(R=0\middle|Y=1\right)}.$$

Propensity score: Term developed by Rosenbaum and Rubin (1983) to describe the conditional probability of assignment to a treatment group versus a control group. In this application, "treatment" may represent assignment to the respondent group or the binary outcome group. This term is used to apply to propensity scores in general. The propensity score is often estimated from logistic or probit regression modeling.

Propensity score adjustment estimator (PSAE): Modified Horvitz-Thompson estimator for NMAR missingness that employs conditional response propensity weight quintiles to weight for nonignorable nonresponse. See equation (5.2.3).

QC*R:* (1) NMAR classification approach of the weighting class adjustment in which adjustment classes are formed from quintiles of the conditional response propensity score based on quintiles of the conditional success propensity score, $\boldsymbol{p}_{qcr}$. (2) Propensity score adjustment estimator approach in which the response rate for a modified Horvitz-Thompson estimator (equation 3.2) is estimated by the conditional response propensity score based on quintiles of the conditional success propensity score, $\boldsymbol{p}_{qcr}$.

QCR:* Propensity score adjustment estimator approach in which the response rate for a modified Horvitz-Thompson estimator (equation 3.2) is estimated by the quintiles of the conditional response propensity score based on quintiles of the conditional success propensity score, $\boldsymbol{p}^*_{qcr}$.

R: (1) NMAR classification approach of the weighting class adjustment in which adjustment classes are formed from quintiles of the NMAR response propensity score, $\boldsymbol{p}_r$. (2) Propensity score adjustment estimator approach in which the response rate for a modified Horvitz-Thompson estimator (equation 3.2) is estimated by the NMAR response propensity scores, $\boldsymbol{p}_r$.

R:* Propensity score adjustment estimator approach in which the response rate for a modified Horvitz-Thompson estimator (equation 3.2) is estimated by the quintiles of the NMAR response propensity scores, $\boldsymbol{p}^*_r$.

Response probability (or response rate): the probability of response but not necessarily a propensity score. The response probability could be the estimated by the observed response rate within an adjustment cell or the ratio of sums of design weights.

Response propensity score: the probability of response estimated from logistic or probit regression. This is a general term that includes conditional response propensity scores.

Response propensity weight: Inverse of the probability of response estimated from logistic or probit regression modeling. This is a general term that includes conditional response propensity weights.

Response propensity weight quintile: Quintile of the vector of response propensity weights from the sample.

SUCC: NMAR approach of the weighting class adjustment in which adjustment classes are formed from quintiles of the conditional success propensity score, $\boldsymbol{p_s}$.

Success propensity score: the probability of success estimated from logistic or probit regression modeling. This is a general term that includes conditional success propensity scores.

Success propensity score quintile: The quintile of the vector of response propensity scores associated with unit *i*. .

Success propensity weight ($\theta_{si} = p_{si}^{-1}$): Inverse of the probability of success estimated from logistic or probit regression modeling. This is a general term that includes conditional success propensity weights.

VL: MAR joint classification approach of the weighting class adjustment in which adjustment classes are formed from quintiles of the MAR response propensity score, $\boldsymbol{p}_{rMAR}$, and quintiles of the MAR success propensity score, $\boldsymbol{p}_{sMAR}$.

APPENDIX B: Proof that p_{cri} is a balancing score for R**.**

Lemma: $p_{cr}\left(p_{si}, X_i\right)$ is a balancing score for **R**.

Proof: Conditioning on all other potentially relevant variables, we see that:

$$P\left[R_i = 1 \middle| Y_i, \mathbf{X}_i, p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right] = \begin{cases} P\left[R_i = 1 \middle| Y_i = 1, \mathbf{X}_i, p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right], y_i = 1 \\ P\left[R_i = 1 \middle| Y_i = 0, \mathbf{X}_i, p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right], y_i = 0 \end{cases}$$

$$= \begin{cases} \dfrac{P\left[Y_i = 1 \middle| R_i = 1, \mathbf{X}_i, p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right] P\left[R_i = 1 \middle| \mathbf{X}_i, p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right]}{P\left[Y_i = 1 \middle| \mathbf{X}_i, p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right]}, y_i = 1 \\ \dfrac{\left\{1 - P\left[Y_i = 1 \middle| R_i = 1, \mathbf{X}_i, p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right]\right\} P\left[R_i = 1 \middle| \mathbf{X}_i, p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right]}{\left\{1 - P\left[Y_i = 1 \middle| \mathbf{X}_i, p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right]\right\}}, y_i = 0 \end{cases}$$

$$= \begin{cases} \dfrac{E\left[E\left(Y_i \middle| R_i = 1, \mathbf{X}_i, p_{si}\right) \middle| R_i = 1, \mathbf{X}_i, p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right] p_{cr}\left(p_{si}, \mathbf{X}_i\right)}{E\left[E\left(Y_i \middle| \mathbf{X}_i, p_{si}\right) \middle| \mathbf{X}_i, p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right]}, y_i = 1 \\ \dfrac{\left\{1 - E\left[E\left(Y_i \middle| R_i = 1, \mathbf{X}_i, p_{si}\right) \middle| R_i = 1, \mathbf{X}_i, p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right]\right\} p_{cr}\left(p_{si}, \mathbf{X}_i\right)}{\left\{1 - E\left[E\left(Y_i \middle| \mathbf{X}_i, p_{si}\right) \middle| X_i, p_{cr}\left(p_{si}, \mathbf{X}_i\right)\right]\right\}}, y_i = 0 \end{cases}$$

$$
=\begin{cases}
\dfrac{E\left[E\left(Y_i\middle|R_i=1,\mathbf{X}_i,p_{si}\right)\middle|R_i=1,\mathbf{X}_i,p_{cr}\left(p_{si},\mathbf{X}_i\right)\right]p_{cr}\left(p_{si},\mathbf{X}_i\right)}{E\left\{E\left[E\left(Y_i\middle|\mathbf{X}_i,R_i\right)\middle|\mathbf{X}_i,p_{si}\right]\middle|\mathbf{X}_i,p_{cr}\left(p_{si},\mathbf{X}_i\right)\right\}}, y_i=1 \\
\dfrac{\left\{1-E\left[E\left(Y_i\middle|R_i=1,\mathbf{X}_i,p_{si}\right)\middle|R_i=1,\mathbf{X}_i,p_{cr}\left(p_{si},\mathbf{X}_i\right)\right]\right\}p_{cr}\left(p_{si},\mathbf{X}_i\right)}{1-E\left\{E\left[E\left(Y_i\middle|\mathbf{X}_i,R_i\right)\middle|\mathbf{X}_i,p_{si}\right]\middle|\mathbf{X}_i,p_{cr}\left(p_{si},\mathbf{X}_i\right)\right\}}, y_i=0
\end{cases}
$$

$$
=\begin{cases}
\dfrac{E\left[p_{si}\middle|R_i=1,\mathbf{X}_i,p_{cr}\left(p_{si},\mathbf{X}_i\right)\right]p_{cr}\left(p_{si},\mathbf{X}_i\right)}{E\left[E\left(p_{si}\middle|\mathbf{X}_i,p_{si}\right)\middle|\mathbf{X}_i,p_{cr}\left(p_{si},\mathbf{X}_i\right)\right]}, y_i=1 \\
\dfrac{\left\{1-E\left[p_{si}\middle|R_i=1,\mathbf{X}_i,p_{cr}\left(p_{si},\mathbf{X}_i\right)\right]\right\}p_{cr}\left(p_{si},\mathbf{X}_i\right)}{1-E\left[E\left(p_{si}\middle|\mathbf{X}_i,p_{si}\right)\middle|\mathbf{X}_i,p_{cr}\left(p_{si},\mathbf{X}_i\right)\right]}, y_i=0
\end{cases}
$$

$$
=\begin{cases}
\dfrac{E\left[p_{si}\middle|R_i=1,\mathbf{X}_i,p_{cr}\left(p_{si},\mathbf{X}_i\right)\right]p_{cr}\left(p_{si},\mathbf{X}_i\right)}{E\left[p_{si}\middle|\mathbf{X}_i,p_{cr}\left(p_{si},\mathbf{X}_i\right)\right]}, y_i=1 \\
\dfrac{\left\{1-E\left[p_{si}\middle|R_i=1,\mathbf{X}_i,p_{cr}\left(p_{si},\mathbf{X}_i\right)\right]\right\}p_{cr}\left(p_{si},\mathbf{X}_i\right)}{1-E\left[\left(p_{si}\middle|\mathbf{X}_i,R_i\right)\middle|\mathbf{X}_i,p_{cr}\left(p_{si},\mathbf{X}_i\right)\right]}, y_i=0
\end{cases}
$$

$$
=\begin{cases}
\dfrac{p_{si}p_{cr}\left(p_{si},\mathbf{X}_i\right)}{p_{si}}, y_i=1 \\
\dfrac{\left(1-p_{si}\right)p_{cr}\left(p_{si},\mathbf{X}_i\right)}{\left(1-p_{si}\right)}, y_i=0
\end{cases}
$$

$= p_{cr}\left(p_{si},\mathbf{X}_i\right)$ for all possible values of y_i.

Therefore, $p_{cr}\left(p_{si},\mathbf{X}_i\right)$ is a balancing score for **R**.

APPENDIX C: Model variable definitions

Variable	Variable Definition
Age1	Licensee age class: <18 years old
Age2	Licensee age class: 18 to 34 years old
Age3	Licensee age class: 35 to 49 years old
Age4	Licensee age class: 50 to 64 years old
ES	Bag limit class: either-sex hunt
Harv	Indicator of harvest success: 1 = harvested an elk
LandPub	Indicator of public landowner hunt: 1 = public land, 0 = private land
Male	Indicator of male licensee: 1 = male, 0 = female
MB	Bag limit class: mature-bull hunt
MBA	Bag limit class: mature-bull or antlerless hunt
Month	Month of the hunt beginning in October at the beginning of the hunt year
NW	Area of the state: NW
NE	Area of the state: NE
Resident	Indicator of NM residency: 1 = NM resident, 0 = out-of-state licensee
Size1	Hunt size: ≤30 licensees in hunt
Size2	Hunt size: 30 to 170 licensees in hunt
SW	Area of the state: SW
TOResp	Indicator of response to original survey: 1 = respondent, 0 = nonrespondent
Wpn1	Weapon type: Center-fire, muzzle-loader, or bow hunts
Wpn2	Weapon type: Bow-only hunts
Wpn3	Weapon type: Muzzle-loader only hunts
Wpn4	Weapon type: Hunts for impaired hunters

APPENDIX D: Success and response model selection results

Table D.1: Explanatory variables for final success propensity models

Model	Success propensity model variables
MAR (**Y** ~ **X**$_s$)	Wpn1, Wpn2, Wpn3, LandPub, NW, NE, SW, MBA, MB, ES, Resident, Month, Wpn1*LandPub, Wpn3*LandPub, Wpn3*NW, Wpn2*NE, Wpn1*MB, Wpn2*ES, Wpn1*Resident, Wpn2*Resident, Wpn3*Resident, Wpn3*Month, NW*LandPub, ES*LandPub, Resident*LandPub, ES*NW, MB*NE, ES*NE, NE*Month, SW*Month, Resident*MB, MBA*Month, MB*Month, ES*Month
Instrumental variable regression (**Y**~ $\mathbf{p}_{rMAR}$, **X**$_s$)	p_{rMAR}, Wpn1, Wpn2, Wpn3, LandPub, NW, NE, SW, MBA, MB, ES, Resident, Month, Wpn1*p_{rMAR}, LandPub*p_{rMAR}, NE*p_{rMAR}, MBA*p_{rMAR}, MB*p_{rMAR}, ES*p_{rMAR}, Month*p_{rMAR}, Wpn2*LandPub, Wpn3*NW, Wpn2*NE, Wpn2*ES, Wpn1*Resident, Wpn2*Resident, Wpn3*Resident, Wpn3*Month, NW*LandPub, SW*LandPub, Resident*LandPub, MBA*NW, ES*NW, MB*NE, ES*NE, Resident*NE, NE*Month, SW*Month, Resident*MB, MBA*Month, MB*Month, ES*Month
Conditional (**Y**~ **R**, **X**$_s$)	TOResp, Wpn1, Wpn2, Wpn3, LandPub, NW, NE, SW, MBA, MB, ES, Resident, Month, TOResp*NW, TOResp*SW, TOResp*MBA, TOResp*Resident, Wpn1*LandPub, Wpn3*LandPub, Wpn3*NW, Wpn2*NE, Wpn3*SW, Wpn2*ES, Wpn1*Resident, Wpn2*Resident, Wpn3*Resident, Wpn2*Month, Wpn3*Month, LandPub*NW, LandPub*ES, NW*ES, NE*MB, NE*ES, NE*Month, SW*Month, MB*Resident, MBA*Month, MB*Month, ES*Month, Resident*Month

Table D.2: Explanatory variables for final response propensity models

Model	Response propensity model variables	
	2001	2003
MAR (WC Adj) (**R** ~ **X**$_r$)	Age1, Age2, Age3, Age4, Size1, Size2, Age2*Size1, Age1*Size2, Age2*Size2	Age1, Age2, Age3, Age4, Size1, Size2, Age1*Size1, Age2*Size1, Age3*Size1, Age1*Size2, Age2*Size2, Age3*Size2, Age4*Size2
MAR (PSAE) (**R** ~ **X**)	Wpn1, Wpn2, Wpn3, LandPub, NW, NE, SW, MBA, MB, ES, Resident, Month, Age1, Age2, Age3, Age4, Size1, Size2, Wpn1*NW, Wpn2*NW, Wpn3*NW, Wpn1*NE, Wpn1*SW, Wpn2*SW, Wpn1*MB, Wpn1*ES, Wpn2*ES, Wpn1*Resident, Wpn2*Resident, Wpn3*Resident, Wpn1*Month, Wpn3*Month, NE*LandPub, SW*LandPub, ES*LandPub, ES*NW, MB*NE, MB*SW, Resident*NW, Resident*NE, NW*Month, Resident*MBA, MBA*Month, MB*Month, LandPub*Age1, SW*Age1, Resident*Age1, Age1*Month, Wpn1*Age2, Wpn2*Age2, Wpn3*Age2, LandPub*Age2, Resident*Age2, Wpn1*Age3, Wpn2*Age3, Wpn3*Age3, LandPub*Age3, ES*Age3, Wpn1*Age4, Wpn2*Age4, Wpn3*Age4, LandPub*Age4, ES*Age4, Wpn1*Size1, Wpn2*Size1, Wpn3*Size1, NE*Size1, MBA*Size1, ES*Size1, Size1*Month, Wpn1*Size2, Wpn2*Size2, NE*Size2, SW*Size2, ES*Size2, Age1*Size2, Age3*Size2	Wpn1, Wpn2, Wpn3, LandPub, NW, NE, SW, MBA, MB, ES, Resident, Month, Age1, Age2, Age3, Age4, Size1, Size2, Wpn2*LandPub, Wpn2*NW, Wpn3*NW, Wpn1*NE, Wpn1*Month, NW*LandPub, SW*LandPub, MB*LandPub, ES*LandPub, Resident*LandPub, MB*NW, MB*NE, ES*NE, MB*SW, NW*Month, NE*Month, Resident*Month, Wpn1*Age1, Wpn2*Age1, Wpn3*Age1, LandPub*Age1, Age1*Month, Wpn2*Age2, MB*Age2, Age2*Month, NE*Age3, Resident*Age3, Age3*Month, SW*Age4, Resident*Age4, Month*Age4, Wpn2*Size1, NE*Size1, SW*Size1, ES*Size1, Size1*Month, NW*Size2, SW*Size2, MBA*Size2, MB*Size2, Age1*Size1

Model	Response propensity model variables	
	2001	**2003**
R (WC Adj) (**R** ~ $\mathbf{X}_r$)	Harv, Age1, Age2, Age3, Age4, Size1, Size2, Age2*Harv, Age3*Harv, Size1*Harv, Size2*Harv, Age2*Size1, Age1*Size2, Age2*Size2	Harv, Age1, Age2, Age3, Size1, Size2, Age2*Harv, Age3*Harv, Harv*Age4, Size1*Harv, Size2*Harv, Age1*Size1, Age2*Size1, Age3*Size1, Age2*Size2, Age3*Size2
R (PSAE) (**R** ~ **Y**, **X**)	Harv, Wpn1, Wpn2, Wpn3, LandPub, NW, NE, SW, MBA, MB, ES, Resident, Month, Age1, Age2, Age3, Age4, Size1, Size2, NW*Harv, Wpn1*LandPub, Wpn1*NW, Wpn2*NW, Wpn3*NW, Wpn1*NE, Wpn1*SW, Wpn2*SW, Wpn1*MB, Wpn1*ES, Wpn2*ES, Wpn1*Month, Wpn3*Month, NE*LandPub, SW*LandPub, ES*LandPub, ES*NW, MB*NE, MB*SW, Resident*NW, Resident*NE, NW*Month, Resident*MBA, MBA*Month, MB*Month, Wpn1*Age1, Wpn2*Age1, Wpn3*Age1, LandPub*Age1, SW*Age1, ES*Age1, Age1*Month, LandPub*Age2, Wpn1*Age3, LandPub*Age3, Wpn1*Age4, LandPub*Age4, Wpn1*Size1, Wpn2*Size1, MBA*Size1, ES*Size1, Size1*Month, Wpn1*Size2, Wpn2*Size2, NE*Size2, SW*Size2, ES*Size2, Age1*Size2, Age3*Size2, Size2*Age4	Harv, Wpn1, Wpn2, Wpn3, LandPub, NW, NE, SW, MBA, MB, ES, Resident, Month, Age1, Age2, Age3, Age4, Size1, Size2, LandPub*Harv, NW*Harv, NE*Harv, Wpn2*LandPub, Wpn1*NW, Wpn1*NE, Wpn3*SW, Wpn1*Month, Wpn3*Month, NW*LandPub, SW*LandPub, MB*LandPub, ES*LandPub, MB*NW, MB*NE, MB*SW, NW*Month, NE*Month, Resident*MB, ES*Month, Wpn1*Age1, Wpn2*Age1, Wpn3*Age1, LandPub*Age1, Age1*Month, Resident*Age3, Resident*Age4, Wpn2*Size1, NE*Size1, SW*Size1, ES*Size1, Size1*Month, NE*Size2, MBA*Size2, MB*Size2, Age1*Size1, Age3*Size2, Size1*Age4

