Extensions of the Penalized Spline Propensity Prediction Method of Imputation

by

Guangyu Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2007

Doctoral Committee:

Professor Roderick J. Little, Chair
Professor Susan A. Murphy
Professor Trivellore E. Raghunathan
Associate Professor Bin Nan

ACKNOWLEGEMENTS

I would like to thank my advisor, Dr. Roderick Little, for his advice, support and intellectual stimulation. I also want to thank my husband, Huanyuan Sheng, for his love.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

ABSTRACT

Little and An (2004) proposed a penalized spline propensity prediction (PSPP) method of imputation of missing values that yields robust model-based inference under the missing at random assumption. The propensity score for a missing variable is estimated and a regression model is fit with the spline of the propensity score as a covariate. The predicted marginal mean of the missing variable is doubly robust (DR) under the misspecification of the imputation model.

In the first part of the thesis, we study properties of a simplified version of the PSPP that does not center the regressors prior to including them in the prediction model. We then extend the PSPP to multivariate data with both continuous and categorical variables so as to yield consistent estimates of both marginal and conditional means. The extended PSPP method is compared with the PSPP method and simple alternatives in a simulation study.

For the second part of the thesis, we compare the PSPP method with several other DR estimators. The PSPP method uses a spline of propensity score to impute the missing values and the resulting estimates have a double robustness property. The DR property can also be achieved by modeling the relationship parametrically, such as the linear in the weight method and calibration method (Firth and Bennett, 1998; Scharfstein, Rotnitzky and Robins, 1999; Robins, Rotnitzky and Zhao, 1994; Scharfstein, Rotnitzky and Robins, 1999). We compare root mean square error (RMSE), width of confidence interval and non-coverage rate of the PSPP method and these alternatives under different mean functions and propensity score functions. We also study the effects of sample size and misspecification of the propensity scores. The PSPP method yields estimates with smaller RMSE and width of confidence interval compared with other methods under most situations. It yields estimates with non-coverage rates close to the 5% nominal level.

For the third part of the thesis, we extend the PSPP methods to the monotone missing data. We propose to impute the missing values based on a stepwise PSPP procedure and simulation studies show the stepwise PSPP method yields consistent marginal and conditional mean estimates. We illustrate the proposed method by applying it to an online weight loss study conducted by Kaiser Permanente. We finish the thesis with a short discussion and future work.

CHAPTER I

INTRODUCTION

Missing data arise in scientific research for many reasons. For example, in a two stage clinical study, a subset of subjects is selected for expensive medical tests and those who have not been selected will have missing values. On the other hand, subjects who have been selected may drop out of the study so that it is impossible to collect test results. No matter what are the reasons, it is important to include the information in the incomplete cases in the analysis to yield efficient estimators and better inferences.

A dataset with missing values can be described by the missing-data pattern, which indicates which observations are present and which are missing. Let $\underline{Y} = (y_{ij})$ be a $(n \times p)$ rectangular data set with the ith row $y_i = (y_{i1},..., y_{ip})$, where y_{ij} is the jth observation for subject i , $j = 1,..., p$. Let $M = (m_{ij})$ be a missing-data indicator matrix with the ith row $m_i = (m_{i1},..., m_{ip})$, such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is present. The matrix M then represents the missing data pattern of the dataset $\underline{Y}$. We assume that (y_i, m_i) are independent over i throughout the dissertation.

We first focus on the univariate missing data, where missingness confines to a single variable. Let $(X_1,..., X_p, Y)$ denote a $(p+1)$-dimensional vector of variables with Y subject to missing values and $X_1,..., X_p$ fully observed covariates. We consider the problem of estimating the mean of *Y*, and the conditional means of *Y* in subclasses defined by a categorical variable, and the regression coefficient of *Y* on a continuous variable.

Estimation of the marginal and conditional means of Y requires the assumptions on the missing-data mechanism, which concerns the relationship between the missingness and the values of the variables in the data matrix. Rubin (1976) treated M as a random matrix and described the missing data mechanism by the conditional distribution of M given $\underline{Y}$, $f(M \mid \underline{Y}, \phi)$, where ϕ denotes unknown parameters. When missingness does not depend on $\underline{Y}$, missing or observed, that is,

$$f(M \mid \underline{Y}, \phi) = f(M \mid \phi) \text{ for all } \underline{Y}, \phi,$$

the data are called missing complete at random (MCAR). If the missingness depends only on Y_{obs}, the observed part of $\underline{Y}$, but not the missing part of $\underline{Y}$, Y_{mis}, that is

$$f(M \mid \underline{Y}, \phi) = f(M \mid Y_{\text{obs}}, \phi) \text{ for all } Y_{\text{mis}}, \phi,$$

then the missing data mechanism is called missing at random (MAR). If the missingness depends on the missing part of the variables, that is,

$$f(M \mid \underline{Y}, \phi) = f(M \mid Y_{\text{obs}}, Y_{\text{mis}}, \phi) \text{ for all } \phi,$$

then the data are called not missing at random (NMAR). MAR is a less restrictive assumption than MCAR. In applications researchers are encouraged to take efforts to render the MAR assumptions plausible by measuring covariates that characterize the nonrespondents (Little and Rubin, 1999). We assume the missing data are MAR throughout the dissertation.

Many methods have been proposed to deal with missing information. A simple approach is complete-case analysis (CC), which deletes units with Y missing, so information contained in the deleted cases is lost. In the context of our problem, CC analysis yields consistent estimates of marginal mean of Y, if the missing-data mechanism is missing complete at random. But it yields biased estimates if the missingness of Y depends on the observed covariates $X_1, ..., X_p$ or depends on Y.

Weighted complete-cases analysis is an alternative of the CC analysis (Little, 1986; Horvitz, and Thompson, 1952; Cochran, 1968). Let r be the number of complete cases, the marginal mean of Y can be derived as $\hat{\mu} = (\sum_{i=1}^{r} w_i y_i) / (\sum_{i=1}^{r} w_i)$ or

$\hat{\mu} = (\sum_{i=1}^{r} w_i y_i)/n$. The weight, w, is the design weight or the probability of being selected in sample surveys without nonresponse. For missing data due to nonresponse the weight is the inverse of the probability of being observed. When the weight is unknown, we can estimate it based on a set of observed variables that characterizes the respondents and nonrespondents. One way is to group subjects into subclasses based on a small set of observed covariates. Within each subgroup the respondents are a random sample of the subpopulation and the weight is the inverse of the proportion of the respondents. When the number of covariates increases, sub-grouping will lead to a large number of subclasses and in this case, propensity weighting will be an alternative (Cochran, 1965; Rosenbaum and Rubin, 1983, 1984, Little, 1986). The propensity score is a scalar function of the observed covariates. One can estimate the propensity score by a logistic or probit regression of *M* on the observed covariates and the weight can be derived as the inverse of the propensity score. The potential draw back of the propensity weighting is that it is may yield estimates with large variances because respondents with very small propensity scores will be assigned huge weights, which may lead to out-of-range estimates for the means (Little and Rubin, 2002).

Complete-case analysis and weighted complete-case analysis delete subjects with missing values thus information contained in the covariates of the incomplete cases are lost. This loss of information may lead to less efficient estimators. To make full use of observed information we can use parametric approach to deal with missing data. For example, we can derive the marginal mean of *Y* based on a linear regression model $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i$, where ε_i is the error term with $\varepsilon_i \sim N(0, \sigma^2)$. We can solve this model by maximum likelihood (ML) approach (Little and Rubin, 2002; Anderson, 1957; Rubin, 1974). The marginal mean of *Y* can be derived as $\hat{Y} = n^{-1}(\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \hat{y}_i)$, with $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ij}$, where $\hat{\beta}_0, ..., \hat{\beta}_p$ are the maximum likelihood estimators based on the complete cases. An alternative to the ML estimators is to add a prior distribution for

the parameters $\beta_0,...,\beta_p$ and σ^2 and derive the posterior distribution of *Y* given the covariates and the unknown parameters, $p(Y \mid X_1,...,X_p,\beta_0,...,\beta_p,\sigma^2)$ (Gelman, Carlin, Stern and Rubin, 1995). Missing values of *Y* and the unknown parameters $\underline{\beta} = (\beta_0,...,\beta_p)$ and σ^2 are drawn iteratively by Gibbs' sampler or by Markov Chain Monte Carlo (MCMC) method (Casella and George, 1992; Geman and Geman, 1984). When the posterior distribution reaches stationary condition after *N*th iteration, *M* sets of data are created such that within each data every missing Y_i is substituted by an independent draw from the posterior distribution. For each dataset a posterior mean of *Y*, $\bar{Y}^{(l)}$, $l = 1,....,M$, is derived as the average of the observed values and the posterior draws. The marginal mean of *Y* is the average the posterior means over the *M* datasets. Usually *M* needs to be a large number. However, if we can assume approximate normality for the posterior distribution of $\underline{\beta}$ and σ^2 given the observed data, $p(\beta_0,...,\beta_p,\sigma^2 \mid X_1,...,X_p)$, we only need to create a small number of datasets to estimate the marginal mean of *Y*, which is the idea of multiple imputation (Little and Rubin, 2002; Rubin, 1978). For each dataset the missing values are replaced by independent posterior draws and the complete data analysis technique is applied to each imputed dataset. The marginal mean of *Y* can be derived using Rubin's combination rules (Rubin 1978, 1987, 1996; Rubin and Schenker, 1986; Barnard and Rubin, 1999). Let $\hat{\mu}_d$ be the estimated marginal mean of the *d*th dataset, $d = 1,...,D$, where D is the total number of imputed datasets, the marginal mean of *Y* is derived as $\hat{\mu} = \sum_{d=1}^{D} \hat{\mu}_d / D$.