Model	Response propensity model variables	
	2001	2003
CR ($\mathbf{R} \sim \mathbf{p}_s$, $\mathbf{X}$)	p_s, Age1, Age2 , Age3, Age4, Size1, Size2, Age3*p_s, Age4*p_s, Size1*p_s, Size2*p_s, Age1*Size1, Age2*Size1	p_s, Age1, Age2, Age3, Age4, Size1, Size2, Age1*p_s, Age2*p_s, Age3*p_s, Age4*p_s, Size1*p_s, Size2*p_s, Age1*Size1, Age2*Size1, Age3*Size1, Age4*Size1, Age1*Size2, Age2*Size2, Age3*Size2, Age4*Size2
QCR ($\mathbf{R} \sim \mathbf{p}_s^*$, $\mathbf{X}$)	p_s^*, Age1, Age2 , Age3, Age4, Size1, Size2, Age1* p_s^*, Age3* p_s^*, Age4* p_s^*, Size1* p_s^*, Size2* p_s^*, Age1*Size1, Age2*Size1, Age3*Size1, Age4*Size1	p_s^*, Age1, Age2 , Age3, Age4, Size1, Size2, Age1* p_s^*, Age2* p_s^*, Age3* p_s^*, Size1* p_s^*, Size2* p_s^*, Age1*Size1, Age2*Size1, Age3*Size1, Age4*Size1

APPENDIX E: Simulation results (Means across 100 simulated populations with 3 samples each)

Table E.1: Mean relative bias for four PSAE approaches (R, R*, MAR, MAR*)

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			R	R*	MA R	MA R*	R	R*	MA R	MA R*	R	R*	MA R	MA R*	R	R*	MA R	MA R*
3.00	0.3	0.3	-0.06	-0.02	0.32	0.36	-0.24	-0.13	-0.30	-0.26	-0.05	0.00	0.44	0.46	-0.31	-0.23	-0.34	-0.31
2.00	0.3	0.3	-0.05	-0.02	0.19	0.21	-0.25	-0.19	-0.34	-0.32	-0.06	-0.03	0.30	0.31	-0.26	-0.18	-0.35	-0.33
1.00	0.3	0.3	-0.04	-0.04	-0.03	-0.01	-0.30	-0.22	-0.46	-0.43	-0.05	-0.04	0.10	0.12	-0.34	-0.25	-0.47	-0.45
0.50	0.3	0.3	-0.10	-0.11	-0.26	-0.24	-0.26	-0.16	-0.58	-0.56	-0.09	-0.09	-0.17	-0.16	-0.38	-0.28	-0.57	-0.56
0.33	0.3	0.3	-0.10	-0.13	-0.37	-0.35	-0.32	-0.17	-0.64	-0.62	-0.12	-0.14	-0.29	-0.28	-0.36	-0.29	-0.60	-0.58
3.00	0.6	0.3	-0.05	0.00	0.20	0.23	-0.25	-0.16	-0.34	-0.32	-0.03	0.02	0.47	0.48	-0.27	-0.19	-0.28	-0.25
2.00	0.6	0.3	-0.05	-0.01	0.13	0.15	-0.28	-0.23	-0.38	-0.36	-0.03	0.00	0.30	0.31	-0.34	-0.30	-0.37	-0.35
1.00	0.6	0.3	-0.05	-0.03	-0.03	0.00	-0.31	-0.23	-0.46	-0.44	-0.07	-0.05	0.05	0.08	-0.22	-0.18	-0.43	-0.41
0.50	0.6	0.3	-0.07	-0.07	-0.21	-0.19	-0.35	-0.25	-0.55	-0.53	-0.06	-0.08	-0.20	-0.19	-0.42	-0.30	-0.58	-0.56
0.33	0.6	0.3	-0.09	-0.08	-0.32	-0.30	-0.30	-0.22	-0.59	-0.58	-0.08	-0.12	-0.32	-0.31	-0.37	-0.23	-0.62	-0.60
3.00	0.9	0.3	-0.05	-0.03	0.01	0.03	-0.28	-0.18	-0.42	-0.39	-0.04	0.01	0.31	0.32	-0.26	-0.17	-0.32	-0.30
2.00	0.9	0.3	-0.05	-0.03	-0.01	0.02	-0.33	-0.23	-0.45	-0.43	-0.04	0.00	0.25	0.25	-0.27	-0.20	-0.37	-0.34
1.00	0.9	0.3	-0.04	-0.02	-0.02	0.00	-0.25	-0.19	-0.46	-0.44	-0.06	-0.04	0.02	0.03	-0.32	-0.22	-0.47	-0.44
0.50	0.9	0.3	-0.05	-0.04	-0.07	-0.06	-0.27	-0.17	-0.48	-0.46	-0.06	-0.06	-0.12	-0.11	-0.30	-0.19	-0.53	-0.50
0.33	0.9	0.3	-0.06	-0.04	-0.11	-0.08	-0.31	-0.22	-0.50	-0.48	-0.07	-0.07	-0.22	-0.21	-0.37	-0.29	-0.58	-0.56
3.00	0.3	0.6	0.00	0.01	0.19	0.19	0.00	0.04	-0.09	-0.09	0.00	0.02	0.18	0.18	-0.02	0.01	-0.14	-0.14
2.00	0.3	0.6	0.00	0.02	0.13	0.13	-0.01	0.01	-0.17	-0.17	0.01	0.03	0.15	0.14	0.00	0.03	-0.16	-0.16
1.00	0.3	0.6	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.29	-0.29	0.00	0.00	0.00	0.00	-0.02	-0.02	-0.30	-0.30
0.50	0.3	0.6	-0.01	-0.03	-0.14	-0.14	-0.03	-0.06	-0.42	-0.42	0.00	-0.02	-0.10	-0.10	-0.02	-0.03	-0.38	-0.38
0.33	0.3	0.6	-0.02	-0.05	-0.22	-0.22	-0.03	-0.07	-0.47	-0.48	-0.02	-0.05	-0.18	-0.18	-0.06	-0.10	-0.47	-0.47
3.00	0.6	0.6	0.00	0.02	0.15	0.15	0.00	0.02	-0.16	-0.16	0.01	0.02	0.22	0.22	0.00	0.03	-0.10	-0.10
2.00	0.6	0.6	0.00	0.01	0.10	0.10	-0.01	0.01	-0.19	-0.19	0.01	0.02	0.16	0.16	-0.02	0.01	-0.18	-0.18
1.00	0.6	0.6	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.29	-0.30	0.00	0.00	0.02	0.02	-0.01	-0.01	-0.29	-0.29
0.50	0.6	0.6	0.00	-0.02	-0.10	-0.10	-0.03	-0.04	-0.39	-0.39	0.00	-0.03	-0.13	-0.13	-0.03	-0.05	-0.41	-0.41

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			R	**R***	**MA R**	**MA R***	**R**	**R***	**MA R**	**MA R***	**R**	**R***	**MA R**	**MA R***	**R**	**R***	**MA R**	**MA R***
0.33	0.6	0.6	0.00	-0.02	-0.15	-0.15	-0.03	-0.04	-0.44	-0.44	-0.01	-0.04	-0.20	-0.20	-0.05	-0.08	-0.48	-0.48
3.00	0.9	0.6	0.00	0.01	0.04	0.04	-0.01	0.00	-0.27	-0.27	0.01	0.02	0.17	0.16	-0.01	0.02	-0.15	-0.16
2.00	0.9	0.6	0.00	0.00	0.02	0.03	-0.02	-0.01	-0.27	-0.27	0.00	0.01	0.11	0.11	-0.01	0.01	-0.20	-0.20
1.00	0.9	0.6	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.29	-0.29	0.00	0.00	0.02	0.02	-0.01	-0.01	-0.28	-0.28
0.50	0.9	0.6	0.00	0.00	-0.02	-0.02	-0.02	-0.02	-0.31	-0.31	0.00	-0.01	-0.08	-0.08	-0.02	-0.03	-0.38	-0.38
0.33	0.9	0.6	0.00	0.00	-0.03	-0.03	-0.02	-0.02	-0.32	-0.33	0.00	-0.02	-0.13	-0.13	-0.03	-0.03	-0.43	-0.43
3.00	0.3	0.9	0.01	0.01	0.06	0.04	0.01	0.03	-0.13	-0.15	0.02	0.01	0.07	0.04	0.01	0.02	-0.16	-0.18
2.00	0.3	0.9	0.01	0.01	0.04	0.02	0.00	0.01	-0.19	-0.21	0.01	0.00	0.04	0.02	0.02	0.02	-0.19	-0.20
1.00	0.3	0.9	0.01	0.00	0.00	-0.01	0.00	0.00	-0.27	-0.29	0.00	-0.01	0.00	-0.01	0.00	0.00	-0.27	-0.28
0.50	0.3	0.9	0.02	0.00	-0.02	-0.03	0.01	-0.02	-0.36	-0.37	0.02	0.00	-0.01	-0.02	0.00	-0.02	-0.35	-0.36
0.33	0.3	0.9	0.01	-0.02	-0.05	-0.06	-0.01	-0.05	-0.41	-0.42	0.00	-0.03	-0.04	-0.05	-0.01	-0.04	-0.39	-0.40
3.00	0.6	0.9	0.01	0.01	0.05	0.03	0.01	0.03	-0.17	-0.19	0.00	0.00	0.05	0.03	0.00	0.02	-0.14	-0.16
2.00	0.6	0.9	0.01	0.01	0.03	0.02	0.01	0.02	-0.21	-0.23	0.01	0.01	0.05	0.02	0.01	0.02	-0.19	-0.21
1.00	0.6	0.9	0.01	0.00	0.01	-0.01	0.01	0.01	-0.27	-0.29	0.00	-0.01	0.00	-0.02	0.00	0.00	-0.27	-0.29
0.50	0.6	0.9	0.01	-0.01	-0.02	-0.03	0.00	-0.01	-0.33	-0.35	0.00	-0.01	-0.03	-0.04	0.00	-0.02	-0.36	-0.37
0.33	0.6	0.9	0.01	-0.01	-0.03	-0.04	0.00	-0.01	-0.37	-0.38	0.00	-0.03	-0.06	-0.07	-0.01	-0.03	-0.40	-0.41
3.00	0.9	0.9	0.01	0.00	0.02	0.00	0.00	0.01	-0.25	-0.27	0.01	0.02	0.05	0.03	0.00	0.02	-0.18	-0.20
2.00	0.9	0.9	0.01	0.00	0.01	0.00	0.00	0.01	-0.26	-0.27	0.01	0.01	0.03	0.02	0.00	0.01	-0.22	-0.23
1.00	0.9	0.9	0.01	0.00	0.01	-0.01	0.00	0.00	-0.27	-0.29	0.00	0.00	0.01	-0.01	0.01	0.00	-0.27	-0.28
0.50	0.9	0.9	0.01	0.00	0.00	-0.01	0.00	0.00	-0.28	-0.30	0.01	-0.01	-0.01	-0.03	0.00	-0.01	-0.34	-0.35
0.33	0.9	0.9	0.01	0.00	0.00	-0.01	0.00	0.00	-0.29	-0.31	0.01	-0.01	-0.03	-0.04	0.00	-0.01	-0.35	-0.37

Table E.2: Mean relative bias for four PSAE approaches (CR, CR*, QCR, QCR*)

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*
3.00	0.3	0.3	0.04	0.09	0.11	0.15	-0.14	-0.03	-0.07	0.05	-0.04	0.00	0.04	0.08	-0.32	-0.24	-0.23	-0.18
2.00	0.3	0.3	0.02	0.05	0.07	0.10	-0.23	-0.18	-0.16	-0.11	-0.05	-0.01	0.01	0.04	-0.25	-0.19	-0.22	-0.15
1.00	0.3	0.3	-0.03	-0.02	-0.02	0.00	-0.28	-0.17	-0.27	-0.17	-0.02	-0.01	-0.01	0.01	-0.32	-0.15	-0.30	-0.13
0.50	0.3	0.3	-0.16	-0.15	-0.17	-0.16	-0.32	-0.19	-0.33	-0.23	-0.05	-0.06	-0.07	-0.08	-0.35	-0.20	-0.37	-0.20
0.33	0.3	0.3	-0.18	-0.19	-0.22	-0.23	-0.38	-0.24	-0.39	-0.26	-0.10	-0.10	-0.14	-0.11	-0.31	-0.20	-0.32	-0.23
3.00	0.6	0.3	0.10	0.14	0.39	0.24	-0.28	-0.14	-0.26	-0.11	0.15	0.19	0.21	0.21	-0.14	0.09	-0.07	0.08
2.00	0.6	0.3	0.01	0.05	0.05	0.08	-0.19	-0.05	-0.13	-0.08	0.09	0.15	0.12	0.17	-0.31	-0.18	-0.28	-0.18
1.00	0.6	0.3	-0.04	-0.03	-0.03	-0.02	-0.33	-0.23	-0.33	-0.25	-0.06	-0.03	-0.05	-0.03	-0.26	-0.17	-0.26	-0.18
0.50	0.6	0.3	-0.18	-0.12	-0.19	-0.13	-0.45	-0.36	-0.44	-0.36	-0.23	-0.19	-0.24	-0.20	-0.50	-0.42	-0.51	-0.44
0.33	0.6	0.3	-0.35	-0.29	-0.35	-0.30	-0.50	-0.44	-0.49	-0.40	-0.39	-0.35	-0.39	-0.35	-0.55	-0.46	-0.57	-0.52
3.00	0.9	0.3	-0.04	-0.02	-0.03	-0.01	-0.25	-0.15	-0.27	-0.15	0.18	0.21	0.20	0.22	-0.13	0.01	-0.15	0.00
2.00	0.9	0.3	-0.03	-0.01	-0.05	-0.02	-0.31	-0.20	-0.31	-0.20	0.10	0.11	0.10	0.10	-0.15	-0.04	-0.16	-0.03
1.00	0.9	0.3	-0.03	-0.01	-0.04	-0.03	-0.24	-0.16	-0.26	-0.16	-0.04	-0.03	-0.05	-0.03	-0.30	-0.18	-0.30	-0.18
0.50	0.9	0.3	-0.04	-0.03	-0.06	-0.05	-0.25	-0.15	-0.26	-0.16	-0.20	-0.18	-0.20	-0.18	-0.38	-0.28	-0.38	-0.32
0.33	0.9	0.3	-0.05	-0.03	-0.08	-0.06	-0.31	-0.22	-0.33	-0.25	-0.31	-0.29	-0.31	-0.30	-0.52	-0.45	-0.52	-0.45
3.00	0.3	0.6	0.01	0.03	0.05	0.06	0.04	0.07	0.10	0.13	-0.05	-0.03	0.00	0.02	-0.08	-0.04	-0.01	0.02
2.00	0.3	0.6	0.01	0.02	0.04	0.05	0.01	0.04	0.06	0.08	-0.02	-0.01	0.01	0.02	-0.04	-0.02	0.02	0.03
1.00	0.3	0.6	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.01	0.01	0.00	-0.01	0.00	-0.01	-0.01	-0.01	0.01	0.00
0.50	0.3	0.6	-0.02	-0.04	-0.04	-0.05	-0.04	-0.06	-0.06	-0.09	0.05	0.03	0.03	0.01	0.08	0.06	0.05	0.02
0.33	0.3	0.6	-0.06	-0.07	-0.08	-0.09	-0.07	-0.08	-0.10	-0.12	0.06	0.04	0.04	0.01	0.04	0.01	0.00	-0.03
3.00	0.6	0.6	0.01	0.08	0.05	0.04	0.04	0.23	0.12	0.16	0.05	0.13	0.08	0.08	0.11	0.31	0.19	0.26
2.00	0.6	0.6	0.00	0.04	0.03	0.02	0.05	0.16	0.09	0.10	0.04	0.07	0.06	0.06	0.06	0.13	0.10	0.09
1.00	0.6	0.6	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	0.00	0.01	0.00	0.01	-0.01	0.00	-0.01	-0.01
0.50	0.6	0.6	-0.11	-0.09	-0.10	-0.09	-0.17	-0.15	-0.18	-0.15	-0.14	-0.12	-0.14	-0.13	-0.20	-0.17	-0.20	-0.17
0.33	0.6	0.6	-0.22	-0.18	-0.21	-0.18	-0.34	-0.31	-0.33	-0.29	-0.27	-0.23	-0.25	-0.21	-0.39	-0.36	-0.38	-0.32
3.00	0.9	0.6	0.00	0.01	0.01	0.01	-0.01	0.00	-0.01	0.00	0.14	0.14	0.15	0.14	0.18	0.20	0.18	0.20