The parametric approach described above is very efficient and yields consistent estimates if the model assumptions are correct. But the drawback is that it is very sensitive to model misspecification. In reality we can never guarantee the model assumptions are correct thus robust estimators are gaining more attention recently. Robins, Rotnitzky and Zhao (1994) and Rotnitzky, Robins and Scharfstein (1998) proposed a class of augmented orthogonal inverse probability-weighed estimators, which combine the features of the parametric prediction with the weighted estimation equations

(WEE). The marginal mean of Y can be derived by calibrating the predictions from a parametric model by adding mean of the weighted residuals,

$$\mu_y = E(E(Y \mid X_1,..., X_p)) + E[\frac{1-M}{\pi(Y)}(Y - E(Y \mid X_1,..., X_p)]$$

where $\pi(Y)$ is the probability of being observed. This leads to a calibration estimator of the form:

$$\hat{\mu} = n^{-1}(\sum_{i=1}^{n} \hat{y}_i) + n^{-1}(\sum_{i=1}^{r} \hat{w}_i(y_i - \hat{y}_i))$$

where $\hat{w}_i = 1/\Pr(M_i = 0 \mid X_1,..., X_P)$ is the estimated weight for the *i*th subject, and $\hat{y}_i$ is the prediction from a parametric model for the *i*th subject. There are three steps for the calibration method. Firstly a parametric model is fit to the complete cases and predictions are derived for all the subjects by substituting the covariates to the regression model. Secondly, the propensity score is estimated by a logistic regression or a probit model of M on $X_1,..., X_p$. Then the marginal mean of Y can be estimated by combining mean of the predictions with mean of the weighted residuals, where residuals are the differences of the observed values and the predicted values for the complete cases. This method has a double robustness property meaning that if either the prediction model is correctly specified or the weight is correctly estimated, the marginal mean of Y is consistent. This class of estimators is further extended in Robins and Rotnitzky (2001), Lunceford and Davidian (2004), Yu and Nan (2006).

An alternative way to achieve robustness is to weaken the model assumptions; for example, we can fit models with robust mean functions. One of the methods is the linear in the weight prediction (LWP). It includes the weight as a linear term in the imputation model (Scharfstein, Rotnitzky and Robins, 1999; Bang and Robins, 2005) as follows.

$$(Y \mid X_1,..., X_p; \beta) \sim N(g(X_1,...X_p; \beta) + \alpha * \hat{W}, \sigma^2)$$

where $\hat{W} = 1/\Pr(R = 1 \mid X_1,..., X_P)$ is the inverse of the estimated propensity score of respondents. Similar approach has been applied in the sample survey setting, where the weights are due to sampling rather than nonresponse (Sarndal, Swensson and Wretman, 2003 ; Firth D. and Bennett, 1998). The linear in the weight prediction method has a similar double robustness property as the calibration estimators meaning that if either the

mean function of *Y* given the covariates are correctly specified or the weight is correctly estimated then the marginal mean of missing variable *Y* will be consistent. Like the calibration method, the first step of fitting a linear in the weight model estimates the propensity score, for example by a logistic regression model or a probit model of *M* on $X_1,...,X_p$; in the second step, a regression of *Y* on the weight and the other covariates is fit parametrically.

Semiparametric and non-parametric method is another approach to yield robust mean functions by capturing the nonlinear relationship between the variables. In particular, with $p = 1$ and single covariate *X*, one version of this approach is to base imputations on the penalized spline model $y_i = s(x_i) + \varepsilon_i$ with truncated polynomial basis

$$s(x) = \beta_0 + \beta_1 x + ... + \beta_q x^q + \sum_{k=1}^{K} \beta_{qk} (x - \kappa_k)_+^q$$

where 1, $x,...,x^q,(x-\kappa_1)_+^q,...,(x-\kappa_k)_+^q$ is known as the truncated power basis of degree $q; \kappa_1 < .... < \kappa_K$ are selected fixed knots and *K* is the total number of knots (Eilers and Marx, 1996; Ruppert, Wand and Carroll, 2003; Ngo and Wand, 2004). The penalized least squares estimator $\hat{\beta} = (\hat{\beta}_0,...,\hat{\beta}_q,\hat{\beta}_{q1},...,\hat{\beta}_{qK})^T$ is obtained by minimizing

$$\sum_{i=1}^{n} \{y_i - \beta_0 - \sum_{j=1}^{q} \beta_j x^j - \sum_{k=1}^{K} \beta_{qk} (x - \kappa_k)_+^q\}^2 + \lambda^{2q} \beta^T D \beta$$

where λ is a smoothing parameter and $D = diag(0_{q+1}, 1_K)$. The fitted values are $\hat{y} = X(X^T X + \lambda^{2q} D)^{-1} X^T y$. This model can be fitted using a number of existing software packages, such as PROC MIXED in SAS (SAS, 1992; Ngo and Wand, 2004; Littell, Milliken, Stroup, and Wolnger. 1996; Ruppert, 2002) and lme() in S-plus (Pinheiro and Bates, 2000). With more than one covariate, one might extend this approach by fitting a multivariate spline. However, such models are subject to the curse of dimensionality when *p* is large, which relates to the difficulty of fitting nonparametric regression functions when the regressor space has high dimension. The Penalized Spline of Propensity Prediction (PSPP; Little and An 2004) method addresses this problem by restricting the spline to a particular function of covariates most sensitive to model misspecification, namely the propensity score. Little and An show that the PSPP method

yields an estimate of the marginal mean of the missing variable with a double robustness (DR) property, which means that the predicted marginal mean of *Y* will be consistent when either the mean function of *Y* given the covariates is correctly specified or the propensity score function is correctly specified. The robustness feature lies in the fact that the parametric function does not have to be correctly specified. A related approach is given by Zeng (2001), who reduces the dimension of the covariates to two, the propensity and a linear predictor, and then models the relationship of the outcome and these two variables by a bivariate nonparametric model.

For the first part of the dissertation, we simplify and extend the PSPP method. Little and An's method requires centering of the covariates before adding them to the model parametrically. We show this centering is not necessary and simplify the PSPP method considerably. We prove that this simplified version has the same DR property as the model proposed by Little and An (2004). The simplified PSPP method is much easier for the practitioners. We then extend the simplified PSPP method to derive the conditional mean(s) of a missing variable given a covariate. A stratified PSPP method is proposed to derive the subgroup means given a categorical covariate. For continuous covariate, we propose a bivariate PSPP method. Both of these extensions consider the interaction of the propensity score and the covariate. Simulations show that these extensions yield consistent conditional means under different mean and propensity structures. We apply the stratified PSPP method to an online weight loss study conducted by Kaiser Permanente (Couper, Peytchev, Little, Strecher and Rothert, 2005).

For the second part of the dissertation, we compare the PSPP method and several alternative doubly robust estimators. The PSPP method is based on the balance property of the propensity score, which means, conditioning on the propensity score and assuming MAR, missingness of *Y* does not depend on the covariates $X_1,...,X_p$ (Rosenbaum and Rubin, 1983). Since we do not know the true relationship of *Y* and the propensity score, we use a spline of the propensity score to impute the missing values and the resulting estimates have a double robustness property. The DR property can also be achieved by modeling the relationship parametrically, such as linear in the weight prediction method

and the calibration estimators. However emphasis in previous research has been on asymptotic properties of the estimates, namely consistency and achieving the semiparametric efficiency bound (Firth and Bennett, 1998; Scharfstein, Rotnitzky and Robins, 1999; Bang and Robins, 2005; Robins, Rotnitzky and Zhao, 1994; Scharfstein, Rotnitzky and Robins, 1999). Consistency is a relatively weak property, and does not guarantee good confidence coverage of inferences in small or moderate sized samples. Semiparametric efficiency is also a fairly weak property since it is asymptotic and does not necessarily guarantee efficiency in finite samples. For the second part of the dissertation we compare root mean square error (RMSE), width of confidence interval and non-coverage rate of the above approaches for a range of sample sizes, when the regression model is misspecified. We also compare these methods when the propensity score is wrongly specified.

In the third part of my dissertation, we extend the PSPP method to the monotone pattern of missing data, where variables can be arranged in a way that if Y_j is missing in a unit then $Y_{j+1,}Y_{j+2}$,…, Y_p are missing as well. Monotone pattern of missing data is common in longitudinal studies when some subjects drop out the study and do not return. We propose to impute the missing values in a stepwise procedure. The marginal propensity score is derived for each part of the missing variables. For the part where marginal propensity score is zero due to the missingness of the precedent variable(s), we cannot apply the PSPP method directly since there is no observed data with this propensity score. In this case we propose to borrow the propensity scores from the previous stages. Imputation of missing values is done in several steps according to the pattern of missing data. The part with least missing information is imputed first and then the imputed data is used to predict the missing values for the next part of data. Simulation studies show that the stepwise procedure yields satisfactory results. We illustrate our method by applying it to an online weight loss study conducted by Kaiser Permanente. We conclude the dissertation with a short discussion and future work in Chapter V.

CHAPTER II

EXTENSIONS OF THE PENALIZED SPLINE PROPENSITY PREDICTION (PSPP) METHOD OF IMPUTATION

Abstract

Little and An (2004) proposed a penalized spline of propensity prediction (PSPP) method of imputation of missing values that yields robust model-based inference under the missing at random assumption. The propensity score for a missing variable is estimated and a regression model is fit that includes the spline of the estimated propensity score as a covariate. The predicted unconditional mean of the missing variable has a double robustness (DR) property under misspecification of the imputation model. We show that a simplified version of PSPP, which does not center other regressors prior to including them in the prediction model, also has the DR property. We also propose two extensions of PSPP, namely stratified PSPP and bivariate PSPP, that extend the DR property to inferences about conditional means. These extended PSPP methods are compared with the PSPP method and simple alternatives in a simulation study and applied to an online weight loss study conducted by Kaiser Permanente.

Keywords: missing at random, propensity, penalized spline.

2.1 Introduction

Missing data problems are common in many applications of statistics. In this paper, we consider univariate nonresponse, where the missingness is confined to a single variable. Let $(Y, X_1, ..., X_p)$ denote a $p+1$ dimensional vector of variables with Y subject to missing values and $X_1, ..., X_p$ fully observed covariates. We consider here the problem of estimating the mean of Y, and the conditional mean of Y in subclasses defined by a categorical X-variable, and the regression coefficient of Y on a continuous X-variable.

Many statistical methods have been proposed for these problems. A simple approach is complete case analysis (CC), which deletes units with Y missing, so information contained in the deleted cases is lost. In the context of our problem, CC analysis yields a consistent estimate of the overall mean of Y if missingness does not depend on any of the variables, and consistent estimate of the conditional mean of Y given a covariate X_1 if the missing-data mechanism depends on X_1, but does not depend on Y or $X_2,...,X_p$. Another approach is to impute missing values based on a parametric model, for example a linear regression model $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i$, where ε_i is the error term with $\varepsilon_i \sim N(0,\sigma^2)$. One can estimate $(\beta_0,...,\beta_p)$ based on the complete cases and predict the missing values of Y by substituting X for that case into the regression equation. This approach is effective when the data are missing at random (Rubin 1976; Little and Rubin, 2002) and the regression model assumptions are correct, but can yield biased results when the model is misspecified. Semiparametric and nonparametric methods weaken the model assumptions and capture the nonlinear relationships between the variables. In particular, with p = 1 and single covariate X, one version of this approach is to base imputations on the penalized spline model $y_i = s(x_i) + \varepsilon_i$ with truncated polynomial basis

$$s(x) = \beta_0 + \beta_1 x + ... + \beta_q x^q + \sum_{k=1}^{K} \beta_{qk} (x - \kappa_k)_+^q \qquad (1)$$

where 1, $x,...,x^q,(x-\kappa_1)_+^q,...,(x-\kappa_k)_+^q$ is known as the truncated power basis of degree $q; \kappa_1 < .... < \kappa_K$ are selected fixed knots and K is the total number of knots (Eilers and Marx, 1996; Ruppert, Wand and Carroll, 2003; Ngo and Wand, 2004). The penalized least squares estimator $\hat{\beta} = (\hat{\beta}_0,...,\hat{\beta}_q,\hat{\beta}_{q1},...,\hat{\beta}_{qK})^T$ is obtained by minimizing

$$\sum_{i=1}^{n} \{y_i - \beta_0 - \sum_{j=1}^{q} \beta_j x^j - \sum_{k=1}^{K} \beta_{qk} (x - \kappa_k)_+^q\}^2 + \lambda^{2q} \beta^T D \beta$$

where λ is a smoothing parameter and D = $diag(0_{q+1},1_K)$. The fitted values are $\hat{y} = X(X^T X + \lambda^{2q} D)^{-1} X^T y$. This model can be fitted using a number of existing software packages, such as PROC MIXED in SAS (SAS, 1992; Ngo and Wand, 2004) and lme() in S-plus (Pinheiro and Bates, 2000). This imputation model is strictly speaking

parametric, but mimics a nonparametric method when K is large, since the form of the relationship between Y and X is very flexible.

With more than one covariate, one might extend this approach by fitting a multivariate spline. However, such models are subject to the curse of dimensionality when p is large, which relates to the difficulty of fitting nonparametric regression functions when the regressor space has high dimension. Penalized Spline of Propensity Prediction (PSPP; Little and An 2004) addresses this problem by restricting the spline to a particular function of covariates most sensitive to model misspecification, namely the propensity score. Little and An show that the PSPP method yields an estimate of the marginal mean of the missing variable with a double robustness (DR) property, described below in section 2.2. We propose a simplification of PSPP that does not center the regressors prior to including them in the prediction model.

Little and An (2004) did not consider whether the PSPP yields robust estimates for other parameters, such as conditional means or regression coefficients. In section 2.3 we provide examples to show that the PSPP method does not in general yield estimates of these parameters with the DR property. This motivates robust extensions of the PSPP method for estimating subgroup means and regression coefficients, which are described in sections 2.4 and 2.5. We apply the proposed methods to an online weight loss study in section 2.6, and section 2.7 presents concluding remarks.

2.2 Penalized Spline of Propensity Prediction (PSPP)

Let $(Y, X_1,..., X_p)$ denote a vector of variables with Y subject to missing values and $X_1,..., X_p$ fully-observed covariates. The missingness of Y depends only on $X_1,..., X_p$, so the missing data mechanism is missing at random (Rubin, 1976). Let M be an indicator variable with $M = 1$ when Y is missing and $M = 0$ when Y is observed. Define the logit of the propensity for Y to be observed as:

$$P^* = \text{logit}\left(\Pr(M = 0 \mid X_1,..., X_p)\right) \quad (2)$$

The key property of the propensity score is that, conditioning on the propensity score and assuming MAR, missingness of Y does not depend on $X_1,...,X_p$ (Rosenbaum and Rubin, 1983). Thus, the mean of Y can be written as

$$\mu_y = E[(1-M)Y] + E[M \times E(Y \mid P^*)] \tag{3}$$

Since the true relationship of Y and the propensity score is unknown, Little and An (2004) proposed to include the propensity score in the imputation model nonparametrically. This motivates the Penalized Spline of Propensity Prediction Method (PSPP), which is based on the following model:

$$\begin{aligned} &(X_2,...,X_p \mid P^*) \sim N_{p-1}((s_2(P^*),...,s_p(P^*)),\Sigma) \\ &(Y \mid P^*, X_2,...,X_p;\beta) \ \sim N(s(P^*) + g(P^*, X_2^*,...X_p^*;\beta),\sigma^2) \end{aligned} \tag{4}$$

where $N_k(\mu,\Sigma)$ denotes the *k*-variate normal distribution with mean μ and covariance matrix Σ, $s_j(P^*) = E(X_j \mid P^*)$, $j = 2,..., p,$ is a spline for the regression of X_j on P^* of the form (1); $X_j^* = X_j - s_j(Y^*)$ is the residual of the spline model and represents the part in X_j not explained by the propensity score; $s(P^*)$ is a spline of Y on P^* of the form (1) and g is a parametric function indexed by unknown parameter β with $g(Y^*,0,...,0;\beta) = 0$ for all β. One of the predictors, here X_1, is omitted from the g - function to avoid multicollinearity. The first step of fitting a PSPP model estimates the propensity score, for example by a logistic regression model of M on $X_1,...,X_p$; in the second step, the regression of Y on P^* is fit as a spline model with the other covariates included in the model parametrically in the g - function.