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*
2.00	0.9	0.6	0.00	0.00	0.00	0.00	-0.02	-0.01	-0.03	-0.02	0.09	0.08	0.09	0.09	0.11	0.11	0.11	0.12
1.00	0.9	0.6	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.02	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01
0.50	0.9	0.6	0.00	0.00	0.00	0.00	-0.02	-0.01	-0.02	-0.01	-0.10	-0.10	-0.10	-0.10	-0.16	-0.15	-0.16	-0.15
0.33	0.9	0.6	0.00	0.00	0.00	0.00	-0.02	-0.01	-0.03	-0.02	-0.15	-0.15	-0.15	-0.15	-0.24	-0.23	-0.24	-0.23
3.00	0.3	0.9	0.02	0.02	0.02	0.02	0.05	0.06	0.08	0.08	0.01	0.00	0.01	0.00	-0.01	0.00	0.02	0.03
2.00	0.3	0.9	0.01	0.01	0.01	0.00	0.04	0.04	0.06	0.06	0.01	0.00	0.00	0.00	0.01	0.01	0.02	0.03
1.00	0.3	0.9	0.01	-0.01	0.01	-0.01	0.00	0.00	0.01	0.00	0.00	-0.02	0.00	-0.02	0.01	0.01	0.01	0.01
0.50	0.3	0.9	0.01	-0.01	0.00	-0.02	-0.04	-0.05	-0.04	-0.06	0.02	-0.01	0.01	-0.02	0.02	0.00	0.01	-0.01
0.33	0.3	0.9	0.00	-0.03	-0.01	-0.04	-0.06	-0.07	-0.07	-0.08	0.01	-0.02	0.00	-0.03	0.06	0.04	0.05	0.02
3.00	0.6	0.9	0.01	0.01	0.02	0.01	0.04	0.08	0.06	0.06	0.01	0.01	0.02	0.01	0.04	0.08	0.06	0.06
2.00	0.6	0.9	0.01	0.00	0.02	0.01	0.03	0.05	0.04	0.04	0.01	0.01	0.02	0.01	0.04	0.05	0.05	0.04
1.00	0.6	0.9	0.00	0.00	0.01	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0.00	-0.01	-0.02	-0.02	-0.02	-0.02
0.50	0.6	0.9	-0.02	-0.02	-0.01	-0.01	-0.12	-0.09	-0.11	-0.10	-0.03	-0.03	-0.02	-0.02	-0.14	-0.13	-0.14	-0.13
0.33	0.6	0.9	-0.05	-0.02	-0.01	0.00	-0.22	-0.16	-0.21	-0.19	-0.08	-0.04	-0.04	-0.03	-0.26	-0.19	-0.25	-0.19
3.00	0.9	0.9	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.04	0.03	0.05	0.03	0.12	0.13	0.13	0.13
2.00	0.9	0.9	0.01	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.03	0.01	0.03	0.02	0.08	0.08	0.08	0.08
1.00	0.9	0.9	0.01	-0.01	0.01	0.00	0.00	-0.01	0.00	0.00	0.00	-0.01	0.00	-0.01	0.00	0.00	0.00	0.00
0.50	0.9	0.9	0.01	0.00	0.01	0.00	0.00	-0.01	0.00	-0.01	-0.02	-0.03	-0.02	-0.03	-0.09	-0.09	-0.09	-0.10
0.33	0.9	0.9	0.00	-0.01	0.01	0.00	-0.01	-0.01	-0.01	-0.01	-0.03	-0.04	-0.03	-0.04	-0.14	-0.14	-0.14	-0.14

Table E3: Mean relative bias for four weighting class adjustment approaches (JC, IVR, R, JR)

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			WC JC	WC IVR	WC R	WC JR	WC JC	WC IVR	WC R	WC JR	WC JC	WC IVR	WC R	WC JR	WC JC	WC IVR	WC R	WC JR
3.00	0.3	0.3	0.23	0.23	0.11	0.06	0.18	0.19	0.22	0.01	0.18	0.20	0.12	0.02	0.13	0.14	0.19	-0.05
2.00	0.3	0.3	0.12	0.13	0.07	0.03	0.11	0.12	0.16	0.00	0.09	0.10	0.10	-0.01	0.09	0.11	0.16	-0.04
1.00	0.3	0.3	-0.04	-0.03	0.05	0.02	-0.02	-0.01	0.03	-0.02	-0.06	-0.03	0.06	0.02	-0.02	-0.01	0.03	-0.01
0.50	0.3	0.3	-0.20	-0.19	-0.04	-0.03	-0.18	-0.18	-0.07	-0.08	-0.18	-0.17	0.01	0.05	-0.18	-0.16	0.00	-0.05
0.33	0.3	0.3	-0.30	-0.28	-0.11	-0.06	-0.26	-0.24	-0.14	-0.11	-0.27	-0.24	-0.05	0.04	-0.22	-0.20	-0.02	-0.04
3.00	0.6	0.3	0.17	0.18	0.10	0.05	0.16	0.17	0.12	0.04	0.25	0.27	0.13	0.08	0.21	0.21	0.24	0.07
2.00	0.6	0.3	0.11	0.12	0.05	0.02	0.11	0.11	0.08	0.03	0.15	0.16	0.06	0.04	0.13	0.14	0.15	0.05
1.00	0.6	0.3	-0.02	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0.00	-0.01	0.00	0.00	0.02	-0.01
0.50	0.6	0.3	-0.15	-0.14	-0.07	-0.05	-0.12	-0.11	-0.11	-0.06	-0.19	-0.19	-0.07	-0.10	-0.16	-0.15	-0.15	-0.08
0.33	0.6	0.3	-0.23	-0.23	-0.10	-0.12	-0.18	-0.18	-0.19	-0.09	-0.29	-0.28	-0.15	-0.15	-0.22	-0.22	-0.19	-0.10
3.00	0.9	0.3	0.04	0.04	0.02	0.02	0.03	0.03	0.00	0.01	0.18	0.18	0.11	0.19	0.15	0.16	0.11	0.16
2.00	0.9	0.3	0.02	0.02	0.01	0.00	0.02	0.02	0.00	0.01	0.11	0.11	0.04	0.11	0.11	0.11	0.09	0.12
1.00	0.9	0.3	0.00	0.00	-0.01	-0.01	0.00	0.00	-0.01	-0.01	0.01	0.01	-0.01	0.01	0.01	0.01	0.00	0.01
0.50	0.9	0.3	-0.03	-0.03	-0.02	-0.01	-0.03	-0.03	-0.03	-0.02	-0.11	-0.11	-0.06	-0.12	-0.10	-0.10	-0.10	-0.10
0.33	0.9	0.3	-0.05	-0.05	-0.02	-0.02	-0.04	-0.04	-0.03	-0.02	-0.20	-0.20	-0.07	-0.20	-0.15	-0.16	-0.16	-0.16
3.00	0.3	0.6	0.16	0.16	0.01	0.01	0.25	0.25	0.08	0.03	0.14	0.14	0.02	0.00	0.19	0.20	0.06	-0.01
2.00	0.3	0.6	0.11	0.11	0.02	0.00	0.14	0.15	0.05	-0.01	0.09	0.09	0.02	0.01	0.13	0.14	0.07	0.00
1.00	0.3	0.6	-0.02	-0.02	0.00	-0.01	-0.01	-0.01	0.01	0.00	-0.03	-0.03	0.00	0.00	-0.02	-0.01	0.01	0.01
0.50	0.3	0.6	-0.15	-0.14	-0.02	-0.02	-0.19	-0.18	-0.04	-0.04	-0.13	-0.13	-0.01	0.05	-0.15	-0.14	-0.02	0.07
0.33	0.3	0.6	-0.22	-0.22	-0.04	-0.06	-0.26	-0.26	-0.08	-0.06	-0.20	-0.20	-0.04	0.05	-0.24	-0.23	-0.07	0.03
3.00	0.6	0.6	0.14	0.14	0.02	0.00	0.18	0.19	0.07	0.02	0.20	0.20	0.03	0.01	0.25	0.25	0.08	0.03
2.00	0.6	0.6	0.09	0.09	0.01	-0.02	0.13	0.13	0.04	0.00	0.12	0.12	0.02	0.01	0.15	0.15	0.04	-0.01
1.00	0.6	0.6	-0.01	-0.01	0.00	-0.02	-0.01	0.00	0.00	-0.01	0.01	0.00	0.00	-0.01	0.00	0.00	0.00	-0.01
0.50	0.6	0.6	-0.10	-0.10	-0.01	-0.05	-0.12	-0.12	-0.03	-0.07	-0.14	-0.14	-0.01	-0.09	-0.17	-0.16	-0.05	-0.09
0.33	0.6	0.6	-0.15	-0.14	-0.01	-0.11	-0.20	-0.20	-0.05	-0.16	-0.20	-0.20	-0.03	-0.17	-0.27	-0.27	-0.11	-0.20
3.00	0.9	0.6	0.03	0.03	0.01	0.00	0.04	0.04	0.02	0.01	0.15	0.15	0.03	0.10	0.18	0.18	0.07	0.17

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			WC JC	**WC IVR**	**WC R**	**WC JR**	**WC JC**	**WC IVR**	**WC R**	**WC JR**	**WC JC**	**WC IVR**	**WC R**	**WC JR**	**WC JC**	**WC IVR**	**WC R**	**WC JR**
2.00	0.9	0.6	0.02	0.02	0.01	0.00	0.02	0.02	0.01	-0.01	0.09	0.09	0.01	0.06	0.11	0.11	0.04	0.10
1.00	0.9	0.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	-0.01	0.01	0.01	0.00	0.00
0.50	0.9	0.6	-0.02	-0.02	0.00	0.00	-0.03	-0.03	-0.01	-0.01	-0.09	-0.09	-0.01	-0.10	-0.12	-0.12	-0.04	-0.13
0.33	0.9	0.6	-0.03	-0.03	-0.01	0.00	-0.05	-0.05	-0.01	-0.02	-0.13	-0.14	-0.01	-0.14	-0.19	-0.19	-0.04	-0.20
3.00	0.3	0.9	0.04	0.04	0.00	-0.01	0.19	0.19	0.02	0.03	0.03	0.03	0.00	-0.01	0.17	0.17	0.02	0.02
2.00	0.3	0.9	0.02	0.02	-0.01	-0.01	0.12	0.12	0.00	0.02	0.01	0.01	-0.01	-0.02	0.12	0.12	0.01	0.01
1.00	0.3	0.9	-0.01	-0.01	-0.01	-0.01	0.01	0.01	0.00	0.00	-0.02	-0.02	-0.03	-0.02	0.01	0.01	-0.01	0.02
0.50	0.3	0.9	-0.03	-0.03	-0.01	-0.01	-0.11	-0.11	-0.01	-0.02	-0.03	-0.03	-0.02	0.00	-0.09	-0.09	-0.02	0.03
0.33	0.3	0.9	-0.05	-0.05	-0.02	-0.02	-0.18	-0.18	-0.04	-0.05	-0.06	-0.06	-0.04	-0.01	-0.14	-0.14	-0.03	0.06
3.00	0.6	0.9	0.03	0.03	0.01	0.00	0.15	0.15	0.03	0.01	0.03	0.03	-0.01	-0.01	0.17	0.17	0.03	0.02
2.00	0.6	0.9	0.02	0.02	0.01	0.00	0.09	0.09	0.01	0.01	0.02	0.03	0.00	0.00	0.12	0.12	0.02	0.01
1.00	0.6	0.9	0.00	0.00	0.01	0.00	0.01	0.01	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	0.00	0.00	0.00	-0.01
0.50	0.6	0.9	-0.02	-0.02	0.00	-0.01	-0.08	-0.08	-0.01	-0.06	-0.03	-0.03	-0.01	-0.02	-0.11	-0.11	-0.01	-0.10
0.33	0.6	0.9	-0.03	-0.03	0.00	-0.01	-0.13	-0.13	-0.01	-0.12	-0.06	-0.06	-0.02	-0.03	-0.19	-0.19	-0.03	-0.18
3.00	0.9	0.9	0.01	0.01	0.01	0.00	0.03	0.03	0.01	0.00	0.04	0.04	0.02	0.03	0.13	0.13	0.02	0.08
2.00	0.9	0.9	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.02	0.02	0.01	0.01	0.08	0.08	0.01	0.05
1.00	0.9	0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00	-0.01
0.50	0.9	0.9	-0.01	0.00	0.00	0.00	-0.02	-0.02	0.00	0.00	-0.02	-0.02	0.00	-0.02	-0.08	-0.08	-0.01	-0.09
0.33	0.9	0.9	-0.01	-0.01	0.00	0.00	-0.03	-0.03	-0.01	0.00	-0.03	-0.03	0.00	-0.03	-0.12	-0.12	-0.01	-0.13

Table E4: Mean relative bias for three weighting class adjustment approaches (CR, QCR, SUCC)

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			CR	**QCR**	**SUCC**	**CR**	**QCR**	**SUCC**	**CR**	**QCR**	**SUCC**	**CR**	**QCR**	**SUCC**
3.00	0.3	0.3	0.15	0.21	0.13	0.17	0.17	0.11	0.06	0.18	0.08	0.11	0.12	0.08
2.00	0.3	0.3	0.10	0.14	0.09	0.14	0.10	0.07	0.04	0.09	0.05	0.10	0.08	0.04
1.00	0.3	0.3	0.00	0.01	0.06	0.02	-0.03	0.04	0.02	-0.06	0.08	0.05	-0.03	0.05
0.50	0.3	0.3	-0.12	-0.13	0.00	-0.14	-0.19	-0.03	-0.02	-0.19	0.10	-0.05	-0.19	-0.01
0.33	0.3	0.3	-0.18	-0.20	-0.03	-0.20	-0.27	-0.09	-0.07	-0.28	0.08	-0.07	-0.23	-0.01
3.00	0.6	0.3	0.12	0.13	0.09	0.14	0.16	0.10	0.19	0.26	0.14	0.19	0.21	0.15
2.00	0.6	0.3	0.08	0.09	0.06	0.10	0.11	0.06	0.12	0.15	0.08	0.13	0.13	0.08
1.00	0.6	0.3	0.00	0.00	0.01	-0.01	-0.01	0.00	0.01	-0.01	0.01	0.01	0.00	0.01
0.50	0.6	0.3	-0.10	-0.11	-0.05	-0.12	-0.12	-0.06	-0.14	-0.19	-0.09	-0.15	-0.16	-0.07
0.33	0.6	0.3	-0.16	-0.18	-0.11	-0.17	-0.18	-0.09	-0.22	-0.29	-0.14	-0.20	-0.22	-0.10
3.00	0.9	0.3	0.03	0.03	0.03	0.03	0.03	0.02	0.23	0.18	0.19	0.19	0.15	0.16
2.00	0.9	0.3	0.02	0.01	0.01	0.02	0.02	0.02	0.14	0.11	0.11	0.15	0.11	0.11
1.00	0.9	0.3	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.01	0.01	0.00	0.01	0.01
0.50	0.9	0.3	-0.01	-0.02	-0.01	-0.03	-0.03	-0.01	-0.15	-0.11	-0.12	-0.13	-0.10	-0.10
0.33	0.9	0.3	-0.02	-0.04	-0.01	-0.04	-0.04	-0.01	-0.26	-0.20	-0.21	-0.19	-0.15	-0.16
3.00	0.3	0.6	0.03	0.06	0.05	0.09	0.22	0.11	-0.03	0.13	0.06	-0.04	0.17	0.06
2.00	0.3	0.6	0.02	0.05	0.04	0.05	0.12	0.05	-0.01	0.08	0.06	0.00	0.10	0.08
1.00	0.3	0.6	-0.01	-0.01	0.01	0.01	-0.03	0.05	0.00	-0.03	0.03	0.00	-0.05	0.08
0.50	0.3	0.6	-0.04	-0.05	0.01	-0.07	-0.22	0.02	0.03	-0.13	0.09	0.04	-0.18	0.14
0.33	0.3	0.6	-0.07	-0.09	-0.02	-0.10	-0.29	0.00	0.03	-0.20	0.11	-0.02	-0.27	0.12
3.00	0.6	0.6	0.05	0.05	0.05	0.09	0.18	0.08	0.09	0.19	0.09	0.13	0.25	0.11
2.00	0.6	0.6	0.02	0.02	0.03	0.05	0.12	0.05	0.05	0.12	0.05	0.06	0.15	0.05
1.00	0.6	0.6	0.00	-0.01	0.00	0.00	-0.01	0.00	0.01	0.00	0.00	0.00	-0.01	0.01
0.50	0.6	0.6	-0.04	-0.04	-0.03	-0.07	-0.13	-0.04	-0.07	-0.14	-0.06	-0.10	-0.18	-0.06
0.33	0.6	0.6	-0.07	-0.07	-0.06	-0.14	-0.21	-0.11	-0.12	-0.20	-0.11	-0.19	-0.28	-0.15
3.00	0.9	0.6	0.02	0.02	0.02	0.03	0.03	0.02	0.15	0.15	0.15	0.21	0.18	0.19
2.00	0.9	0.6	0.01	0.01	0.01	0.01	0.02	0.01	0.10	0.09	0.09	0.13	0.12	0.12