The predicted mean of *Y* from model (4) has the following DR property:

Theorem 1. Let $\hat{\mu}_y$ be the prediction estimator for (3) based on model (4), and assume MAR. Then $\hat{\mu}_y$ is a consistent estimator of μ_y if either (a) the mean of Y given $(P^*, X_2,...,X_p)$ in model (4) is correctly specified, or (b1) the propensity P^* is correctly specified, and (b2) $E(X_j \mid P^*) = s_j(P^*)$ for $j = 2,..., p$ and $E(Y \mid P^*) = s(P^*)$. The

robustness feature derives from the fact that the regression function g does not have to be correctly specified (Little and An, 2004).

The covariates $X_2^*,...,X_p^*$ in this theorem are centered by regressing $X_2,...,X_p$ on splines of P^* and taking residuals. A simpler method adds $X_2,...,X_p$ directly to the regression, without centering. We now show that this method also has the DR property:

Theorem 2. The PSPP method based on model (4) can be simplified as follows:

$$(Y \mid P^*, X_2,..., X_p; \beta) \sim N(s(P^*) + g(P^*, X_2,...X_p; \beta), \sigma^2) \quad (5)$$

that is, the covariates $X_2,...,X_p$ enter the parametric function g without centering. Let $\hat{\mu}_y$ be the prediction estimator for (3) based on model (5), and assume MAR, then $\hat{\mu}_y$ has the same DR property as that derived from model (4) (see appendix for proof). For this reason, we focus on the uncentered version of the PSPP method for the remainder of the paper.

2.3 PSPP is not doubly robust for subgroup means.

The DR property of PSPP for estimating the marginal mean of *Y* does not extend to estimates of conditional means, such as means in subgroups defined by a categorical covariate X_1. The next two examples illustrate this statement. The first example illustrates the intuitively obvious fact that for estimating the conditional mean of *Y* given X_1, the PSPP method needs to include X_1 as a predictor in the model for *Y*. The second example illustrates that inclusion of X_1 as a predictor in the model for *Y* is not sufficient to avoid bias with the PSPP method. This limitation is then addressed with the extended versions of the method.

Example 1. PSPP for estimating a conditional mean: including the subgroup variable in the model for *Y* is necessary. We simulate 500 datasets with 500 subjects, with categorical covariate X_1, continuous covariate X_2 and continuous response variable *Y*, where X_1, X_2 are independent with $X_1 \sim$ multinomial $(0.5, 0.3, 0.2)$, $X_2 \sim N(0,1)$, and

$$Y \mid X_1, X_2 \sim N\left(\mu(X_1, X_2), 1\right),$$
$$\mu(X_1, X_2) = I[X_1 = 1] + 3 \times I[X_1 = 2] + 5 \times I[X_1 = 3] + 10 X_2$$

where *I*[] denotes an indicator for the event in the parenthesis. We create missing values of *Y* from the response propensity model:

$$\text{logit}\,(P(M = 0 \mid X_1, X_2)) = X_2 + 0.5 * I[X_1 = 1] - 0.5 * I[X_1 = 2]$$

We impute the missing values of *Y* using predicted means from the following methods:

(a) A correctly-specified ANCOVA model of Y given X_1, X_2, which we denote $[X_1 + X_2]$.

(b) An incorrectly specified regression model for *Y* that omits X_2, namely $[X_1]$.

(c) The PSPP Method with null *g* function, which we denote $[s(P^*_{correct})]$. The propensity score $P^*_{correct}$ is modeled as an additive function of X_1 and X_2 and hence is correctly specified and conditions on X_1.

(d) Model (c) with X_1 included, namely $[s(P^*_{correct}) + X_1]$. This model correctly specifies the mean of Y given the covariates, since it includes the main effects of X_1 and X_2.

(e) The PSPP Method with null *g* function and incorrectly specified propensity score, modeled as a linear function of X_2 alone, which we denote $[s(P^*_{wrong})]$.

(f) Model (e) with X_1 included, namely $[s(P^*_{wrong}) + X_1]$. This model correctly specifies the mean of Y given the covariates, since it includes the main effects of X_1 and X_2.

For all the penalized spline methods in this paper, we choose 20 equally spaced fixed knots and a truncated linear basis. We estimate the marginal mean of *Y* and the conditional means of *Y* given X_1 as the average of observed and imputed values from these methods. For comparison purposes, we also show estimates from the data before deletion (BD) and estimates based on the complete cases (CC). Empirical bias (Bias), empirical standard deviation (STD) and root mean square error (RMSE) over the 500 replications are summarized in Tables 2.1A and 2.1B. CC analysis yields estimates with large biases and RMSEs. The correctly specified ANCOVA model (a) yields unbiased estimates close to the BD estimates. The wrongly specified ANCOVA model (b) yields

biased parameter estimates, with large biases and RMSEs. For the PSPP method, inclusion of X_1 in the model is important for subgroup mean estimation. Without X_1 in the model, the PSPP method (c) yields small empirical bias for the marginal mean estimate and a large empirical bias for the conditional means of *Y* given X_1, even though the propensity score model is correct and conditions on both X_1 and X_2; including X_1 in the PSPP method (d) yields estimates of the marginal mean of *Y* and conditional means of *Y* given X_1 with small empirical biases, and STDs and RMSEs very close to those of BD. When neither the propensity score nor the mean function is correctly specified, the PSPP method (e) yields biased results; but the bias is removed in model (f) by including X_1, since then the regression is correctly specified.

Example 2. PSPP for estimating a conditional mean: including the subgroup variable in the model for *Y* is not sufficient. We now generate X_1 and X_2 as in Example 1; but the mean of *Y* given X_1 and X_2 is simulated to include both a quadratic term in X_2 and interactions between X_1 and X_2:

$$Y \mid X_1, X_2 \sim N(\mu(X_1, X_2), 1),$$
$$\mu(X_1, X_2) = I[X_1 = 1] + 3 \times I[X_1 = 2] + 5 \times I[X_1 = 3] + 10X_2 + X_2^2$$
$$-1 + 4 \times I[X_1 = 1] \times X_2 - 10 \times I[X_1 = 2] \times X_2$$

The logistic regression of *M* is additive in X_1 and a quadratic function of X_2:

$$\text{logit}(P(M = 0 \mid X_1, X_2)) = 0.5 \times I[X_1 = 1] - 0.25 \times I[X_1 = 2] + 0.25 \times X_2^2 + 0.5 \times X_2 - 0.5$$

We simulate 500 datasets with sample size of 1000 each. We impute the missing *Y* as predicted means from the following methods:

(a) A correctly-specified regression model for *Y*, namely $[X_1 + X_2 + X_1 \times X_2 + X_2^2]$.

(b) The PSPP model with null *g*-function, namely $[s(P^*)]$. The propensity score P^* is modeled as an additive function of X_1, X_2 and X_2^2 and hence is correctly specified.

(c) PSPP with X_1 included, that is, $[s(P^*) + X_1]$.

(d) PSPP with X_2 and X_2^2 included, namely, $[s(P^*) + X_2 + X_2^2]$.

The correctly-specified ANCOVA model yields estimates with small empirical bias and RMSE close to BD (Table 2.2). CC analysis and the wrongly specified ANCOVA model yield biased estimates. The PSPP methods (b) – (d) yield estimates for the marginal mean of *Y* with small empirical bias, but are clearly biased for the conditional means of *Y* given X_1 and *Y* given X_2. In particular, unlike Example 1, adding X_1 to the *g*-function does not correct the misspecification of the mean of *Y* given X_1, since the estimates of the conditional means are still biased.

In the second example, the PSPP method $[s(P^*) + X_1]$ assumes that for different levels of X_1, the splines of *Y* on P^* have the same shape; since the true model includes the interaction between X_1 and X_2, this assumption is violated, and it is this fact that leads to bias for the conditional means. One solution is to include the interaction of propensity score and X_1 into the model, yielding a stratified PSPP method discussed in the next section.

2.4 Stratified Penalized Spline Propensity Prediction for subgroup means

Let $I_c = 1$ if $X_1 = c$; $I_c = 0$ if $X_1 \neq c$, $c = 1,...,C$, where *C* is the total number of categories of X_1. The stratified PSPP method is based on the following model:

$$(Y \mid P^*, X_1,..., X_p; \beta) \sim N(\sum_{c=1}^{C} I_c s_c(P^*) + g(P^*, X_1, X_2,..., X_{p-1}; \beta), \sigma^2) \qquad (6)$$

Where *g* is a parametric function indexed by unknown parameter β as before, with X_p dropped to avoid multicollinearity; $I_c s_c(P^*) = I_c(\gamma_{0c} + \sum_{j=1}^{q} \gamma_{jc}(P^*)^j + \sum_{k=1}^{K} \gamma_{qkc}(P^* - \kappa_k)_+^q)$ is the fitted curves for the *c*th level of X_1. Within each level of X_1,

$$E(Y \mid P^*, X_1 = c, X_2,..., X_{p-1}; \beta) = s_c(P^*) + g(P^*, X_1 = c, X_2,..., X_{p-1}; \beta).$$

Note that this method is not the same as applying PSPP within strata defined by X_1, since the *g*-function does not necessarily include the interactions of X_1 with the other covariates. This method yields consistent estimates for the conditional means of *Y* given

X_1. The marginal mean of *Y* is a weighted average of conditional means, which again has the double robustness property (see appendix for proof).