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			CR	QCR	SUCC	CR	QCR	SUCC	CR	QCR	SUCC	CR	QCR	SUCC
1.00	0.9	0.6	0.00	0.00	0.01	0.00	-0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00
0.50	0.9	0.6	0.00	0.00	0.00	-0.01	-0.03	0.00	-0.10	-0.09	-0.10	-0.15	-0.12	-0.13
0.33	0.9	0.6	0.00	0.00	0.00	-0.01	-0.05	-0.01	-0.15	-0.14	-0.15	-0.22	-0.19	-0.20
3.00	0.3	0.9	0.02	0.02	0.01	0.06	0.16	0.07	0.01	0.03	0.00	0.00	0.13	0.06
2.00	0.3	0.9	0.01	0.01	0.00	0.04	0.10	0.05	0.00	0.01	0.00	0.01	0.08	0.05
1.00	0.3	0.9	0.00	0.00	0.00	0.00	-0.01	0.03	-0.01	-0.02	0.00	0.01	-0.02	0.05
0.50	0.3	0.9	-0.01	-0.01	0.00	-0.04	-0.13	0.00	0.00	-0.03	0.01	0.00	-0.12	0.07
0.33	0.3	0.9	-0.02	-0.03	-0.01	-0.07	-0.21	-0.02	-0.01	-0.06	0.00	0.04	-0.17	0.11
3.00	0.6	0.9	0.02	0.02	0.01	0.07	0.14	0.07	0.01	0.03	0.00	0.08	0.16	0.07
2.00	0.6	0.9	0.01	0.01	0.01	0.04	0.09	0.04	0.01	0.02	0.00	0.06	0.11	0.05
1.00	0.6	0.9	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
0.50	0.6	0.9	-0.01	-0.01	-0.01	-0.05	-0.09	-0.05	-0.02	-0.03	-0.02	-0.08	-0.12	-0.07
0.33	0.6	0.9	-0.01	-0.02	-0.01	-0.08	-0.14	-0.07	-0.04	-0.06	-0.04	-0.12	-0.20	-0.12
3.00	0.9	0.9	0.01	0.01	0.01	0.02	0.03	0.02	0.04	0.04	0.03	0.14	0.13	0.13
2.00	0.9	0.9	0.00	0.00	0.00	0.01	0.02	0.02	0.02	0.02	0.02	0.09	0.08	0.08
1.00	0.9	0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00
0.50	0.9	0.9	0.00	0.00	0.00	-0.01	-0.02	0.00	-0.03	-0.02	-0.02	-0.09	-0.08	-0.10
0.33	0.9	0.9	0.00	0.00	0.00	-0.01	-0.03	0.00	-0.04	-0.03	-0.04	-0.14	-0.12	-0.14

Table E5: Mean relative bias for three weighting class adjustment approaches (INT, JCR, JQCR)

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			INT	JCR	JQCR	INT	JCR	JQCR	INT	JCR	JQCR	INT	JCR	JQCR
3.00	0.3	0.3	0.08	0.08	0.08	0.03	0.02	0.02	0.03	0.04	0.03	-0.04	-0.04	-0.03
2.00	0.3	0.3	0.04	0.04	0.04	0.02	0.01	0.01	0.00	0.00	0.00	-0.03	-0.04	-0.04
1.00	0.3	0.3	0.02	0.02	0.02	-0.01	-0.02	-0.02	0.01	0.02	0.01	0.00	-0.01	-0.01
0.50	0.3	0.3	-0.04	-0.03	-0.04	-0.07	-0.07	-0.08	0.05	0.05	0.05	-0.04	-0.05	-0.06
0.33	0.3	0.3	-0.07	-0.06	-0.06	-0.11	-0.11	-0.11	0.03	0.04	0.04	-0.03	-0.04	-0.04
3.00	0.6	0.3	0.08	0.08	0.08	0.07	0.07	0.08	0.13	0.13	0.13	0.11	0.12	0.13
2.00	0.6	0.3	0.03	0.04	0.04	0.05	0.05	0.05	0.06	0.06	0.06	0.06	0.06	0.06
1.00	0.6	0.3	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0.00	-0.01	-0.01
0.50	0.6	0.3	-0.06	-0.05	-0.05	-0.06	-0.06	-0.06	-0.10	-0.09	-0.09	-0.07	-0.08	-0.08
0.33	0.6	0.3	-0.11	-0.11	-0.10	-0.09	-0.09	-0.09	-0.14	-0.14	-0.14	-0.09	-0.10	-0.10
3.00	0.9	0.3	0.02	0.02	0.02	0.01	0.01	0.01	0.20	0.20	0.20	0.17	0.17	0.17
2.00	0.9	0.3	0.01	0.01	0.01	0.01	0.01	0.01	0.12	0.12	0.11	0.12	0.12	0.12
1.00	0.9	0.3	0.00	0.00	0.00	-0.01	-0.01	-0.01	0.01	0.01	0.01	0.01	0.01	0.01
0.50	0.9	0.3	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02	-0.12	-0.12	-0.12	-0.10	-0.10	-0.10
0.33	0.9	0.3	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.21	-0.21	-0.21	-0.16	-0.16	-0.16
3.00	0.3	0.6	0.03	0.03	0.03	0.07	0.06	0.06	0.02	0.01	0.02	0.03	0.00	0.01
2.00	0.3	0.6	0.02	0.02	0.02	0.03	0.01	0.01	0.02	0.01	0.02	0.04	0.01	0.01
1.00	0.3	0.6	-0.02	-0.01	-0.01	0.01	-0.01	0.00	0.00	0.00	0.00	0.03	0.02	0.02
0.50	0.3	0.6	-0.04	-0.03	-0.03	-0.02	-0.05	-0.05	0.02	0.04	0.04	0.08	0.06	0.05
0.33	0.3	0.6	-0.08	-0.07	-0.06	-0.04	-0.06	-0.06	0.01	0.03	0.04	0.06	0.03	0.03
3.00	0.6	0.6	0.05	0.05	0.05	0.07	0.07	0.08	0.09	0.08	0.08	0.10	0.10	0.10
2.00	0.6	0.6	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.04	0.05	0.04	0.03	0.03
1.00	0.6	0.6	-0.01	-0.01	-0.01	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.50	0.6	0.6	-0.04	-0.03	-0.03	-0.05	-0.04	-0.05	-0.06	-0.07	-0.06	-0.06	-0.07	-0.07
0.33	0.6	0.6	-0.06	-0.06	-0.06	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.15	-0.15	-0.14
3.00	0.9	0.6	0.02	0.02	0.02	0.01	0.01	0.01	0.15	0.15	0.15	0.19	0.19	0.19
2.00	0.9	0.6	0.01	0.01	0.01	0.00	0.00	0.00	0.09	0.09	0.09	0.12	0.12	0.12

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			INT	JCR	JQCR	INT	JCR	JQCR	INT	JCR	JQCR	INT	JCR	JQCR
1.00	0.9	0.6	0.00	0.00	0.00	0.00	-0.01	0.00	-0.01	-0.01	-0.01	0.00	0.00	0.00
0.50	0.9	0.6	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.10	-0.10	-0.10	-0.13	-0.13	-0.13
0.33	0.9	0.6	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.15	-0.15	-0.15	-0.20	-0.20	-0.20
3.00	0.3	0.9	0.00	0.00	0.00	0.07	0.06	0.06	-0.01	-0.01	-0.01	0.05	0.03	0.03
2.00	0.3	0.9	-0.01	-0.01	-0.01	0.05	0.04	0.04	-0.01	-0.01	-0.01	0.05	0.03	0.02
1.00	0.3	0.9	-0.01	-0.01	-0.01	0.02	0.00	0.00	-0.02	-0.02	-0.02	0.03	0.01	0.02
0.50	0.3	0.9	-0.01	-0.01	-0.01	-0.02	-0.03	-0.03	-0.01	-0.01	-0.01	0.04	0.02	0.03
0.33	0.3	0.9	-0.03	-0.03	-0.03	-0.05	-0.06	-0.06	-0.02	-0.02	-0.02	0.07	0.05	0.06
3.00	0.6	0.9	0.01	0.01	0.01	0.07	0.07	0.07	0.00	0.00	0.00	0.08	0.07	0.07
2.00	0.6	0.9	0.00	0.00	0.00	0.04	0.04	0.04	0.00	0.00	0.00	0.05	0.05	0.05
1.00	0.6	0.9	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
0.50	0.6	0.9	-0.01	-0.01	-0.01	-0.05	-0.05	-0.05	-0.02	-0.02	-0.02	-0.07	-0.07	-0.07
0.33	0.6	0.9	-0.01	-0.01	-0.01	-0.08	-0.07	-0.07	-0.04	-0.04	-0.04	-0.11	-0.12	-0.12
3.00	0.9	0.9	0.00	0.00	0.01	0.02	0.02	0.02	0.03	0.03	0.03	0.13	0.13	0.13
2.00	0.9	0.9	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.02	0.02	0.08	0.08	0.08
1.00	0.9	0.9	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
0.50	0.9	0.9	0.00	0.00	0.00	0.00	0.00	0.00	-0.02	-0.02	-0.02	-0.10	-0.10	-0.10
0.33	0.9	0.9	0.00	0.00	0.00	0.00	0.00	0.00	-0.04	-0.04	-0.04	-0.14	-0.14	-0.14

Table E6: Mean relative bias of MCAR, MCAR2, and DSS estimates

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			MCAR	MCAR2	DSS	MCAR	MCAR2	DSS	MCAR	MCAR2	DSS	MCAR	MCAR2	DSS
3.00	0.3	0.3	0.28	0.28	0.00	0.21	0.21	0.00	0.25	0.25	0.01	0.18	0.18	0.00
2.00	0.3	0.3	0.17	0.17	0.00	0.14	0.14	0.00	0.15	0.15	0.00	0.13	0.12	0.01
1.00	0.3	0.3	0.00	0.01	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.00
0.50	0.3	0.3	-0.15	-0.15	0.00	-0.16	-0.16	0.00	-0.13	-0.13	0.01	-0.14	-0.14	0.00
0.33	0.3	0.3	-0.26	-0.26	0.00	-0.23	-0.23	0.00	-0.21	-0.21	0.00	-0.19	-0.19	0.01
3.00	0.6	0.3	0.18	0.18	-0.01	0.17	0.17	0.00	0.27	0.27	0.00	0.22	0.22	-0.01
2.00	0.6	0.3	0.13	0.13	0.00	0.11	0.11	0.00	0.16	0.16	0.00	0.14	0.14	0.00
1.00	0.6	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.01
0.50	0.6	0.3	-0.13	-0.13	0.00	-0.11	-0.11	0.00	-0.17	-0.17	0.00	-0.15	-0.15	0.00
0.33	0.6	0.3	-0.21	-0.21	-0.01	-0.17	-0.17	0.00	-0.26	-0.26	0.00	-0.21	-0.21	-0.01
3.00	0.9	0.3	0.04	0.04	0.00	0.03	0.03	0.00	0.18	0.18	0.00	0.15	0.15	0.00
2.00	0.9	0.3	0.02	0.02	0.00	0.02	0.02	0.00	0.11	0.11	0.00	0.11	0.11	0.00
1.00	0.9	0.3	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.00
0.50	0.9	0.3	-0.03	-0.03	0.00	-0.03	-0.03	0.00	-0.11	-0.11	0.00	-0.10	-0.10	0.00
0.33	0.9	0.3	-0.04	-0.04	0.00	-0.03	-0.03	0.00	-0.19	-0.19	0.00	-0.15	-0.15	0.00
3.00	0.3	0.6	0.18	0.18	0.00	0.26	0.26	0.00	0.16	0.16	0.00	0.21	0.21	0.00
2.00	0.3	0.6	0.12	0.12	0.00	0.15	0.16	-0.01	0.11	0.11	0.01	0.16	0.16	0.01
1.00	0.3	0.6	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00
0.50	0.3	0.6	-0.13	-0.13	0.00	-0.18	-0.18	0.00	-0.11	-0.11	0.00	-0.13	-0.13	0.01
0.33	0.3	0.6	-0.20	-0.20	0.00	-0.25	-0.25	0.00	-0.18	-0.18	0.00	-0.23	-0.23	0.00
3.00	0.6	0.6	0.14	0.14	0.00	0.19	0.19	0.00	0.20	0.20	0.01	0.25	0.25	0.00
2.00	0.6	0.6	0.10	0.10	0.00	0.13	0.13	0.00	0.12	0.12	0.00	0.15	0.16	-0.01
1.00	0.6	0.6	-0.01	-0.01	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00
0.50	0.6	0.6	-0.09	-0.09	0.00	-0.12	-0.12	0.00	-0.13	-0.13	0.00	-0.16	-0.16	0.00
0.33	0.6	0.6	-0.14	-0.14	0.00	-0.20	-0.20	0.00	-0.19	-0.18	0.00	-0.27	-0.27	-0.01
3.00	0.9	0.6	0.03	0.03	0.00	0.04	0.04	0.00	0.14	0.14	0.00	0.18	0.18	0.00

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			MCAR	MCAR2	DSS	MCAR	MCAR2	DSS	MCAR	MCAR2	DSS	MCAR	MCAR2	DSS
2.00	0.9	0.6	0.02	0.02	0.00	0.02	0.02	0.00	0.09	0.09	0.00	0.11	0.11	0.00
1.00	0.9	0.6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00
0.50	0.9	0.6	-0.02	-0.02	0.00	-0.03	-0.03	0.00	-0.09	-0.09	0.00	-0.12	-0.12	0.00
0.33	0.9	0.6	-0.03	-0.03	0.00	-0.04	-0.04	0.00	-0.13	-0.14	0.00	-0.18	-0.18	0.00
3.00	0.3	0.9	0.04	0.04	0.00	0.19	0.18	0.01	0.04	0.04	0.01	0.16	0.16	0.00
2.00	0.3	0.9	0.03	0.03	0.00	0.11	0.12	0.00	0.02	0.02	0.00	0.11	0.12	0.01
1.00	0.3	0.9	0.00	0.00	0.00	0.00	-0.01	0.00	-0.01	-0.01	-0.01	0.00	-0.01	0.00
0.50	0.3	0.9	-0.02	-0.02	0.01	-0.11	-0.12	0.01	-0.02	-0.02	0.01	-0.11	-0.10	0.00
0.33	0.3	0.9	-0.04	-0.04	0.00	-0.19	-0.19	0.00	-0.04	-0.04	0.00	-0.15	-0.16	0.01
3.00	0.6	0.9	0.03	0.03	0.00	0.14	0.14	0.00	0.03	0.03	-0.01	0.17	0.17	0.00
2.00	0.6	0.9	0.02	0.02	0.00	0.09	0.09	0.00	0.03	0.03	0.00	0.12	0.12	0.00
1.00	0.6	0.9	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	0.00
0.50	0.6	0.9	-0.02	-0.02	0.00	-0.08	-0.08	0.00	-0.03	-0.03	0.00	-0.11	-0.12	0.00
0.33	0.6	0.9	-0.03	-0.03	0.00	-0.13	-0.13	0.00	-0.05	-0.05	-0.01	-0.19	-0.19	0.00
3.00	0.9	0.9	0.01	0.01	0.00	0.03	0.03	0.00	0.03	0.03	0.00	0.13	0.13	0.00
2.00	0.9	0.9	0.00	0.00	0.00	0.02	0.02	0.00	0.02	0.02	0.00	0.08	0.08	0.00
1.00	0.9	0.9	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00	0.00
0.50	0.9	0.9	0.00	0.00	0.00	-0.02	-0.02	0.00	-0.02	-0.02	0.00	-0.08	-0.08	0.00
0.33	0.9	0.9	-0.01	-0.01	0.00	-0.03	-0.03	0.00	-0.03	-0.03	0.00	-0.12	-0.12	0.00

Table E7: Mean confidence interval coverage for four PSAE approaches (R, R*, MAR, MAR*)

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			R	R*	MAR	MAR*	R	R*	MAR	MAR*	R	R*	MAR	MAR*	R	R*	MAR	MAR*
3.00	0.3	0.3	0.92	0.93	0.94	0.93	0.40	0.46	0.46	0.52	0.91	0.95	0.91	0.91	0.40	0.51	0.48	0.55
2.00	0.3	0.3	0.92	0.94	0.96	0.96	0.37	0.47	0.37	0.41	0.91	0.93	0.93	0.93	0.36	0.51	0.42	0.47
1.00	0.3	0.3	0.90	0.92	0.88	0.89	0.30	0.47	0.20	0.25	0.89	0.91	0.85	0.86	0.32	0.50	0.25	0.27
0.50	0.3	0.3	0.86	0.88	0.60	0.62	0.31	0.57	0.08	0.09	0.88	0.88	0.63	0.65	0.29	0.50	0.13	0.15
0.33	0.3	0.3	0.82	0.85	0.44	0.47	0.27	0.56	0.04	0.05	0.84	0.85	0.50	0.51	0.31	0.56	0.10	0.11
3.00	0.6	0.3	0.94	0.96	1.00	0.99	0.43	0.43	0.26	0.28	0.94	0.96	0.78	0.76	0.47	0.46	0.43	0.50
2.00	0.6	0.3	0.96	0.98	0.99	0.99	0.40	0.43	0.20	0.22	0.95	0.99	0.93	0.91	0.37	0.43	0.25	0.32
1.00	0.6	0.3	0.95	0.97	0.95	0.95	0.44	0.50	0.09	0.11	0.91	0.90	0.89	0.88	0.35	0.50	0.14	0.17
0.50	0.6	0.3	0.90	0.92	0.68	0.71	0.30	0.46	0.05	0.07	0.88	0.91	0.62	0.65	0.26	0.47	0.04	0.05
0.33	0.6	0.3	0.85	0.85	0.40	0.45	0.41	0.52	0.00	0.02	0.82	0.84	0.43	0.43	0.31	0.52	0.03	0.04
3.00	0.9	0.3	0.99	0.99	0.99	0.99	0.32	0.51	0.08	0.08	0.95	0.97	0.85	0.82	0.38	0.42	0.24	0.28
2.00	0.9	0.3	0.96	0.97	0.98	0.98	0.37	0.53	0.04	0.07	0.96	0.97	0.89	0.88	0.42	0.49	0.17	0.24
1.00	0.9	0.3	0.98	0.98	0.98	0.98	0.38	0.54	0.04	0.06	0.94	0.96	0.85	0.88	0.37	0.49	0.06	0.11
0.50	0.9	0.3	0.97	0.96	0.95	0.95	0.35	0.47	0.04	0.05	0.91	0.92	0.73	0.71	0.37	0.46	0.04	0.05
0.33	0.9	0.3	0.97	0.98	0.92	0.90	0.39	0.50	0.02	0.05	0.88	0.89	0.61	0.63	0.26	0.40	0.02	0.03
3.00	0.3	0.6	0.99	1.00	0.73	0.75	1.00	1.00	0.84	0.85	0.99	0.99	0.87	0.87	0.99	0.99	0.78	0.76
2.00	0.3	0.6	0.99	1.00	0.90	0.90	1.00	1.00	0.63	0.63	0.99	1.00	0.88	0.91	0.99	1.00	0.74	0.72
1.00	0.3	0.6	1.00	1.00	0.97	0.97	0.99	1.00	0.22	0.23	0.99	0.99	0.96	0.96	0.98	0.99	0.35	0.34
0.50	0.3	0.6	1.00	0.99	0.77	0.77	0.99	0.98	0.01	0.01	1.00	1.00	0.84	0.84	0.99	0.99	0.15	0.14
0.33	0.3	0.6	0.99	0.98	0.48	0.45	0.99	0.99	0.01	0.01	1.00	0.99	0.63	0.63	0.97	0.97	0.05	0.04
3.00	0.6	0.6	1.00	1.00	0.66	0.67	1.00	1.00	0.48	0.48	1.00	1.00	0.50	0.50	1.00	1.00	0.80	0.78
2.00	0.6	0.6	1.00	1.00	0.91	0.91	1.00	1.00	0.34	0.32	0.99	0.99	0.69	0.70	1.00	1.00	0.49	0.51
1.00	0.6	0.6	1.00	1.00	0.99	0.98	1.00	1.00	0.07	0.06	1.00	1.00	0.95	0.94	1.00	1.00	0.14	0.13
0.50	0.6	0.6	1.00	1.00	0.82	0.81	0.99	1.00	0.00	0.00	1.00	1.00	0.64	0.64	0.99	0.99	0.00	0.00
0.33	0.6	0.6	1.00	1.00	0.52	0.53	0.99	0.99	0.00	0.00	1.00	0.99	0.45	0.42	0.99	0.98	0.00	0.00
3.00	0.9	0.6	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.02	1.00	1.00	0.53	0.56	1.00	1.00	0.49	0.48

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			R	R*	MA R	MA R*	R	R*	MA R	MA R*	R	R*	MA R	MA R*	R	R*	MA R	MA R*
2.00	0.9	0.6	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.03	1.00	1.00	0.74	0.74	1.00	1.00	0.30	0.29
1.00	0.9	0.6	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.02	1.00	1.00	0.93	0.94	1.00	1.00	0.07	0.06
0.50	0.9	0.6	1.00	1.00	1.00	1.00	1.00	1.00	0.02	0.01	1.00	1.00	0.77	0.77	1.00	1.00	0.01	0.01
0.33	0.9	0.6	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	0.57	0.57	1.00	1.00	0.00	0.00
3.00	0.3	0.9	0.99	0.99	0.98	0.99	1.00	1.00	0.65	0.53	0.97	0.96	0.93	0.95	0.99	0.99	0.65	0.60
2.00	0.3	0.9	0.98	0.97	0.97	0.97	1.00	1.00	0.38	0.27	0.98	0.98	0.95	0.96	0.99	0.99	0.54	0.46
1.00	0.3	0.9	0.99	0.99	0.98	0.98	0.99	0.99	0.12	0.08	0.97	0.96	0.95	0.95	0.99	0.99	0.27	0.18
0.50	0.3	0.9	1.00	1.00	0.99	0.98	1.00	0.99	0.02	0.02	0.97	0.96	0.96	0.95	0.99	0.99	0.07	0.04
0.33	0.3	0.9	0.99	0.97	0.93	0.90	0.99	0.99	0.00	0.00	0.99	0.97	0.90	0.90	0.99	0.99	0.04	0.03
3.00	0.6	0.9	1.00	0.99	0.96	0.97	1.00	1.00	0.18	0.07	0.98	0.98	0.92	0.95	1.00	1.00	0.49	0.34
2.00	0.6	0.9	1.00	1.00	0.98	1.00	1.00	1.00	0.07	0.02	1.00	1.00	0.96	0.98	1.00	1.00	0.26	0.15
1.00	0.6	0.9	1.00	1.00	1.00	0.99	1.00	1.00	0.01	0.00	0.99	0.98	0.94	0.94	1.00	1.00	0.03	0.01
0.50	0.6	0.9	1.00	1.00	0.99	0.97	1.00	1.00	0.01	0.00	1.00	0.99	0.95	0.94	1.00	1.00	0.01	0.00
0.33	0.6	0.9	0.99	0.98	0.96	0.92	1.00	1.00	0.00	0.00	0.99	0.95	0.85	0.84	1.00	0.99	0.00	0.00
3.00	0.9	0.9	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.99	0.99	0.91	0.95	1.00	1.00	0.15	0.05
2.00	0.9	0.9	1.00	1.00	1.00	1.00	1.00	1.00	0.01	0.00	0.99	1.00	0.95	0.97	1.00	1.00	0.05	0.02
1.00	0.9	0.9	1.00	1.00	1.00	1.00	1.00	1.00	0.01	0.00	1.00	1.00	0.96	0.95	1.00	1.00	0.02	0.00
0.50	0.9	0.9	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00	0.99	0.95	0.94	1.00	1.00	0.01	0.00
0.33	0.9	0.9	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	0.95	0.91	1.00	1.00	0.00	0.00

Table E8: Mean confidence interval coverage for four PSAE approaches (CR, CR*, QCR, QCR*)

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*
3.00	0.3	0.3	0.97	0.96	0.98	0.98	0.46	0.59	0.54	0.64	0.81	0.81	0.89	0.92	0.40	0.47	0.46	0.57
2.00	0.3	0.3	0.93	0.95	0.95	0.95	0.45	0.57	0.50	0.64	0.83	0.84	0.88	0.90	0.38	0.52	0.43	0.55
1.00	0.3	0.3	0.91	0.93	0.90	0.94	0.35	0.55	0.35	0.54	0.82	0.84	0.85	0.86	0.38	0.60	0.40	0.61
0.50	0.3	0.3	0.75	0.78	0.74	0.76	0.32	0.55	0.30	0.51	0.82	0.83	0.79	0.80	0.38	0.56	0.34	0.56
0.33	0.3	0.3	0.68	0.72	0.61	0.66	0.22	0.47	0.21	0.43	0.80	0.80	0.76	0.80	0.36	0.62	0.33	0.56
3.00	0.6	0.3	0.91	0.91	0.89	0.92	0.43	0.63	0.48	0.60	0.89	0.92	0.89	0.91	0.52	0.65	0.53	0.68
2.00	0.6	0.3	0.95	0.95	0.96	0.96	0.45	0.64	0.45	0.61	0.92	0.91	0.91	0.93	0.42	0.64	0.44	0.64
1.00	0.6	0.3	0.95	0.95	0.94	0.95	0.42	0.50	0.43	0.48	0.90	0.88	0.89	0.90	0.37	0.49	0.37	0.47
0.50	0.6	0.3	0.70	0.75	0.68	0.76	0.22	0.27	0.23	0.29	0.60	0.69	0.60	0.61	0.20	0.28	0.19	0.26
0.33	0.6	0.3	0.39	0.45	0.38	0.44	0.17	0.26	0.19	0.29	0.35	0.43	0.36	0.41	0.11	0.21	0.12	0.17
3.00	0.9	0.3	0.98	0.97	0.94	0.97	0.37	0.56	0.35	0.53	0.99	0.99	1.00	1.00	0.61	0.82	0.61	0.79
2.00	0.9	0.3	0.97	0.98	0.95	0.95	0.41	0.59	0.37	0.60	0.99	0.99	0.99	0.99	0.63	0.81	0.60	0.78
1.00	0.9	0.3	0.98	0.99	0.97	0.98	0.45	0.60	0.41	0.59	0.93	0.95	0.92	0.93	0.47	0.54	0.45	0.56
0.50	0.9	0.3	0.98	0.98	0.97	0.97	0.41	0.51	0.34	0.50	0.66	0.68	0.66	0.70	0.30	0.35	0.28	0.36
0.33	0.9	0.3	0.97	0.96	0.95	0.93	0.48	0.55	0.40	0.52	0.44	0.45	0.42	0.45	0.17	0.22	0.16	0.21
3.00	0.3	0.6	0.99	0.99	0.99	0.98	0.94	0.95	0.95	0.94	0.92	0.93	0.96	0.96	0.88	0.90	0.92	0.91
2.00	0.3	0.6	0.99	0.99	0.99	0.99	0.97	0.97	0.98	0.99	0.93	0.95	0.95	0.96	0.90	0.92	0.95	0.96
1.00	0.3	0.6	1.00	0.99	0.99	0.99	0.95	0.95	0.97	0.97	0.95	0.94	0.93	0.93	0.94	0.93	0.93	0.93
0.50	0.3	0.6	0.99	0.99	0.98	0.96	0.95	0.93	0.91	0.91	0.97	0.97	0.97	0.96	0.97	0.97	0.96	0.97
0.33	0.3	0.6	0.96	0.95	0.94	0.93	0.91	0.91	0.90	0.88	0.97	0.96	0.95	0.94	0.96	0.95	0.94	0.94
3.00	0.6	0.6	0.93	0.94	0.94	0.95	0.88	0.91	0.86	0.88	0.89	0.87	0.87	0.87	0.87	0.88	0.87	0.86
2.00	0.6	0.6	0.99	0.98	0.99	0.99	0.94	0.95	0.96	0.98	0.95	0.94	0.94	0.95	0.92	0.92	0.93	0.93
1.00	0.6	0.6	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.98	0.98	0.99	0.97	0.98	0.98	0.96	0.97	0.98
0.50	0.6	0.6	0.73	0.79	0.73	0.82	0.65	0.71	0.62	0.69	0.67	0.73	0.67	0.69	0.65	0.71	0.64	0.69
0.33	0.6	0.6	0.37	0.47	0.38	0.51	0.23	0.30	0.19	0.30	0.25	0.40	0.30	0.42	0.22	0.28	0.20	0.25
3.00	0.9	0.6	1.00	0.99	1.00	1.00	0.98	0.99	0.98	0.98	0.77	0.73	0.72	0.72	0.91	0.86	0.92	0.87

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			CR	**CR***	**QCR**	**QCR***	**CR**	**CR***	**QCR**	**QCR***	**CR**	**CR***	**QCR**	**QCR***	**CR**	**CR***	**QCR**	**QCR***
2.00	0.9	0.6	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.98	0.94	0.95	0.93	0.94	0.98	0.98	0.99	0.98
1.00	0.9	0.6	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
0.50	0.9	0.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.81	0.80	0.78	0.70	0.72	0.70	0.75
0.33	0.9	0.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.53	0.52	0.51	0.51	0.37	0.41	0.36	0.38
3.00	0.3	0.9	1.00	0.99	0.99	0.99	0.91	0.90	0.89	0.89	0.95	0.95	0.96	0.97	0.89	0.90	0.92	0.93
2.00	0.3	0.9	0.98	0.98	0.98	0.97	0.93	0.92	0.93	0.92	0.96	0.95	0.95	0.95	0.92	0.93	0.95	0.94
1.00	0.3	0.9	0.98	0.99	0.99	0.98	0.91	0.93	0.94	0.93	0.96	0.96	0.97	0.96	0.89	0.89	0.90	0.90
0.50	0.3	0.9	1.00	0.99	1.00	0.99	0.88	0.88	0.88	0.88	0.97	0.95	0.98	0.95	0.92	0.92	0.91	0.91
0.33	0.3	0.9	0.98	0.96	0.98	0.94	0.92	0.89	0.86	0.86	0.98	0.96	0.97	0.94	0.93	0.92	0.93	0.91
3.00	0.6	0.9	1.00	0.99	0.99	0.99	0.88	0.89	0.88	0.88	0.95	0.97	0.96	0.97	0.88	0.89	0.88	0.87
2.00	0.6	0.9	0.99	0.99	1.00	1.00	0.92	0.92	0.91	0.90	0.98	0.99	0.98	0.99	0.91	0.90	0.92	0.91
1.00	0.6	0.9	0.99	1.00	1.00	1.00	0.95	0.97	0.96	0.96	0.96	0.96	0.96	0.97	0.95	0.97	0.97	0.96
0.50	0.6	0.9	0.94	0.98	1.00	0.98	0.72	0.80	0.73	0.74	0.95	0.96	0.98	0.96	0.64	0.71	0.66	0.69
0.33	0.6	0.9	0.81	0.92	0.98	0.96	0.36	0.49	0.35	0.43	0.78	0.88	0.88	0.91	0.32	0.38	0.31	0.38
3.00	0.9	0.9	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.98	0.98	0.97	0.98	0.80	0.78	0.78	0.77
2.00	0.9	0.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99	0.98	0.99	0.97	0.96	0.96	0.95
1.00	0.9	0.9	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.99	0.99	1.00	0.99	0.99	1.00	1.00	0.99
0.50	0.9	0.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.95	0.97	0.96	0.83	0.84	0.83	0.81
0.33	0.9	0.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.94	0.96	0.94	0.55	0.56	0.55	0.53

Table E9: Mean confidence interval coverage for four weighting class adjustment approaches (JC, IVR, R, JR)

Odds Ratio	**ps**	**pr**	**Population 1**				**Population 2**				**Population 3**				**Population 4**			
			WC JC	**WC IVR**	**WC R**	**WC JR**	**WC JC**	**WC IVR**	**WC R**	**WC JR**	**WC JC**	**WC IVR**	**WC R**	**WC JR**	**WC JC**	**WC IVR**	**WC R**	**WC JR**
3.00	0.3	0.3	0.58	0.57	0.86	0.92	0.79	0.77	0.80	0.85	0.81	0.76	0.93	0.85	0.86	0.85	0.90	0.78
2.00	0.3	0.3	0.85	0.85	0.86	0.95	0.91	0.87	0.85	0.90	0.91	0.86	0.93	0.86	0.88	0.83	0.92	0.81
1.00	0.3	0.3	0.90	0.92	0.95	0.96	0.90	0.89	0.92	0.85	0.87	0.86	0.95	0.77	0.92	0.90	0.93	0.89
0.50	0.3	0.3	0.59	0.63	0.92	0.95	0.72	0.72	0.79	0.81	0.73	0.70	0.93	0.75	0.79	0.74	0.78	0.83
0.33	0.3	0.3	0.25	0.32	0.87	0.90	0.52	0.57	0.69	0.79	0.49	0.54	0.90	0.76	0.68	0.68	0.79	0.80
3.00	0.6	0.3	0.37	0.35	0.75	0.94	0.53	0.50	0.50	0.82	0.42	0.39	0.77	0.87	0.64	0.62	0.67	0.86
2.00	0.6	0.3	0.68	0.67	0.85	0.98	0.74	0.70	0.70	0.84	0.79	0.71	0.90	0.82	0.84	0.81	0.83	0.89
1.00	0.6	0.3	0.93	0.94	0.94	0.96	0.91	0.92	0.91	0.87	0.94	0.92	0.94	0.88	0.92	0.89	0.91	0.87
0.50	0.6	0.3	0.58	0.62	0.85	0.96	0.76	0.71	0.82	0.85	0.57	0.60	0.87	0.79	0.75	0.73	0.77	0.84
0.33	0.6	0.3	0.23	0.22	0.76	0.78	0.50	0.50	0.57	0.81	0.23	0.27	0.75	0.63	0.53	0.53	0.59	0.79
3.00	0.9	0.3	0.62	0.62	0.86	0.88	0.74	0.71	0.80	0.83	0.28	0.29	0.66	0.28	0.51	0.52	0.57	0.51
2.00	0.9	0.3	0.83	0.88	0.94	0.93	0.81	0.81	0.84	0.84	0.66	0.66	0.87	0.71	0.69	0.68	0.72	0.72
1.00	0.9	0.3	0.91	0.92	0.95	0.96	0.93	0.88	0.92	0.89	0.93	0.94	0.92	0.95	0.94	0.93	0.94	0.95
0.50	0.9	0.3	0.89	0.88	0.95	0.95	0.90	0.86	0.92	0.86	0.71	0.69	0.83	0.68	0.80	0.79	0.80	0.80
0.33	0.9	0.3	0.74	0.73	0.95	0.86	0.86	0.84	0.92	0.91	0.30	0.31	0.78	0.35	0.62	0.63	0.67	0.61
3.00	0.3	0.6	0.64	0.69	0.94	0.95	0.43	0.43	0.93	0.85	0.83	0.81	0.95	0.88	0.69	0.70	0.99	0.83
2.00	0.3	0.6	0.83	0.83	0.96	0.96	0.77	0.74	0.95	0.90	0.89	0.89	0.97	0.90	0.85	0.80	0.97	0.86
1.00	0.3	0.6	0.94	0.94	0.97	0.96	0.94	0.94	0.98	0.85	0.96	0.95	0.98	0.88	0.91	0.91	0.98	0.80
0.50	0.3	0.6	0.67	0.69	0.97	0.92	0.55	0.58	0.95	0.81	0.80	0.79	0.99	0.78	0.76	0.77	0.97	0.73
0.33	0.3	0.6	0.38	0.35	0.96	0.85	0.26	0.27	0.94	0.86	0.54	0.55	0.97	0.77	0.52	0.54	0.95	0.80
3.00	0.6	0.6	0.33	0.36	0.92	0.93	0.21	0.20	0.85	0.86	0.39	0.42	0.93	0.87	0.29	0.29	0.86	0.88
2.00	0.6	0.6	0.67	0.70	0.93	0.95	0.51	0.50	0.93	0.89	0.78	0.77	0.95	0.92	0.70	0.72	0.94	0.91
1.00	0.6	0.6	0.97	0.96	0.98	0.98	0.96	0.96	0.98	0.92	0.97	0.97	0.99	0.95	0.94	0.96	0.99	0.90
0.50	0.6	0.6	0.64	0.66	0.96	0.85	0.59	0.62	0.94	0.79	0.66	0.67	0.97	0.67	0.58	0.61	0.95	0.79
0.33	0.6	0.6	0.34	0.37	0.91	0.70	0.20	0.19	0.91	0.53	0.37	0.40	0.94	0.49	0.17	0.20	0.87	0.50
3.00	0.9	0.6	0.61	0.62	0.95	0.91	0.60	0.56	0.86	0.80	0.23	0.23	0.93	0.58	0.15	0.16	0.83	0.29