Example 2 continued

Row (e) in Table 2.2 shows the results of applying stratified PSPP to the data in Example 2. The empirical bias is small for the marginal mean of *Y* and the subgroup means of *Y* given X_1, and the RMSE for these parameters is only slightly larger than for BD. Thus stratified PSPP has fixed the bias for the subgroup means in the PSPP methods. On the other hand the empirical bias remains large for the coefficients of the regression of *Y* on X_2. For those parameters we need another extension of PSPP, which we now describe.

2.5 A Bivariate PSPP Method for estimating the conditional mean of *Y* given a continuous covariate.

In this section we consider estimating the conditional mean of *Y* given a continuous variable X_2, based on a regression model for *Y* given X_2. To estimate the regression coefficients in this case we need to assume that the regression of *Y* on X_2 is correctly specified; for concreteness we assume it is linear with mean $E(Y \mid X_2) = \beta_0 + \beta_1 X_2 + \beta_2 X_2^2$. To yield consistent parameter estimates for the regression coefficients, we now include the interaction of propensity score and X_2 in the model for predicting the missing values of *Y*. Specifically, we propose the following bivariate PSPP method, based on the model:

$$(Y \mid P^*, X_1, X_2, ..., X_p; \beta) \sim N(s(P^*, X_2) + g(P^*, X_1, X_2, ..., X_{p-1}; \beta), \sigma^2) \qquad (7)$$

where g is a parametric function; $s(P^*, X_2)$ is a bivariate P-spline of P^* and X_2. Estimation of the bivariate smoothing function $s(P^*, X_2)$ requires bivariate basis functions, which can be derived in several different ways. A natural extension of the truncated linear basis for one dimension is to form all the pair-wise products of the basis functions. The resulting bivariate basis is called the tensor product basis (Ruppert, Wand and Carroll, 2003). With this basis, the bivariate function $s(P^*, X_2)$ can be written as

$$s(P^*, X_2) = \alpha_0 + \alpha_1 P^* + \sum_{k=1}^{K_1} \gamma_{1k}(P^* - \kappa_{1k})_+ + \alpha_2 X_2 + \sum_{k'=1}^{K_2} \gamma_{2k'}(X_2 - \kappa_{2k'})_+ + \alpha_3 P^* X_2$$
$$+ \sum_{k=1}^{K_1} \gamma_{3k} X_2 (P^* - \kappa_{1k})_+ + \sum_{k'=1}^{K_2} \gamma_{4k'} P^* (X_2 - \kappa_{2k'})_+ + \sum_{k=1}^{K_1} \sum_{k'=1}^{K_2} \gamma_{5kk'} (P^* - \kappa_{1k})_+ (X_2 - \kappa_{2k'})_+$$

where $\kappa_{11} < ... < \kappa_{1K_1}$ and $\kappa_{21} < ... < \kappa_{2K_2}$ are selected fixed knots for propensity score and X_2 respectively. In this paper we choose 5 equally spaced knots for each variable when fitting the bivariate splines using a tensor product basis.

Example 2 Continued

Row (f) in Table 2.2 shows estimates of the parameters when missing values are imputed using the bivariate PSPP method. This method yields estimates of the coefficients of the regression of Y on X_2 with small empirical biases and RMSEs only slightly higher than those of BD analysis.

The conditional means of Y given X_1 from bivariate PSPP are biased. To get consistent estimates of both the conditional means of Y given X_1 and conditional mean of Y given X_2, a model is needed that includes the interaction between the propensity score and X_1 and the interaction between the propensity score and X_2. This motivates the following combination of the stratified PSPP and bivariate PSPP models:

$$(Y \mid P^*, X_1, ..., X_p; \beta) \sim N(\sum_{c=1}^{C} I_c s_c(P^*) + s(P^*, X_2) + g(P^*, X_2, ..., X_p; \beta), \sigma^2)$$

where $I_c s_c(P^*)$ and $s(P^*, X_2)$ are defined as in sections 4 and 5 respectively. When we applied this method to the second simulation, a small number (8) of the 500 samples failed to converge, but results for the other samples indicate that empirical bias from this model is small for both the conditional mean of Y given X_1 and the conditional mean of Y given X_2 (Table 2.2, row (g)).

2.6 An Example: Online Weight Loss Study

To illustrate our proposed approach, we consider data from an online weight loss study conducted by Kaiser Permanente (Couper et al., 2005). The study randomized approximately 4,000 subjects to the treatment or the control group. For the treatment group, the weight loss information provided online was tailored to the subjects based on their answers to an initial survey, which contained baseline measurements such as baseline weight, motivation to weight loss, etc; for the control group, information provided online was the same for all the subjects. At 3 months, a second survey was sent to all of the participants, which collected follow-up measurements such as current weight. Our goal is to compare the short-term treatment effects; in particular, we compare the reduction of the body mass index (BMI), defined as difference of 3-month BMI and baseline BMI.

There were 2059 subjects in the treatment group and 1956 subjects in the control group at the baseline. At 3 month 623 subjects in the treatment group and 611 subjects in the control group responded to the second survey. We assume the data are missing at random. Subjects in the treatment group who remained in the study have much lower baseline BMI than those who dropped out ($P<0.001$), but this differences is not seen in the control group ($P=0.47$); On the other hand, for the control group subjects who remained in the study have better baseline health, as measured by the number of previous diseases, than those who dropped out of the study ($P<0.01$); this differences was not seen in the treatment group ($P=0.56$). These differences suggest that interactions between treatment and baseline covariates need to be included when estimating the propensity scores.

We estimate the propensity scores by a logistic regression, with the inclusive criterion of retaining all variables with P-values less than 0.20. The final model includes the following covariates: baseline BMI; number of previous disease; baseline self care; which is harder–eating less or being active; baseline exercise support; baseline activity level; baseline eating topology; education; ethnic identity; treatment; interaction of treatment and baseline BMI; interaction of treatment and baseline eating topology;

interaction of treatment and baseline activity level; interaction of treatment and number of previous disease; interaction of treatment and which is harder–eating less or being active.

We apply the PSPP method and the stratified PSPP method to the data as follows:

(a) PSPP method with null *g*-function, denoted as $[s(P^*)]$, where P^* is the propensity scores defined in section 2.

(b) Model (a) with treatment as a covariate, denoted as $[s(P^*) + \text{treatment}]$.

(c) Model (b) with baseline covariates, denoted as $[s(P^*) + \text{treatment} + g(\text{baseline vars})]$.

(d) Stratified PSPP method with null *g*-function, denoted as $[\sum_{c=1}^{2} I_c s_c(P^*)]$.

(e) Model (d) with baseline covariates, denoted as $[\sum_{c=1}^{2} I_c s_c(P^*) + g(\text{baseline vars})]$.

The baseline covariates in the *g*-function of model (c) and (e) include: ethnic identity; baseline medical advice; baseline eating topology; baseline cardio exercise; baseline activity level; baseline BMI; number of previous disease; number of weigh loss methods tried; motivation of weigh loss; which is harder–eating less or being active.

Results are summarized in Table 2.3. Empirical Standard errors (SE) and the corresponding confidence intervals are obtained from 200 bootstrap samples. The treatment group has a larger reduction of BMI after 3 month (-0.91 (0.09)) compared to the control group (-0.45 (0.10)) based on the complete case analysis. The stratified PSPP method (model d and e) and the PSPP method with the treatment as a covariate (model b and c) yield similar results, with the reduction of BMI ranging from -0.95 to -1.01 for the treatment group and -0.40 to -0.46 in the control group. The 95% confidence intervals for the treatment group do not overlap with the control group suggesting a treatment effect on the weight loss (model b, c, d, e). On the other hand, the PSPP method without treatment as a covariate does not shown the treatment effect (95% CI (-0.96, -0.65) for the treatment; 95% CI (-0.76, -0.47) for the control). Adding *g* function into the model does not affect bias but improves efficiency (model c and e).

2.7 Discussion

We have shown that the PSPP method yields an estimate of the marginal mean of *Y* with a double robustness property, without the need to center the covariates in the *g* function. However the PSPP method lacks this property for conditional mean estimation. We have proposed two extensions of PSPP that extend the double robustness property to conditional means, namely stratified PSPP for a categorical predictor, and bivariate PSPP for a continuous predictor. The key property of these extensions is that they include in the prediction model the interaction of the propensity score and the conditioning variable that defines the estimand of interest. Simulations are presented as empirical evidence of the robustness of these extensions.

We estimate the bivariate function $s(P^*, X_2)$ using a P-spline with a tensor product basis, but other spline fitting methods could also be applied. One choice is to use a thin plate spline (Green and Silverman, 1994; Wood, 1999). To estimate $s(P^*, X_2)$, we need to find the function $g = g(\underline{t}) = g(t_1, t_2)$ minimizing

$$\sum_{i=1}^{n}(y_i - g(t_1,t_2))^2 + \lambda\iint[(\frac{\partial^2 g}{\partial {t_1}^2})^2 + 2(\frac{\partial^2 g}{\partial t_1 \partial t_2})^2 + (\frac{\partial^2 g}{\partial {t_2}^2})^2]dt_1 dt_2$$

where the *g* function has the form

$$g(\underline{t}) = \theta_0 + \sum_{j=1}^{M}\theta_j\phi_j(\underline{t}) + \sum_{j=1}^{n}\delta_j E_2(\underline{t} - \underline{t_j})$$

with $E_2(s) = \frac{1}{2^3\pi}\|s\|^2 \ln(\|s\|)$; $\phi_j(t)$ are linearly independent functions of $\underline{t}$ with $\underline{t} \in R^2$ and λ is the smoothing parameter. This model can be fit using the tpspline procedure from SAS (SAS, 1992; Ngo and Wand 2004; Wand 2003). We also fitted thin plate splines for the simulation study in section 5 but found some samples failed to yield estimates due to negative variance estimates. For the other samples the results from the tpsline procedure are comparable to those from a P-spline with a tensor product basis.

More generally, a PSPP method that yields doubly robust estimates of the conditional mean of *Y* given a subset of the covariates $(X_1,...,X_s)$, $s < p$, requires

inclusion of the interactions between the propensity score and $(X_1,...,X_s)$; clearly the curse of dimensionality comes increasingly into play as the size of *s* increases. A natural question is whether these propensity score methods can be extended to yield robust estimates for the regression given the complete set of covariates, such as, $(X_1,...,X_p)$. We note that in our setting the cases with *Y* missing contribute no information to this regression, so there is no gain in developing an imputation model. If it is the covariates rather than the outcome that have missing values, however, then the incomplete cases do include information, and it remains an open question whether propensity methods can be used to increase the robustness of inference in such situations. This question deserves future study.

We use a smoothing spline function to model the relationship between *Y* and the propensity score and our method has a DR property. The DR property can also be achieved by modeling the relationship parametrically. One method is to include the inverse of the propensity score as a linear term in the imputation model (Firth and Bennett, 1998; Bang and Robins, 2005). Another approach is to calibrate the predictions from a parametric model by adding means of the weighted residuals, with weights equal to inverse of the propensity scores (Robins, Rotnitzky and Zhao, 1994; Scharfstein, Rotnitzky and Robins, 1999). We are currently conducting simulations to compare the performance of these methods with the PSPP method, and results will be reported in a future paper.

Acknowledgements: this research is supported by CECCR Center grant P50 CA101451. We thank Trivellore Raghunathan for assistance with Theorem 2.

Table 2.1 Example 1: Empirical Bias, Standard Deviation (SD) and Root Mean Squared Error (RMSE) for (A) Marginal mean of Y , and (B) Conditional Mean of Y given X_1. Entries are multiplied by 100.

(A) Marginal Mean of Y

Methods	Bias	STD	RMSE
BD	0	45	35
CC	368	58	368
(a)Correct ANCOVA $[X_1 + X_2]$	0	45	36
(b)Wrong ANCOVA $[X_1]$	398	58	398
(c)PSPP $[s(P^*_{correct})]$	-2	50	40
(d)PSPP $[s(P^*_{correct}) + X_1]$	0	45	36
(e)PSPP $[s(P^*_{wrong})]$	-20	47	41
(f)PSPP $[s(P^*_{wrong}) + X_1]$	0	45	36

(B) Conditional Mean of Y given X_1

Methods	$X_1 = 1$			$X_1 = 2$			$X_1 = 3$		
	Bias	STD	RMSE	Bias	STD	RMSE	Bias	STD	RMSE
BD	-3	63	51	6	82	67	-1	98	77
CC	328	78	328	505	118	505	416	124	416
(a) Correct ANCOVA $[X_1, X_2]$	-3	64	51	7	83	68	-1	98	78
(b)Wrong ANCOVA $[X_1]$	328	78	328	505	118	505	416	124	416
(c)PSPP $[s(P^*_{correct})]$	213	69	214	-271	111	271	-139	141	162
(d)PSPP $[s(P^*_{correct}) + X_1]$	-3	64	51	7	84	69	-1	99	78
(e)PSPP $[s(P^*_{wrong})]$	43	65	63	-44	84	76	-145	107	154
(f)PSPP $[s(P^*_{wrong}) + X_1]$	-3	64	52	7	84	68	-1	99	78

Table 2.2 Example 2: Empirical Bias, Root Mean Squared Error (RMSE) and Coverage rate (Cov) for (A) Marginal mean of Y, (B) Conditional Mean of Y given X_1, and (C) Intercept and Slopes for Regression of Y on X_2, X_2^2. Entries are multiplied by 100.

Methods	Overall Mean			Conditional mean given X_1									Coefficients of conditional mean given X_2								
				X_1=1			X_1=2			X_1=3			Intercept			X_2			X_2^2		
	Bias	RMSE	Cov	Bias	RMSE	Cov	Bias	RMSE	Cov	Bias	RMSE	Cov	Bias	RMSE	Cov	Bias	RMSE	Cov	Bias	RMSE	Cov
BD	-1	30	93	-1	53	93	0	8	94	0	56	95	0	20	95	-2	26	95	1	25	95
CC	199	199	7	280	280	15	28	29	71	274	274	38	-21	33	89	81	84	58	-17	35	89
(a) Correct Model $[X_1 + X_2 + X_1 \times X_2 + X_2^2]$	-1	30	93	-1	53	93	0	10	94	1	58	95	0	19	95	-2	26	95	1	24	95
(b) PSPP $[s(P^*)]$	-3	34	95	112	116	66	-91	93	43	-154	156	53	-94	94	20	-163	163	1	93	93	20
(c) PSPP $[s(P^*) + X_1]$	-3	33	94	47	69	89	-121	122	28	53	90	94	-100	100	14	-139	139	17	98	99	19
(d) PSPP $([s(P^*) + X_2 + X_2^2])$	-3	33	95	50	70	89	-1	47	97	-137	146	72	37	43	78	17	40	92	-39	48	78
(e) Stratified PSPP ($Y = \sum I_c s_c(P^*)$)	-1	31	94	0	54	95	-4	13	92	4	60	96	-81	81	41	-86	87	60	81	81	43
(f) Bivariate PSPP ($Y = s(P^*, X_2)$)	-1	30	94	10	54	94	18	30	98	-60	87	93	2	22	97	2	29	96	-3	26	96
(g) Stratified_Bivariate PSPP ($Y = \sum I_c s_c(P^*) + s(P^*, X_2)$)	-2	30	94	-2	53	93	-7	16	98	5	59	96	4	23	96	4	30	95	-5	29	95

Table 2.3 BMI reduction within groups

Method	Treatment		Control	
	Mean (SE)	95% CI	Mean (SE)	95% CI
Complete Case Analysis	-0.91 (0.09)	(-1.09, -0.73)	-0.45 (0.10)	(-0.64, -0.25)
(a) PSPP $[s(P^*)]$	-0.80 (0.08)	(-0.96, -0.65)	-0.61 (0.07)	(-0.76, -0.47)
(b) PSPP $[s(P^*) + \text{treatment}]$	-0.95 (0.11)	(-1.16, -0.74)	-0.46 (0.10)	(-0.66, -0.26)
(c) PSPP $[s(P^*) + \text{treatment} + g(\text{baseline covariates})]$	-0.97 (0.10)	(-1.16, -0.78)	-0.46 (0.09)	(-0.64, -0.27)
(d) Stratified PSPP $[\sum_{c=1}^{2} I_c s_c(P^*)]$	-1.01 (0.11)	(-1.22, -0.79)	-0.40 (0.10)	(-0.59, -0.21)
(e) Stratified PSPP $[\sum_{c=1}^{2} I_c s_c(P^*) + g(\text{baseline covariates})]$	-1.00 (0.10)	(-1.20, -0.80)	-0.42 (0.09)	(-0.60, -0.23)

*SE and CI denote empirical standard error and confidence interval. SE and 95% CI are based on 200 bootstrap samples.

Appendix
(1)Proof of Theorem 2.