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			WC JC	WC IVR	WC R	WC JR	WC JC	WC IVR	WC R	WC JR	WC JC	WC IVR	WC R	WC JR	WC JC	WC IVR	WC R	WC JR
2.00	0.9	0.6	0.82	0.83	0.99	0.93	0.81	0.80	0.94	0.91	0.67	0.68	0.96	0.84	0.56	0.58	0.92	0.65
1.00	0.9	0.6	0.96	0.96	0.99	0.95	0.91	0.91	0.98	0.86	0.97	0.98	0.98	0.96	0.96	0.96	0.99	0.94
0.50	0.9	0.6	0.91	0.90	0.98	0.99	0.83	0.84	0.99	0.89	0.71	0.72	0.96	0.58	0.59	0.55	0.95	0.58
0.33	0.9	0.6	0.81	0.82	0.98	0.95	0.67	0.70	0.98	0.83	0.42	0.41	0.91	0.36	0.25	0.24	0.91	0.25
3.00	0.3	0.9	0.99	0.99	0.99	0.99	0.55	0.56	0.92	0.85	0.95	0.97	0.99	0.95	0.74	0.74	0.95	0.80
2.00	0.3	0.9	0.97	0.98	0.98	0.97	0.78	0.79	0.94	0.87	0.97	0.96	0.99	0.94	0.84	0.84	0.96	0.85
1.00	0.3	0.9	0.98	0.99	0.99	0.98	0.96	0.96	0.97	0.87	0.95	0.96	0.97	0.96	0.95	0.94	0.98	0.80
0.50	0.3	0.9	0.98	0.98	0.99	0.99	0.78	0.78	0.95	0.84	0.93	0.95	0.97	0.94	0.86	0.88	0.99	0.78
0.33	0.3	0.9	0.93	0.93	0.95	0.98	0.52	0.53	0.93	0.81	0.92	0.91	0.98	0.94	0.74	0.74	0.99	0.73
3.00	0.6	0.9	0.94	0.96	0.93	0.98	0.26	0.27	0.87	0.84	0.98	0.97	0.97	0.98	0.48	0.50	0.92	0.87
2.00	0.6	0.9	0.97	0.97	0.96	0.99	0.67	0.69	0.94	0.83	1.00	0.99	0.99	0.99	0.74	0.75	0.96	0.89
1.00	0.6	0.9	1.00	0.99	0.99	1.00	0.97	0.96	0.97	0.85	0.98	0.98	0.97	0.98	0.99	0.98	0.99	0.90
0.50	0.6	0.9	0.99	0.99	0.99	1.00	0.79	0.79	0.91	0.80	0.98	0.99	0.99	0.98	0.76	0.74	0.96	0.65
0.33	0.6	0.9	0.96	0.96	0.96	0.99	0.47	0.47	0.88	0.58	0.92	0.91	0.95	0.95	0.37	0.38	0.91	0.44
3.00	0.9	0.9	0.95	0.95	0.95	0.96	0.63	0.65	0.96	0.83	0.96	0.96	0.94	0.97	0.35	0.37	0.92	0.65
2.00	0.9	0.9	0.97	0.97	0.99	0.99	0.80	0.81	0.97	0.82	0.98	0.98	0.97	0.98	0.72	0.70	0.96	0.86
1.00	0.9	0.9	0.97	0.97	0.97	0.97	0.97	0.97	0.99	0.86	0.99	0.99	0.99	1.00	0.96	0.96	0.96	0.94
0.50	0.9	0.9	0.98	0.99	0.99	0.99	0.94	0.94	0.98	0.90	0.97	0.97	0.97	0.98	0.78	0.78	0.94	0.65
0.33	0.9	0.9	0.98	0.99	0.99	0.98	0.84	0.88	0.98	0.93	0.97	0.97	0.98	0.96	0.47	0.47	0.86	0.42

Table E10: Mean confidence interval coverage for three weighting class adjustment approaches (CR, QCR, SUCC)

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			CR	QCR	SUCC	CR	QCR	SUCC	CR	QCR	SUCC	CR	QCR	SUCC
3.00	0.3	0.3	0.92	0.82	0.84	0.82	0.80	0.83	0.90	0.79	0.84	0.89	0.86	0.86
2.00	0.3	0.3	0.95	0.92	0.91	0.88	0.90	0.87	0.90	0.90	0.86	0.90	0.87	0.86
1.00	0.3	0.3	0.97	0.95	0.97	0.92	0.89	0.89	0.89	0.87	0.74	0.90	0.91	0.93
0.50	0.3	0.3	0.88	0.85	0.97	0.82	0.70	0.84	0.86	0.72	0.77	0.80	0.76	0.87
0.33	0.3	0.3	0.80	0.76	0.94	0.70	0.50	0.83	0.87	0.48	0.76	0.83	0.67	0.81
3.00	0.6	0.3	0.78	0.76	0.84	0.63	0.54	0.69	0.71	0.40	0.73	0.71	0.65	0.76
2.00	0.6	0.3	0.88	0.85	0.94	0.76	0.74	0.78	0.82	0.78	0.82	0.86	0.83	0.86
1.00	0.6	0.3	0.99	0.98	0.98	0.91	0.92	0.89	0.90	0.94	0.89	0.91	0.92	0.88
0.50	0.6	0.3	0.89	0.84	0.96	0.80	0.76	0.86	0.74	0.57	0.79	0.77	0.75	0.84
0.33	0.6	0.3	0.65	0.58	0.78	0.61	0.51	0.81	0.56	0.24	0.66	0.62	0.53	0.78
3.00	0.9	0.3	0.72	0.69	0.82	0.77	0.75	0.78	0.22	0.27	0.25	0.46	0.51	0.49
2.00	0.9	0.3	0.87	0.84	0.92	0.81	0.81	0.81	0.61	0.65	0.67	0.66	0.69	0.71
1.00	0.9	0.3	0.92	0.93	0.95	0.89	0.92	0.88	0.91	0.93	0.93	0.93	0.94	0.95
0.50	0.9	0.3	0.95	0.92	0.95	0.88	0.89	0.87	0.61	0.70	0.67	0.78	0.81	0.78
0.33	0.9	0.3	0.97	0.90	0.88	0.95	0.85	0.90	0.20	0.29	0.32	0.60	0.62	0.60
3.00	0.3	0.6	0.99	0.97	0.96	0.92	0.51	0.79	0.94	0.85	0.85	0.91	0.74	0.82
2.00	0.3	0.6	0.99	0.97	0.96	0.94	0.83	0.88	0.95	0.90	0.88	0.93	0.87	0.83
1.00	0.3	0.6	1.00	0.99	0.98	0.93	0.94	0.81	0.95	0.94	0.85	0.90	0.88	0.75
0.50	0.3	0.6	0.98	0.96	0.96	0.90	0.44	0.88	0.96	0.80	0.72	0.95	0.67	0.68
0.33	0.3	0.6	0.94	0.91	0.94	0.85	0.22	0.88	0.92	0.54	0.71	0.92	0.43	0.76
3.00	0.6	0.6	0.91	0.91	0.92	0.77	0.25	0.73	0.82	0.41	0.80	0.78	0.32	0.82
2.00	0.6	0.6	0.97	0.98	0.96	0.86	0.61	0.85	0.93	0.76	0.93	0.91	0.71	0.88
1.00	0.6	0.6	0.99	0.99	0.99	0.91	0.97	0.91	0.98	0.96	0.94	0.94	0.93	0.87
0.50	0.6	0.6	0.96	0.96	0.98	0.84	0.58	0.90	0.85	0.64	0.78	0.84	0.56	0.85
0.33	0.6	0.6	0.89	0.89	0.92	0.65	0.18	0.71	0.73	0.37	0.64	0.60	0.18	0.63
3.00	0.9	0.6	0.87	0.81	0.78	0.73	0.64	0.75	0.23	0.24	0.26	0.13	0.13	0.14
2.00	0.9	0.6	0.96	0.95	0.90	0.87	0.86	0.86	0.66	0.69	0.67	0.48	0.55	0.51

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			CR	QCR	SUCC	CR	QCR	SUCC	CR	QCR	SUCC	CR	QCR	SUCC
1.00	0.9	0.6	0.98	0.96	0.96	0.88	0.91	0.80	0.97	0.97	0.96	0.94	0.96	0.94
0.50	0.9	0.6	0.99	0.99	0.98	0.90	0.83	0.85	0.68	0.71	0.64	0.50	0.60	0.55
0.33	0.9	0.6	0.98	0.98	0.97	0.88	0.66	0.82	0.35	0.40	0.35	0.18	0.26	0.22
3.00	0.3	0.9	1.00	1.00	0.99	0.89	0.67	0.84	0.97	0.96	0.96	0.90	0.84	0.74
2.00	0.3	0.9	0.99	0.99	0.98	0.91	0.86	0.86	0.98	0.96	0.95	0.92	0.90	0.80
1.00	0.3	0.9	1.00	1.00	0.99	0.93	0.96	0.88	0.98	0.94	0.95	0.90	0.93	0.74
0.50	0.3	0.9	1.00	1.00	0.99	0.88	0.70	0.84	0.98	0.94	0.94	0.90	0.82	0.71
0.33	0.3	0.9	0.99	0.99	0.98	0.87	0.39	0.83	0.99	0.92	0.94	0.88	0.61	0.65
3.00	0.6	0.9	0.98	0.98	0.99	0.79	0.32	0.76	0.98	0.97	0.97	0.86	0.56	0.85
2.00	0.6	0.9	0.99	0.99	1.00	0.83	0.72	0.81	0.99	1.00	0.99	0.88	0.77	0.87
1.00	0.6	0.9	1.00	1.00	1.00	0.93	0.97	0.91	0.98	0.98	0.98	0.96	0.98	0.91
0.50	0.6	0.9	0.99	0.99	1.00	0.86	0.76	0.86	1.00	0.98	0.98	0.82	0.71	0.73
0.33	0.6	0.9	0.98	0.99	1.00	0.79	0.43	0.83	0.95	0.92	0.94	0.69	0.31	0.62
3.00	0.9	0.9	0.96	0.95	0.95	0.86	0.71	0.75	0.94	0.95	0.96	0.32	0.36	0.35
2.00	0.9	0.9	0.99	0.99	0.98	0.87	0.85	0.82	0.97	0.98	0.98	0.69	0.70	0.72
1.00	0.9	0.9	0.97	0.97	0.97	0.93	0.97	0.88	1.00	1.00	0.99	0.98	0.97	0.95
0.50	0.9	0.9	0.99	0.99	0.99	0.93	0.93	0.90	0.97	0.97	0.97	0.71	0.79	0.67
0.33	0.9	0.9	0.99	0.99	0.98	0.94	0.82	0.94	0.96	0.97	0.96	0.41	0.44	0.39

Table E11: Mean confidence interval coverage for three weighting class adjustment approaches (INT, JCR, JQCR)

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			INT	JCR	JQCR	INT	JCR	JQCR	INT	JCR	JQCR	INT	JCR	JQCR
3.00	0.3	0.3	0.89	0.90	0.89	0.86	0.86	0.86	0.85	0.85	0.87	0.80	0.82	0.82
2.00	0.3	0.3	0.95	0.95	0.95	0.92	0.90	0.90	0.86	0.86	0.85	0.82	0.83	0.82
1.00	0.3	0.3	0.96	0.96	0.96	0.86	0.85	0.85	0.75	0.77	0.77	0.90	0.89	0.89
0.50	0.3	0.3	0.93	0.94	0.95	0.84	0.82	0.82	0.75	0.77	0.75	0.86	0.84	0.84
0.33	0.3	0.3	0.91	0.91	0.91	0.81	0.80	0.80	0.75	0.76	0.75	0.81	0.80	0.80
3.00	0.6	0.3	0.90	0.90	0.87	0.84	0.76	0.76	0.77	0.79	0.79	0.84	0.80	0.79
2.00	0.6	0.3	0.95	0.95	0.95	0.82	0.82	0.82	0.82	0.83	0.83	0.87	0.87	0.87
1.00	0.6	0.3	0.97	0.97	0.97	0.88	0.87	0.87	0.88	0.86	0.86	0.88	0.87	0.87
0.50	0.6	0.3	0.96	0.96	0.96	0.86	0.86	0.87	0.77	0.78	0.78	0.85	0.85	0.85
0.33	0.6	0.3	0.80	0.80	0.82	0.81	0.81	0.81	0.65	0.67	0.68	0.80	0.79	0.79
3.00	0.9	0.3	0.87	0.86	0.86	0.83	0.83	0.82	0.24	0.23	0.23	0.46	0.47	0.46
2.00	0.9	0.3	0.93	0.94	0.93	0.84	0.84	0.84	0.66	0.66	0.67	0.68	0.68	0.68
1.00	0.9	0.3	0.96	0.96	0.96	0.89	0.89	0.88	0.93	0.93	0.93	0.94	0.95	0.94
0.50	0.9	0.3	0.96	0.96	0.95	0.87	0.85	0.85	0.69	0.68	0.68	0.79	0.79	0.79
0.33	0.9	0.3	0.87	0.86	0.86	0.90	0.90	0.90	0.34	0.33	0.33	0.59	0.61	0.61
3.00	0.3	0.6	0.97	0.99	0.98	0.85	0.86	0.84	0.87	0.87	0.86	0.86	0.85	0.85
2.00	0.3	0.6	0.96	0.97	0.96	0.91	0.90	0.91	0.89	0.91	0.91	0.89	0.85	0.87
1.00	0.3	0.6	0.96	0.96	0.96	0.87	0.83	0.84	0.89	0.86	0.85	0.80	0.80	0.80
0.50	0.3	0.6	0.95	0.95	0.95	0.85	0.82	0.82	0.83	0.78	0.79	0.76	0.74	0.78
0.33	0.3	0.6	0.90	0.91	0.92	0.84	0.84	0.85	0.79	0.79	0.79	0.82	0.79	0.78
3.00	0.6	0.6	0.91	0.92	0.92	0.80	0.80	0.79	0.81	0.82	0.83	0.84	0.82	0.82
2.00	0.6	0.6	0.98	0.97	0.97	0.89	0.90	0.90	0.93	0.94	0.93	0.88	0.88	0.90
1.00	0.6	0.6	0.99	0.99	0.99	0.92	0.91	0.91	0.95	0.95	0.95	0.90	0.89	0.88
0.50	0.6	0.6	0.98	0.97	0.98	0.89	0.90	0.91	0.80	0.76	0.78	0.85	0.87	0.87
0.33	0.6	0.6	0.93	0.91	0.91	0.72	0.73	0.73	0.67	0.65	0.67	0.69	0.65	0.68
3.00	0.9	0.6	0.87	0.86	0.87	0.80	0.80	0.80	0.27	0.26	0.27	0.16	0.15	0.16
2.00	0.9	0.6	0.95	0.95	0.95	0.87	0.90	0.89	0.69	0.68	0.70	0.55	0.54	0.54

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			INT	JCR	JQCR	INT	JCR	JQCR	INT	JCR	JQCR	INT	JCR	JQCR
1.00	0.9	0.6	0.96	0.95	0.96	0.86	0.85	0.83	0.96	0.95	0.97	0.93	0.94	0.94
0.50	0.9	0.6	0.98	0.98	0.98	0.89	0.86	0.85	0.63	0.64	0.62	0.56	0.56	0.54
0.33	0.9	0.6	0.99	0.97	0.97	0.84	0.82	0.81	0.34	0.32	0.32	0.22	0.23	0.24
3.00	0.3	0.9	0.99	0.99	0.99	0.85	0.87	0.87	0.96	0.97	0.97	0.83	0.82	0.83
2.00	0.3	0.9	0.98	0.97	0.97	0.86	0.89	0.89	0.95	0.95	0.95	0.84	0.84	0.85
1.00	0.3	0.9	1.00	0.99	0.99	0.89	0.88	0.86	0.96	0.96	0.96	0.82	0.82	0.77
0.50	0.3	0.9	0.99	0.99	0.99	0.86	0.80	0.80	0.95	0.95	0.95	0.79	0.78	0.76
0.33	0.3	0.9	0.97	0.97	0.96	0.84	0.78	0.78	0.95	0.94	0.95	0.75	0.69	0.71
3.00	0.6	0.9	0.98	0.99	0.98	0.78	0.78	0.78	0.97	0.97	0.98	0.86	0.85	0.85
2.00	0.6	0.9	1.00	1.00	1.00	0.82	0.81	0.83	0.99	0.99	0.99	0.87	0.89	0.89
1.00	0.6	0.9	1.00	1.00	1.00	0.89	0.90	0.91	0.98	0.98	0.98	0.93	0.90	0.91
0.50	0.6	0.9	0.99	0.99	0.99	0.88	0.86	0.85	0.99	0.98	0.99	0.77	0.73	0.76
0.33	0.6	0.9	0.99	0.99	0.99	0.82	0.82	0.82	0.94	0.96	0.96	0.65	0.62	0.63
3.00	0.9	0.9	0.96	0.95	0.96	0.79	0.83	0.82	0.95	0.95	0.95	0.39	0.36	0.36
2.00	0.9	0.9	0.98	0.99	0.98	0.82	0.84	0.83	0.98	0.98	0.98	0.74	0.72	0.70
1.00	0.9	0.9	0.97	0.98	0.98	0.89	0.87	0.86	1.00	1.00	1.00	0.96	0.95	0.95
0.50	0.9	0.9	0.99	0.99	0.99	0.91	0.89	0.88	0.97	0.96	0.97	0.66	0.70	0.69
0.33	0.9	0.9	0.98	0.98	0.98	0.93	0.93	0.92	0.96	0.97	0.96	0.39	0.38	0.36