Lemma 1: Let

$$X_1 = \begin{pmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{1n} \end{pmatrix},\ X_2 = \begin{pmatrix} x_{21} & \cdots & x_{M1} & (x_1 * x_2)_1 & \cdots & (x_1 * x_M)_1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{2n} & \cdots & x_{Mn} & (x_1 * x_2)_n & \cdots & (x_1 * x_M)_n \end{pmatrix},\ Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

Where X_1 is a matrix containing variable x_1; X_2 contains the other covariates and interactions of x_1 and the other covariates; Y is a vector of the response variable. Let

$$X_{(1)} = \begin{pmatrix} f_1((X_1)_1) & \cdots & f_{N1}((X_1)_1) \\ \vdots & \ddots & \vdots \\ f_1((X_1)_n) & \cdots & f_{N1}((X_1)_n) \end{pmatrix} \text{ and } X_{(2)} = \begin{pmatrix} g_1((X_2)_1) & \cdots & g_{N2}((X_2)_1) \\ \vdots & \ddots & \vdots \\ g_1((X_2)_n) & \cdots & g_{N2}((X_2)_n) \end{pmatrix} \text{ be}$$

matrices that contain functions of X_1 and X_2 as columns.

Suppose we have the following models:

(a) Linear regression model of Y given X_1, X_2 with $E_A(Y \mid X_1, X_2) = X_{(1)} * \underline{\gamma}_1 + X_{(2)} * \underline{\gamma}_2$, where $E_A(Y \mid X_1, X_2)$ is the conditional mean of Y given the covariates X_1, X_2 under the assumed model. Let $\underline{\hat{\gamma}}_1, \underline{\hat{\gamma}}_2$ be the least squares estimates of $\underline{\gamma}_1$ and $\underline{\gamma}_2$, the predicted values of Y is written as $\hat{Y}_1(X_1, X_2) = X_{(1)} * \underline{\hat{\gamma}}_1 + X_{(2)} * \underline{\hat{\gamma}}_2$.

(b) Linear regression model of Y given X_1 with $E_A(Y \mid X_1) = X_{(1)} * \underline{\beta}_1$, where $E_A(Y \mid X_1)$ is the conditional mean of Y given the covariates X_1. Let $\underline{\hat{\beta}}_1$ be the least squares estimate of $\underline{\beta}_1$, the predicted values of Y is $\hat{Y}_2(X_1) = X_{(1)} * \underline{\hat{\beta}}_1$.

(c) Linear regression model of $g_i(X_2)$ given X_1 with $E_A(g_i(X_2) \mid X_1) = X_{(1)} * \underline{\delta}_i$, $i = 1, ..., N2$, where $E_A(g_i(X_2) \mid X_1)$ is the conditional mean of $g_i(X_2)$ given the covariates X_1. Let $\underline{\hat{\delta}}_i$ be the least squares estimates of $\underline{\delta}_i$, the predicted values of $g_i(X_2)$ is $\hat{g}_i(X_2) = X_{(1)} * \underline{\hat{\delta}}_i$. Let $\hat{X}_{(2)} = [\hat{g}_1(X_2), ..., \hat{g}_{N2}(X_2)]$ and $\underline{\hat{\delta}} = [\underline{\hat{\delta}}_1, ..., \underline{\hat{\delta}}_{N2}]$. Substitute $\hat{X}_{(2)}$ into $\hat{Y}_1(X_1, X_2)$ of model (a) and obtain $\hat{Y}_2^*(X_1) = X_{(1)} * \underline{\hat{\gamma}}_1 + \hat{X}_{(2)} * \underline{\hat{\gamma}}_2$.

Then $\hat{Y}_2^*(X_1) = \hat{Y}_2(X_1)$.

Proof:

To prove $\hat{Y}_2^*(X_1) = \hat{Y}_2(X_1)$, we need to show that $\underline{\hat{\beta}}_1 = \underline{\hat{\gamma}}_1 + \underline{\hat{\delta}} * \underline{\hat{\gamma}}_2$.

Let $X = [X_{(1)}, X_{(2)}]$, $H = X(X^T X)^{-1} X^T$, $H_1 = X_{(1)}(X_{(1)}^T X_{(1)})^{-1} X_{(1)}^T$.

From model (a): $Y = X_{(1)}\underline{\hat{\gamma}}_1 + X_{(2)}\underline{\hat{\gamma}}_2 + (I - H)Y$ (1)
Multiply (1) by $I - H_1$ and obtain
$(I - H_1)Y = (I - H_1)X_{(1)}\underline{\hat{\gamma}}_1 + (I - H_1)X_{(2)}\underline{\hat{\gamma}}_2 + (I - H_1)(I - H)Y$

Noting that:
(i) $(I - H_1)Y = Y - X_{(1)}\underline{\hat{\beta}}_1$
(ii) $(I - H_1)X_{(1)} = 0$
(iii) $(I - H_1)X_{(2)} = X_{(2)} - X_{(1)}\underline{\hat{\delta}}$
(iv) $H_1(I - H)Y = X_{(1)}(X_{(1)}^T X_{(1)})^{-1} X_{(1)}^T (I - H)Y = 0$ since that
$X^T(I - H)Y = X^T Y - X^T(X(X^T X)^{-1} X^T)Y = 0$

We have
$Y - X_{(1)}\underline{\hat{\beta}}_1 = (X_{(2)} - X_{(1)}\underline{\hat{\delta}})\underline{\hat{\gamma}}_2 + (I - H)Y$, which means
$Y = X_{(1)}(\underline{\hat{\beta}}_1 - \underline{\hat{\delta}} * \underline{\hat{\gamma}}_2) + X_{(2)}\underline{\hat{\gamma}}_2 + (I - H)Y$

So $\underline{\hat{\beta}}_1 - \underline{\hat{\delta}} * \underline{\hat{\gamma}}_2 = \underline{\hat{\gamma}}_1 \rightarrow \underline{\hat{\beta}}_1 = \underline{\hat{\delta}} * \underline{\hat{\gamma}}_2 + \underline{\hat{\gamma}}_1$

Now need to show for the penalized smoothing splines we have the same property.

Corollary:

(d) Regress Y on X_1, X_2 with $E_A(Y \mid X_1, X_2) = s_1(X_1; \underline{\gamma}_1) + g(X_2; \underline{\gamma}_2) = s_1(X_1; \underline{\gamma}_1) + X_{(2)}\underline{\gamma}_2$, where $E_A(Y \mid X_1, X_2)$ is the conditional mean of Y given the covariates X_1, X_2 under the assumed model; $s_1(X_1; \underline{\gamma}_1)$ is a spline of X_1 indexed by the parameter $\underline{\gamma}_1$; $g(X_2; \underline{\gamma}_2)$ is a parametric function indexed by the parameter $\underline{\gamma}_2$. Let $\underline{\hat{\gamma}}_1, \underline{\hat{\gamma}}_2$ be the restricted maximum likelihood estimates of $\underline{\gamma}_1$ and $\underline{\gamma}_2$, the predicted values of Y is written as $\hat{Y}_1(X_1, X_2) = s_1(X_1; \underline{\hat{\gamma}}_1) + X_{(2)} * \underline{\hat{\gamma}}_2$.

(e) Regress Y on X_1 with $E_A(Y \mid X_1) = s_1(X_1; \underline{\beta}_1)$, where $E_A(Y \mid X_1)$ is the conditional mean of Y given the covariates X_1 under the assumed mode; $s_1(X_1; \underline{\beta}_1)$ is a spline of X_1 indexed by the parameter $\underline{\beta}_1$. Let $\underline{\hat{\beta}}_1$ be the restricted maximum likelihood estimates of $\underline{\beta}_1$, the predicted values of Y is $\hat{Y}_2(X_1) = s_1(X_1; \underline{\hat{\beta}}_1)$.

(f) Regress $g_i(X_2)$ on X_1 with $E_A(g_i(X_2)\mid X_1)=s_1(X_1;\underline{\delta}_i)$, $i=1,...,N2$, where $E_A(g_i(X_2)\mid X_1)$ is the conditional mean of $g_i(X_2)$ given the covariates X_1 under the assumed model; $s_1(X_1;\underline{\delta}_i)$ is a spline of X_1 indexed by the parameter $\underline{\delta}_i$. Let $\hat{\underline{\delta}}_i$ be the restricted maximum likelihood estimates of $\underline{\delta}_i$, the predicted values of $g_i(X_2)$ is $\hat{g}_i(X_2)=s(X_1;\hat{\underline{\delta}}_i)$. Let $\hat{X}_{(2)}=[\hat{g}_1(X_2),...,\hat{g}_{N2}(X_2)]$ and $\hat{\underline{\delta}}=[\hat{\underline{\delta}}_1,...,\hat{\underline{\delta}}_{N2}]$. Substitute $\hat{X}_{(2)}$ into $\hat{Y}_1(X_1,X_2)$ of model (a) and obtain $\hat{Y}_2^*(X_1)=s(X_1;\hat{\underline{\gamma}}_1)+\hat{X}_{(2)}*\hat{\underline{\gamma}}_2$.