Table E12: Mean confidence interval coverage of MCAR, MCAR2, and DSS estimates

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			MCAR	MCAR2	DSS	MCAR	MCAR2	DSS	MCAR	MCAR2	DSS	MCAR	MCAR2	DSS
3.00	0.3	0.3	0.53	0.20	0.98	0.84	0.35	0.98	0.78	0.35	0.97	0.89	0.43	0.98
2.00	0.3	0.3	0.82	0.40	0.95	0.90	0.47	0.96	0.90	0.50	0.96	0.93	0.46	0.97
1.00	0.3	0.3	0.94	0.65	0.96	0.95	0.57	0.96	0.93	0.63	0.94	0.95	0.59	0.96
0.50	0.3	0.3	0.78	0.40	0.95	0.84	0.37	0.94	0.84	0.57	0.95	0.85	0.46	0.94
0.33	0.3	0.3	0.49	0.14	0.93	0.66	0.25	0.93	0.75	0.38	0.94	0.80	0.44	0.91
3.00	0.6	0.3	0.31	0.08	0.94	0.52	0.15	0.95	0.36	0.11	0.96	0.66	0.22	0.93
2.00	0.6	0.3	0.61	0.22	0.93	0.76	0.31	0.93	0.73	0.31	0.99	0.83	0.40	0.96
1.00	0.6	0.3	0.95	0.67	0.94	0.94	0.59	0.92	0.95	0.63	0.96	0.94	0.55	0.94
0.50	0.6	0.3	0.66	0.26	0.96	0.80	0.38	0.97	0.64	0.31	0.96	0.80	0.41	0.94
0.33	0.6	0.3	0.29	0.06	0.95	0.59	0.15	0.95	0.33	0.09	0.94	0.61	0.22	0.92
3.00	0.9	0.3	0.56	0.31	0.89	0.76	0.39	0.85	0.29	0.05	0.93	0.55	0.16	0.92
2.00	0.9	0.3	0.85	0.52	0.91	0.81	0.45	0.89	0.67	0.28	0.93	0.73	0.29	0.94
1.00	0.9	0.3	0.91	0.64	0.93	0.94	0.59	0.97	0.94	0.64	0.95	0.96	0.56	0.94
0.50	0.9	0.3	0.94	0.55	0.97	0.93	0.47	0.95	0.69	0.33	0.93	0.81	0.39	0.96
0.33	0.9	0.3	0.85	0.37	0.98	0.93	0.47	0.97	0.30	0.09	0.96	0.67	0.20	0.98
3.00	0.3	0.6	0.60	0.41	0.96	0.49	0.15	0.95	0.83	0.59	0.96	0.77	0.44	0.96
2.00	0.3	0.6	0.79	0.58	0.95	0.80	0.42	0.98	0.89	0.69	0.96	0.86	0.56	0.96
1.00	0.3	0.6	0.95	0.81	0.95	0.94	0.78	0.96	0.94	0.83	0.94	0.94	0.70	0.95
0.50	0.3	0.6	0.71	0.51	0.94	0.66	0.33	0.95	0.80	0.64	0.94	0.85	0.58	0.92
0.33	0.3	0.6	0.44	0.25	0.94	0.41	0.15	0.94	0.65	0.45	0.90	0.63	0.31	0.92
3.00	0.6	0.6	0.24	0.10	0.94	0.18	0.06	0.92	0.29	0.16	0.95	0.28	0.09	0.94
2.00	0.6	0.6	0.55	0.37	0.93	0.47	0.17	0.94	0.68	0.45	0.93	0.68	0.30	0.95
1.00	0.6	0.6	0.93	0.85	0.95	0.95	0.77	0.94	0.95	0.78	0.94	0.94	0.71	0.93
0.50	0.6	0.6	0.60	0.39	0.95	0.59	0.24	0.96	0.59	0.37	0.95	0.61	0.27	0.97
0.33	0.6	0.6	0.31	0.14	0.96	0.18	0.03	0.95	0.32	0.13	0.94	0.18	0.05	0.90
3.00	0.9	0.6	0.54	0.36	0.89	0.56	0.30	0.88	0.17	0.09	0.94	0.12	0.04	0.96

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			MCAR	MCAR2	DSS	MCAR	MCAR2	DSS	MCAR	MCAR2	DSS	MCAR	MCAR2	DSS
2.00	0.9	0.6	0.75	0.60	0.96	0.79	0.53	0.92	0.57	0.39	0.93	0.54	0.25	0.92
1.00	0.9	0.6	0.93	0.82	0.92	0.92	0.70	0.93	0.95	0.78	0.96	0.95	0.73	0.98
0.50	0.9	0.6	0.93	0.73	0.98	0.86	0.56	0.95	0.62	0.41	0.94	0.53	0.25	0.96
0.33	0.9	0.6	0.82	0.61	0.95	0.75	0.39	0.95	0.31	0.12	0.96	0.21	0.05	0.98
3.00	0.3	0.9	0.97	0.91	0.98	0.54	0.39	0.97	0.93	0.88	0.94	0.80	0.58	0.97
2.00	0.3	0.9	0.92	0.90	0.94	0.78	0.57	0.97	0.94	0.90	0.95	0.87	0.67	0.95
1.00	0.3	0.9	0.96	0.94	0.96	0.93	0.80	0.95	0.96	0.92	0.93	0.94	0.82	0.92
0.50	0.3	0.9	0.96	0.90	0.96	0.74	0.53	0.96	0.94	0.91	0.95	0.83	0.68	0.95
0.33	0.3	0.9	0.88	0.86	0.95	0.47	0.30	0.94	0.91	0.87	0.95	0.74	0.52	0.96
3.00	0.6	0.9	0.88	0.82	0.93	0.19	0.10	0.94	0.92	0.86	0.93	0.40	0.22	0.96
2.00	0.6	0.9	0.92	0.89	0.94	0.57	0.37	0.96	0.96	0.93	0.98	0.68	0.47	0.96
1.00	0.6	0.9	0.94	0.89	0.94	0.95	0.84	0.96	0.95	0.91	0.95	0.98	0.88	0.96
0.50	0.6	0.9	0.94	0.89	0.95	0.66	0.44	0.96	0.95	0.91	0.98	0.67	0.47	0.96
0.33	0.6	0.9	0.88	0.85	0.94	0.34	0.18	0.97	0.84	0.79	0.95	0.29	0.13	0.95
3.00	0.9	0.9	0.89	0.85	0.93	0.61	0.46	0.92	0.86	0.83	0.97	0.26	0.13	0.95
2.00	0.9	0.9	0.92	0.90	0.94	0.76	0.60	0.93	0.92	0.87	0.93	0.60	0.41	0.96
1.00	0.9	0.9	0.92	0.89	0.94	0.94	0.84	0.94	0.96	0.91	0.96	0.93	0.84	0.95
0.50	0.9	0.9	0.97	0.92	0.97	0.90	0.74	0.95	0.92	0.88	0.95	0.67	0.44	0.96
0.33	0.9	0.9	0.95	0.91	0.95	0.82	0.61	0.97	0.88	0.85	0.97	0.35	0.19	0.95

Table E13: Mean RMSE for four PSAE approaches (R, R*, MAR, MAR*)

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			R	R*	MAR	MAR*	R	R*	MAR	MAR*	R	R*	MAR	MAR*	R	R*	MAR	MAR*
3.00	0.3	0.3	66	68	140	154	141	139	109	102	56	57	143	146	102	94	86	83
2.00	0.3	0.3	74	74	112	119	144	135	121	116	60	60	116	115	104	99	91	88
1.00	0.3	0.3	84	82	83	84	149	140	151	146	67	66	85	87	110	103	111	108
0.50	0.3	0.3	90	87	104	101	164	150	185	181	72	70	72	71	114	105	130	127
0.33	0.3	0.3	101	97	131	128	170	162	203	200	78	74	88	86	121	107	137	134
3.00	0.6	0.3	108	116	182	200	246	242	218	206	82	91	250	252	176	165	137	127
2.00	0.6	0.3	110	113	149	164	249	234	239	228	90	93	185	187	186	170	168	160
1.00	0.6	0.3	118	117	118	122	260	233	289	276	96	95	113	117	194	175	196	188
0.50	0.6	0.3	136	132	162	157	280	249	334	325	120	111	127	125	206	182	248	242
0.33	0.6	0.3	150	146	215	207	276	259	362	353	128	120	163	159	213	194	267	260
3.00	0.9	0.3	148	143	146	152	377	336	394	370	113	123	262	263	257	248	222	209
2.00	0.9	0.3	151	147	148	155	373	323	418	397	115	118	222	225	256	236	245	231
1.00	0.9	0.3	148	145	144	148	374	327	426	410	123	120	130	133	270	242	302	288
0.50	0.9	0.3	154	150	153	150	387	349	444	426	139	135	146	142	280	258	337	326
0.33	0.9	0.3	160	156	166	161	380	335	454	438	153	147	180	175	299	271	367	358
3.00	0.3	0.6	33	34	74	73	44	48	46	47	29	30	53	53	38	39	43	44
2.00	0.3	0.6	35	36	58	57	47	49	64	65	31	31	48	47	40	41	46	47
1.00	0.3	0.6	38	38	38	38	52	52	95	96	33	32	32	32	45	44	72	72
0.50	0.3	0.6	42	41	55	55	61	59	135	136	35	34	38	38	48	47	88	88
0.33	0.3	0.6	45	45	76	76	66	62	151	151	37	36	49	49	54	52	106	106
3.00	0.6	0.6	46	48	107	107	64	69	106	106	39	41	106	105	51	56	59	60
2.00	0.6	0.6	48	49	80	79	66	68	124	125	40	42	83	81	55	57	85	85
1.00	0.6	0.6	51	51	51	50	72	72	184	184	44	44	45	45	61	60	128	127
0.50	0.6	0.6	55	54	77	76	81	80	234	234	49	48	70	70	71	69	179	179
0.33	0.6	0.6	58	57	102	102	88	86	266	267	52	51	93	93	78	77	207	208
3.00	0.9	0.6	59	59	69	68	85	85	245	248	47	49	121	118	65	70	108	110

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			R	R*	MA R	MA R*	R	R*	MA R	MA R*	R	R*	MA R	MA R*	R	R*	MA R	MA R*
2.00	0.9	0.6	59	59	64	65	85	84	251	252	49	50	91	90	67	69	133	134
1.00	0.9	0.6	60	60	60	59	87	86	268	270	52	52	54	53	73	73	183	185
0.50	0.9	0.6	61	60	62	62	89	88	285	288	57	55	71	71	82	81	241	241
0.33	0.9	0.6	62	61	65	64	92	90	297	299	59	58	96	97	89	87	272	272
3.00	0.3	0.9	27	27	34	31	33	35	50	55	24	23	29	26	28	29	41	45
2.00	0.3	0.9	28	27	31	28	34	35	65	70	24	23	26	24	29	30	47	50
1.00	0.3	0.9	29	28	28	27	37	36	89	94	24	23	24	23	31	31	63	66
0.50	0.3	0.9	30	28	28	28	41	39	115	119	25	24	24	24	34	33	79	82
0.33	0.3	0.9	31	28	31	31	43	42	129	132	26	24	25	26	35	34	87	89
3.00	0.6	0.9	35	36	46	40	45	48	108	118	30	30	38	33	37	39	69	75
2.00	0.6	0.9	36	35	42	37	46	47	130	142	31	31	38	33	39	40	85	93
1.00	0.6	0.9	37	35	36	35	49	49	165	174	32	31	32	31	42	42	119	125
0.50	0.6	0.9	38	36	37	39	53	52	203	212	33	32	33	35	47	46	154	160
0.33	0.6	0.9	39	36	39	42	56	54	223	231	34	34	39	41	50	49	173	178
3.00	0.9	0.9	41	39	43	39	56	56	232	248	36	37	50	42	45	48	119	130
2.00	0.9	0.9	41	39	42	39	56	56	234	248	36	36	43	38	47	48	141	151
1.00	0.9	0.9	42	39	41	39	57	57	248	264	37	36	37	36	50	49	171	182
0.50	0.9	0.9	42	40	40	40	58	57	259	272	38	37	37	39	54	53	213	221
0.33	0.9	0.9	42	40	40	41	59	58	265	279	39	37	40	43	56	55	227	235

Table E14: Mean RMSE for four PSAE approaches (CR, CR*, QCR, QCR*)

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*
3.00	0.3	0.3	80	89	90	101	142	146	143	157	59	62	63	69	103	95	98	91
2.00	0.3	0.3	81	85	86	91	135	122	134	123	62	62	65	66	104	96	100	94
1.00	0.3	0.3	84	82	85	85	144	134	144	133	69	66	70	68	107	102	106	101
0.50	0.3	0.3	92	90	93	91	167	154	168	153	74	70	73	69	112	103	113	106
0.33	0.3	0.3	106	102	110	107	179	165	182	169	77	74	77	75	121	107	124	110
3.00	0.6	0.3	222	236	500	313	250	232	242	235	173	179	204	194	189	252	197	238
2.00	0.6	0.3	138	154	156	165	254	250	261	237	138	159	148	168	174	159	173	158
1.00	0.6	0.3	123	122	126	125	260	233	259	234	97	100	98	100	189	177	189	178
0.50	0.6	0.3	160	157	161	154	315	301	319	299	137	131	139	133	232	219	235	220
0.33	0.6	0.3	233	216	235	219	343	323	346	322	185	176	186	178	262	247	264	254
3.00	0.9	0.3	151	148	164	167	374	330	380	343	190	207	221	227	219	229	220	236
2.00	0.9	0.3	150	155	154	156	362	314	370	321	152	153	163	150	230	217	235	230
1.00	0.9	0.3	146	146	148	147	366	312	373	324	124	121	124	123	262	234	264	235
0.50	0.9	0.3	152	151	154	150	385	345	391	353	167	160	168	160	307	273	308	276
0.33	0.9	0.3	160	159	162	158	377	326	384	336	219	209	219	211	352	315	352	316
3.00	0.3	0.6	35	36	40	41	51	56	61	67	30	29	29	30	41	41	40	42
2.00	0.3	0.6	36	37	40	40	50	53	55	59	30	30	31	31	39	39	41	42
1.00	0.3	0.6	38	38	38	37	52	52	53	52	33	32	32	32	45	44	46	45
0.50	0.3	0.6	43	42	43	43	62	61	62	61	40	37	38	35	58	54	54	50
0.33	0.3	0.6	47	47	48	50	67	66	68	69	43	40	40	37	60	57	57	53
3.00	0.6	0.6	66	101	76	70	118	255	159	190	59	93	70	67	112	215	150	185
2.00	0.6	0.6	56	68	61	58	108	179	133	135	53	61	59	55	85	123	105	96
1.00	0.6	0.6	52	51	52	51	74	74	75	75	45	44	45	44	63	63	63	62
0.50	0.6	0.6	82	75	80	77	126	118	128	119	75	69	75	72	102	96	103	97
0.33	0.6	0.6	141	122	133	121	213	197	205	190	120	107	115	106	171	160	169	161
3.00	0.9	0.6	60	64	63	62	91	97	95	97	103	108	109	107	145	155	152	157

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*	CR	CR*	QCR	QCR*
2.00	0.9	0.6	60	60	61	60	88	87	90	89	78	76	79	77	109	103	110	113
1.00	0.9	0.6	60	60	61	60	87	86	89	88	52	52	52	52	74	73	75	74
0.50	0.9	0.6	62	61	62	61	89	89	90	90	79	79	80	81	121	117	121	117
0.33	0.9	0.6	63	62	63	62	93	93	94	93	106	105	107	107	162	157	164	161
3.00	0.3	0.9	28	28	28	27	39	41	44	45	23	23	23	23	28	28	29	30
2.00	0.3	0.9	28	27	28	27	38	38	41	41	24	23	23	23	29	29	30	30
1.00	0.3	0.9	28	27	28	27	37	36	37	36	24	23	24	23	32	31	32	31
0.50	0.3	0.9	29	28	28	28	41	41	41	42	25	24	25	24	35	33	35	33
0.33	0.3	0.9	29	29	29	30	45	45	45	46	26	24	26	25	41	38	39	37
3.00	0.6	0.9	37	35	39	37	62	85	72	70	32	32	34	32	50	66	56	56
2.00	0.6	0.9	37	35	38	35	58	65	63	60	32	31	33	31	48	53	52	48
1.00	0.6	0.9	36	35	37	35	50	50	50	50	32	31	32	31	43	43	43	43
0.50	0.6	0.9	38	37	37	37	85	75	83	80	35	33	33	33	73	68	72	71
0.33	0.6	0.9	45	39	39	38	138	115	133	126	44	37	38	35	115	103	111	104
3.00	0.9	0.9	41	39	43	39	58	59	60	59	47	40	49	42	94	98	99	98
2.00	0.9	0.9	41	39	42	39	57	57	59	58	41	37	42	37	72	72	73	72
1.00	0.9	0.9	41	40	41	39	57	57	58	57	37	36	37	36	50	50	51	50
0.50	0.9	0.9	41	40	42	40	59	58	59	59	38	40	39	40	76	76	76	77
0.33	0.9	0.9	41	40	42	40	60	59	60	60	42	44	42	44	98	98	98	99

Table E15: Mean RMSE for four weighting class adjustment approaches (JC, IVR, R, JR)

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			WC JC	WC IVR	WC R	WC JR	WC JC	WC IVR	WC R	WC JR	WC JC	WC IVR	WC R	WC JR	WC JC	WC IVR	WC R	WC JR
3.00	0.3	0.3	84	84	61	43	75	77	95	49	55	58	54	32	54	54	72	42
2.00	0.3	0.3	56	58	57	41	60	63	82	48	41	42	53	33	49	52	69	42
1.00	0.3	0.3	42	42	56	42	48	48	63	49	37	35	50	36	44	44	56	44
0.50	0.3	0.3	73	71	55	45	73	72	63	55	52	51	50	40	56	54	55	47
0.33	0.3	0.3	100	96	64	49	93	89	71	62	68	63	50	42	64	61	55	47
3.00	0.6	0.3	114	116	82	65	109	113	92	72	118	123	77	69	104	107	120	75
2.00	0.6	0.3	82	85	64	56	83	86	79	65	79	84	62	57	80	82	92	67
1.00	0.6	0.3	49	49	56	51	56	57	65	58	48	49	57	50	55	57	67	58
0.50	0.6	0.3	101	97	70	61	92	90	94	70	93	92	65	66	85	84	89	68
0.33	0.6	0.3	144	144	83	88	121	123	130	81	130	128	87	83	107	109	101	74
3.00	0.9	0.3	42	42	37	33	38	39	38	36	123	124	88	125	109	110	89	114
2.00	0.9	0.3	31	31	35	30	35	35	39	34	83	84	60	83	88	89	83	90
1.00	0.9	0.3	27	28	36	29	33	33	42	35	48	49	56	50	56	56	65	57
0.50	0.9	0.3	39	38	38	30	44	45	50	37	85	88	65	89	84	85	91	87
0.33	0.9	0.3	53	56	40	32	51	53	54	38	133	135	67	139	111	114	121	118
3.00	0.3	0.6	61	59	27	28	85	86	47	35	41	40	26	23	53	54	39	27
2.00	0.3	0.6	46	46	31	29	58	59	44	34	34	34	29	24	43	45	41	27
1.00	0.3	0.6	32	32	33	32	35	35	43	36	27	27	31	26	30	30	39	30
0.50	0.3	0.6	55	54	33	34	68	67	44	40	38	38	31	30	44	42	39	36
0.33	0.3	0.6	74	75	33	38	89	88	49	43	50	50	31	32	59	58	42	36
3.00	0.6	0.6	91	90	32	40	118	119	60	55	93	92	31	38	115	116	54	52
2.00	0.6	0.6	65	64	32	40	86	86	52	49	64	64	32	38	78	77	47	48
1.00	0.6	0.6	38	38	35	39	43	43	45	45	37	38	35	38	42	43	45	44
0.50	0.6	0.6	70	70	31	50	86	85	47	64	68	69	32	54	81	81	50	59
0.33	0.6	0.6	95	95	29	77	129	129	50	108	90	91	33	80	121	121	64	97
3.00	0.9	0.6	35	34	25	23	39	40	33	28	98	98	34	72	122	121	63	113

Odds Ratio	ps	pr	Population 1				Population 2				Population 3				Population 4			
			WC JC	WC IVR	WC R	WC JR	WC JC	WC IVR	WC R	WC JR	WC JC	WC IVR	WC R	WC JR	WC JC	WC IVR	WC R	WC JR
2.00	0.9	0.6	27	27	24	22	29	30	29	29	67	67	33	52	82	82	49	76
1.00	0.9	0.6	22	22	23	21	25	25	28	26	38	38	35	38	43	43	45	44
0.50	0.9	0.6	30	31	22	20	38	37	28	26	68	68	31	74	87	88	47	92
0.33	0.9	0.6	37	37	22	19	51	50	28	28	93	93	28	94	125	125	47	136
3.00	0.3	0.9	30	30	29	26	69	69	25	29	24	24	27	22	46	46	24	23
2.00	0.3	0.9	28	29	30	27	49	49	27	29	23	23	27	23	37	37	26	24
1.00	0.3	0.9	28	28	30	27	31	31	32	31	23	23	28	23	26	26	29	26
0.50	0.3	0.9	29	29	30	28	46	46	31	33	23	24	27	23	33	33	29	28
0.33	0.3	0.9	32	32	30	28	63	63	32	36	26	26	28	23	40	40	29	31
3.00	0.6	0.9	38	38	31	32	95	95	32	38	35	36	31	32	82	82	30	37
2.00	0.6	0.9	35	35	32	32	67	67	31	37	34	35	31	32	63	63	31	36
1.00	0.6	0.9	33	33	32	33	37	37	34	38	33	33	32	33	36	37	34	37
0.50	0.6	0.9	36	36	31	33	61	62	29	53	35	35	32	33	59	60	30	56
0.33	0.6	0.9	38	38	30	33	86	86	27	80	41	41	32	35	86	86	30	85
3.00	0.9	0.9	19	20	21	18	33	33	23	22	40	40	31	36	88	88	31	63
2.00	0.9	0.9	19	19	20	19	27	27	22	21	35	35	31	34	63	63	31	46
1.00	0.9	0.9	19	19	20	18	21	21	22	21	34	34	31	33	37	37	33	38
0.50	0.9	0.9	20	20	20	18	26	27	21	19	36	36	30	35	64	64	28	66
0.33	0.9	0.9	20	20	20	18	34	34	22	19	40	40	30	37	86	86	27	87

Table E16: Mean RMSE for three weighting class adjustment approaches (CR, QCR, SUCC)

Odds Ratio	**ps**	**pr**	**Population 1**			**Population 2**			**Population 3**			**Population 4**		
			CR	**QCR**	**SUCC**	**CR**	**QCR**	**SUCC**	**CR**	**QCR**	**SUCC**	**CR**	**QCR**	**SUCC**
3.00	0.3	0.3	72	87	56	83	72	59	49	54	38	60	53	47
2.00	0.3	0.3	63	72	49	77	59	53	48	41	36	61	48	45
1.00	0.3	0.3	54	55	46	63	49	51	48	37	40	57	44	48
0.50	0.3	0.3	65	66	44	73	75	52	49	53	45	55	58	47
0.33	0.3	0.3	78	81	47	84	95	59	51	69	46	56	65	48
3.00	0.6	0.3	94	95	76	103	109	84	102	119	82	104	104	87
2.00	0.6	0.3	75	79	61	86	83	69	79	80	63	85	80	69
1.00	0.6	0.3	57	57	49	65	57	57	58	48	50	67	55	57
0.50	0.6	0.3	83	88	59	95	92	69	83	94	63	87	85	67
0.33	0.6	0.3	115	123	84	119	121	81	109	130	80	102	107	74
3.00	0.9	0.3	42	42	36	41	38	36	150	124	128	131	110	112
2.00	0.9	0.3	36	35	30	40	35	35	100	83	85	110	88	89
1.00	0.9	0.3	34	35	27	39	33	32	57	48	49	65	56	56
0.50	0.9	0.3	36	42	27	50	45	31	113	85	91	104	83	88
0.33	0.9	0.3	39	52	28	55	53	33	175	133	141	137	111	119
3.00	0.3	0.6	35	40	35	51	78	48	29	38	27	37	49	31
2.00	0.3	0.6	36	40	33	46	52	38	30	32	29	37	37	33
1.00	0.3	0.6	36	36	32	44	37	39	32	27	27	39	32	35
0.50	0.3	0.6	38	39	33	49	76	39	33	39	34	42	49	45
0.33	0.3	0.6	42	45	34	54	95	40	33	51	39	41	66	44
3.00	0.6	0.6	51	53	52	74	114	70	56	91	55	77	113	67
2.00	0.6	0.6	43	43	43	59	81	54	46	62	44	57	75	51
1.00	0.6	0.6	39	39	38	47	43	44	40	37	38	48	42	44
0.50	0.6	0.6	48	49	44	66	88	54	50	69	47	64	86	52
0.33	0.6	0.6	59	59	54	98	134	84	67	90	63	93	124	79
3.00	0.9	0.6	29	30	29	36	37	31	103	98	98	136	122	127
2.00	0.9	0.6	25	26	25	30	28	26	71	66	66	93	83	86

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			CR	QCR	SUCC	CR	QCR	SUCC	CR	QCR	SUCC	CR	QCR	SUCC
1.00	0.9	0.6	23	23	22	28	26	24	39	38	38	46	43	44
0.50	0.9	0.6	22	22	20	27	39	23	73	68	76	104	87	96
0.33	0.9	0.6	21	21	19	28	51	23	102	93	103	147	125	135
3.00	0.3	0.9	32	32	27	39	60	38	28	24	22	28	38	27
2.00	0.3	0.9	32	32	27	37	44	35	28	23	22	30	31	27
1.00	0.3	0.9	32	32	28	35	31	33	28	23	22	31	26	28
0.50	0.3	0.9	32	32	28	38	51	32	29	24	23	31	36	31
0.33	0.3	0.9	33	33	28	40	71	33	29	26	23	32	45	37
3.00	0.6	0.9	36	36	33	58	89	58	35	35	33	52	79	50
2.00	0.6	0.9	35	35	33	46	63	46	35	34	33	47	60	44
1.00	0.6	0.9	34	34	33	39	37	38	35	33	33	39	37	37
0.50	0.6	0.9	35	35	33	51	64	48	36	35	33	51	62	49
0.33	0.6	0.9	35	35	34	64	89	59	38	41	36	65	90	63
3.00	0.9	0.9	21	21	19	28	31	28	42	40	39	94	89	91
2.00	0.9	0.9	21	21	19	25	25	25	36	35	35	65	63	62
1.00	0.9	0.9	20	20	19	23	22	21	34	34	34	38	37	38
0.50	0.9	0.9	20	20	18	23	28	20	38	36	37	70	65	71
0.33	0.9	0.9	20	20	18	25	36	20	42	40	41	94	87	94

Table E17: Mean RMSE for three weighting class adjustment approaches (INT, JCR, JQCR)

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			INT	JCR	JQCR	INT	JCR	JQCR	INT	JCR	JQCR	INT	JCR	JQCR
3.00	0.3	0.3	48	47	47	50	50	50	34	34	34	42	42	42
2.00	0.3	0.3	42	42	42	49	48	48	33	34	34	42	42	42
1.00	0.3	0.3	42	42	42	49	49	49	35	36	35	44	44	44
0.50	0.3	0.3	45	45	45	54	55	55	40	40	40	46	47	47
0.33	0.3	0.3	49	49	49	62	62	62	41	42	42	47	48	47
3.00	0.6	0.3	74	74	74	78	77	78	81	80	81	83	83	85
2.00	0.6	0.3	58	58	59	68	67	67	60	60	60	68	68	68
1.00	0.6	0.3	51	51	51	58	58	58	50	50	50	58	58	58
0.50	0.6	0.3	61	60	61	69	69	69	66	65	65	67	67	68
0.33	0.6	0.3	85	84	83	81	81	81	80	81	79	73	74	74
3.00	0.9	0.3	34	34	34	36	36	35	132	132	132	118	118	118
2.00	0.9	0.3	30	30	30	34	34	34	87	87	86	93	92	92
1.00	0.9	0.3	29	29	29	35	35	35	50	50	50	57	57	57
0.50	0.9	0.3	30	30	29	36	36	36	90	90	91	87	87	87
0.33	0.9	0.3	32	31	31	37	37	38	140	141	141	118	117	117
3.00	0.3	0.6	32	33	33	43	40	40	25	25	25	29	28	28
2.00	0.3	0.6	31	31	31	36	35	35	26	26	26	30	28	29
1.00	0.3	0.6	32	32	32	36	36	36	26	26	26	32	31	31
0.50	0.3	0.6	35	35	35	39	41	41	29	30	30	38	36	35
0.33	0.3	0.6	41	40	39	42	43	43	29	30	31	38	36	36
3.00	0.6	0.6	51	50	51	68	66	68	56	54	54	68	68	68
2.00	0.6	0.6	42	42	42	52	52	52	45	45	45	51	50	51
1.00	0.6	0.6	39	39	39	44	45	45	38	38	38	44	44	44
0.50	0.6	0.6	45	45	45	56	54	54	48	49	48	54	55	55
0.33	0.6	0.6	55	55	54	84	84	83	62	63	61	78	80	77
3.00	0.9	0.6	27	27	27	29	29	29	99	100	99	126	128	127
2.00	0.9	0.6	24	24	24	27	28	28	67	66	66	84	86	86

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			INT	JCR	JQCR	INT	JCR	JQCR	INT	JCR	JQCR	INT	JCR	JQCR
1.00	0.9	0.6	22	22	22	26	26	26	39	39	39	44	44	44
0.50	0.9	0.6	20	20	20	25	25	24	75	77	77	95	95	96
0.33	0.9	0.6	20	20	20	26	25	25	103	105	105	134	134	134
3.00	0.3	0.9	27	27	28	38	36	36	23	23	23	27	26	26
2.00	0.3	0.9	28	28	28	34	33	33	23	23	23	27	26	26
1.00	0.3	0.9	28	28	28	32	32	32	23	23	23	27	26	26
0.50	0.3	0.9	28	28	28	33	34	34	23	23	23	29	28	28
0.33	0.3	0.9	29	29	29	36	38	37	23	23	24	32	31	31
3.00	0.6	0.9	33	33	33	56	56	56	33	33	33	51	50	50
2.00	0.6	0.9	33	33	33	46	44	45	33	33	33	44	44	43
1.00	0.6	0.9	33	33	33	38	38	38	33	33	33	38	38	38
0.50	0.6	0.9	34	34	34	49	50	50	34	34	34	48	49	49
0.33	0.6	0.9	34	34	34	61	60	60	36	36	36	62	65	64
3.00	0.9	0.9	19	19	19	26	26	26	40	39	40	89	91	91
2.00	0.9	0.9	19	19	19	24	24	24	35	35	35	62	62	63
1.00	0.9	0.9	19	19	19	21	21	21	34	34	34	38	38	38
0.50	0.9	0.9	19	19	19	20	20	20	37	37	37	72	72	72
0.33	0.9	0.9	19	18	18	20	21	21	41	41	41	95	96	96

Table E18: Mean RMSE of MCAR, MCAR2, and DSS estimates

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			MCAR	MCAR2	DSS	MCAR	MCAR2	DSS	MCAR	MCAR2	DSS	MCAR	MCAR2	DSS
3.00	0.3	0.3	99	89	24	87	69	23	70	60	22	66	45	21
2.00	0.3	0.3	70	57	23	72	49	23	54	39	21	59	35	21
1.00	0.3	0.3	46	22	22	56	22	22	41	20	20	50	20	20
0.50	0.3	0.3	64	53	21	72	54	21	48	34	19	55	35	19
0.33	0.3	0.3	90	83	21	89	76	21	60	50	18	63	47	19
3.00	0.6	0.3	120	113	22	114	102	22	124	116	24	109	95	24
2.00	0.6	0.3	89	80	23	86	69	23	85	73	24	83	63	24
1.00	0.6	0.3	48	24	24	59	24	24	49	24	24	59	24	24
0.50	0.6	0.3	91	80	24	91	73	24	87	78	23	84	66	23
0.33	0.6	0.3	133	126	25	118	104	24	121	114	23	105	92	23
3.00	0.9	0.3	43	38	12	41	31	12	122	116	21	109	97	22
2.00	0.9	0.3	33	24	13	38	26	12	82	72	22	89	74	22
1.00	0.9	0.3	29	14	14	35	14	14	48	24	23	58	24	23
0.50	0.9	0.3	39	28	16	45	28	16	85	73	24	85	66	24
0.33	0.9	0.3	53	44	17	51	35	16	130	123	24	113	98	24
3.00	0.3	0.6	65	61	24	91	85	24	46	41	21	59	51	21
2.00	0.3	0.6	50	44	23	63	54	23	38	32	21	51	41	21
1.00	0.3	0.6	31	22	22	38	22	22	28	20	20	34	20	20
0.50	0.3	0.6	51	46	22	66	59	21	36	31	19	43	34	19
0.33	0.3	0.6	70	67	21	85	81	21	47	44	19	59	54	18
3.00	0.6	0.6	91	89	23	118	115	22	92	89	25	116	110	25
2.00	0.6	0.6	66	62	23	86	81	23	63	58	24	77	70	24
1.00	0.6	0.6	33	24	24	41	24	24	34	24	24	41	24	24
0.50	0.6	0.6	65	61	24	83	75	24	65	61	24	79	71	23
0.33	0.6	0.6	90	86	25	126	122	25	86	82	23	119	115	23
3.00	0.9	0.6	35	32	12	40	36	12	95	93	22	119	115	22

Odds Ratio	ps	pr	Population 1			Population 2			Population 3			Population 4		
			MCAR	MCAR2	DSS	MCAR	MCAR2	DSS	MCAR	MCAR2	DSS	MCAR	MCAR2	DSS
2.00	0.9	0.6	27	23	13	30	25	13	64	60	23	80	74	22
1.00	0.9	0.6	20	14	14	25	14	14	33	23	23	40	24	23
0.50	0.9	0.6	27	23	15	36	28	16	65	61	24	86	79	24
0.33	0.9	0.6	34	30	16	49	43	17	91	88	25	123	118	25
3.00	0.3	0.9	28	26	23	67	62	24	24	22	20	45	41	21
2.00	0.3	0.9	26	24	23	47	43	23	22	21	20	38	33	21
1.00	0.3	0.9	24	22	22	30	22	22	21	20	20	27	20	20
0.50	0.3	0.9	24	23	22	47	42	22	22	20	20	35	30	19
0.33	0.3	0.9	27	25	22	65	62	21	23	22	20	42	39	19
3.00	0.6	0.9	32	31	24	91	89	23	29	27	24	79	76	25
2.00	0.6	0.9	29	27	24	64	61	23	28	27	24	60	55	24
1.00	0.6	0.9	25	24	24	32	24	24	26	24	24	32	24	24
0.50	0.6	0.9	28	27	24	59	55	24	28	27	24	58	55	24
0.33	0.6	0.9	30	28	24	85	83	25	34	32	24	85	83	24
3.00	0.9	0.9	16	16	14	32	30	13	33	31	23	86	85	23
2.00	0.9	0.9	15	14	14	26	23	13	27	25	23	60	57	23
1.00	0.9	0.9	15	14	14	19	14	14	25	24	24	32	23	23
0.50	0.9	0.9	16	15	15	26	22	15	28	27	24	61	57	24
0.33	0.9	0.9	16	15	15	34	30	16	32	31	24	83	80	24