Then $\hat{Y}_2^*(X_1)\to\hat{Y}_2(X_1)$ as $n\to\infty$.

Proof:

Consider the penalized spline with the linear basis:

Let $X_{(1)}=\begin{pmatrix}1 & x_{11} & (x_{11}-k_1)_+ & \cdots & (x_{1n}-k_k)_+\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_{1n} & (x_{1n}-k_1)_+ & \cdots & (x_{1n}-k_k)_+\end{pmatrix}$, $X_{(2)}=\begin{pmatrix}g_1((X_2)_1) & \cdots & g_{N2}((X_2)_1)\\ \vdots & \ddots & \vdots\\ g_1((X_2)_n) & \cdots & g_{N2}((X_2)_n)\end{pmatrix}$,

$X=\begin{pmatrix}X_{(1)} & X_{(2)}\end{pmatrix}$

Then model (a) is:

$E_A(Y\mid X_1,X_2)=s_1(X_1;\underline{\gamma}_1)+g(X_2;\underline{\gamma}_2)=\gamma_0+\gamma_1*x_1+\sum_{k=1}^{K}\gamma_{1k}(x_1-k_k)_+ + X_{(2)}\underline{\gamma}_2$, the fitting criterion is to minimize $\|y-X\gamma\|^2+\lambda^2\underline{\gamma}^T D\underline{\gamma}$, where $\gamma=(\gamma_0,\gamma_1,\underline{\gamma}_2,\gamma_{11},...,\gamma_{1K})^T$. Using mixed model presentation and by the restricted maximal likelihood, the fitted values are $\hat{Y}_1(X_1,X_2;\hat{\lambda})=X(X^TX+\hat{\lambda}^2D)^{-1}X^TY$, $\hat{\lambda}$ is the estimated penalty and $D=\text{diag}(0_{2+N2},1_K)$. When $n\to\infty$, $\hat{\lambda}\to 0$ and $\hat{Y}_1(X_1,X_2;\hat{\lambda})\to\hat{Y}_1(X_1,X_2;0)=X(X^TX)^{-1}X^TY$, the least squares estimates of model (a).

Similarly, for model (b), $E_A(Y\mid X_1)=s_1(X_1;\beta_1)=\beta_0+\beta_1x_{11}+\sum_{k=1}^{K}\beta_{1k}(x_{11}-k_k)_+$, as $n\to\infty$, $\hat{Y}_2(X_1;\hat{\lambda})\to\hat{Y}_2(X_1;0)=X_{(1)}(X_{(1)}^TX_{(1)})^{-1}X_{(1)}^TY$.

For model (c), $E_A(g_i(X_2)\mid X_1)=s_1(X_1;\underline{\delta}_i)=\delta_{i0}+\delta_{i1}x_{11}+\sum_{k=1}^{K}\delta_{i1k}(x_{11}-k_k)_+$, as $n\to\infty$, $\hat{g}_i(X_2;\hat{\lambda})\to\hat{g}_i(X_2;0)=X_{(1)}(X_{(1)}^TX_{(1)})^{-1}X_{(1)}^Tg_i(X_2)$ and $\hat{X}_{(2)}(\hat{\lambda})\to\hat{X}_{(2)}(0)$

By lemma 1, $\hat{Y}_2^*(X_1;0) = \hat{Y}_2(X_1;0)$, then,
$\hat{Y}_2^*(X_1;\hat{\lambda}) = s_1(X_1;\underline{\hat{\gamma}}_1,\hat{\lambda}) + \hat{X}_{(2)}(\hat{\lambda})\underline{\hat{\gamma}}_2 \to s_1(X_1;\underline{\hat{\gamma}}_1,0) + \hat{X}_{(2)}(0)\underline{\hat{\gamma}}_2 = \hat{Y}_2(X_1;0)$ as $n \to \infty$.

From model (b) $\hat{Y}_2(X_1;\hat{\lambda}) \to \hat{Y}_2(X_1;0) = X_{(1)}(X_{(1)}^T X_{(1)})^{-1} X_{(1)}^T Y$ as $n \to \infty$.

So $\hat{Y}_2^*(X_1) \to \hat{Y}_2(X_1)$ as $n \to \infty$ and the proof is complete.

Based on the corollary, the simplified PSPP method yields consistent marginal mean of the missing variable even when the *g* function is not correctly specified.

We prove the case when the *g* function is linear. We can approximate a nonlinear *g* function using a linear form and the corollary can be applied directly.

(2) Proof of consistency of the stratified PSPP method under correct specification of the propensity score

Model:

$$Y \sim N(I_1 s_1(P^*) + ... + I_C s_C(P^*) + g(P^*, X_2, ..., X_p), \sigma^2)$$

For each level of $X_1 = c$, $\hat{Y}_i\_stratified = \hat{s}_c(P_i^*;\alpha) + \hat{g}(P_i^*, X_{i,2}, ..., X_{i,p-1};\beta)$ (1) is the predicted value for the *i*th subject.

The mean function has the form of a spline on the propensity score with subgroup *c*, plus a parametric function of the other covariates.

Let $g_i(\underline{X})$ be the ith component in the *g* function in (1) with $E_A(g_i(\underline{X}) \mid P^*) = s_{gi}(P^*;\underline{\delta}_i)$, $i = 1, ..., N2$, where $E_A(g_i(\underline{X}) \mid P^*)$ is the conditional mean of $g_i(\underline{X})$ given P^*; $s_{gi}(P^*;\underline{\delta}_i)$ is a spline of P^* indexed by the parameter $\underline{\delta}_i$. Let $\underline{\hat{\delta}}_i$ be the restricted maximum likelihood estimates of $\underline{\delta}_i$, the predicted values of $g_i(\underline{X})$ is $\hat{g}_i(\underline{X}) = s(P_i^*;\underline{\hat{\delta}}_i)$. Let

$\hat{X}_{(2)} = [\hat{g}_1(\underline{X}),...,\hat{g}_{N2}(\underline{X})]$ and $\hat{\underline{\delta}} = [\hat{\underline{\delta}}_1,...,\hat{\underline{\delta}}_{N2}]$. Substitute $\hat{X}_{(2)}$ into (1) and obtain

$\hat{Y}_i^*(X_1 = c) = \hat{s}_c(Y^*;\hat{\alpha}) + \hat{X}_{(2)} * \hat{\underline{\delta}}$. (2)

Since within each subgroup *c*, $E_c(Y) = s(P^*;\gamma)$ or $\hat{Y}_i(X_1 = c) = \hat{s}(P_i^*;\hat{\gamma})$ (3)

which is consistent for the conditional mean of Y in subgroup *c* by the balancing property of propensity score.

By corollary, $\hat{Y}_i^*(X_1 = c) \to \hat{Y}_i\ (X_1 = c)$ as $n \to \infty$.

When the propensity is incorrectly specified while the mean function is correctly specified, the predicted conditional means are consistent. Thus the stratified PSPP method has the DR property as described in Section 2.2.

CHAPTER III
A COMPARATIVE STUDY OF THE PENALIZED SPLINE PROPENSITY PREDICTION METHOD WITH ALTERNATIVE DOUBLY ROBUST ESTIMATORS

Abstract

The goal of this paper is to compare several doubly robust (DR) estimators of the mean when missing data exist. An estimator is doubly robust if either the regression of the missing variable on the observed variables or the missing data mechanism is correctly specified. One method is to include the inverse of the propensity score as a linear term in the imputation model (Firth and Bennett, 1998; Scharfstein, Rotnitzky and Robins, 1999; Bang and Robins, 2005). Another method is to calibrate the predictions from a parametric model by adding mean of the weighted residuals (Robins, Rotnitzky and Zhao, 1994; Scharfstein, Rotnitzky and Robins, 1999). The Penalized Spline Propensity Prediction (PSPP) model includes the propensity score into the model nonparametrically (Little and An, 2004; Zhang and Little, 2005). All these methods have consistency properties under misspecification of regression models, but their efficiency and confidence coverage has received little attention. In this paper we compare root mean square error (RMSE), width of confidence interval and non-coverage rate of these methods under different mean functions and propensity score functions. We also study the effects of sample size and misspecification of the propensity scores. The PSPP method yields estimates with smaller RMSE and width of confidence interval compared with other methods under most situations. It also yields estimates with non-coverage rates close to the 5% nominal level.

3.1 Introduction

Missing data problems are very common for statistical research. In this paper we focus on the univariate missing data, where missingness confines to a single variable. Let $(X_1,...,X_p,Y)$ denote a $(p+1)$ -dimensional vector of variables with Y subject to missingvalues and $X_1,...,X_p$ fully observed covariates. We consider the problem of estimating the mean of Y, $E(Y)$.

The sample mean of Y based on the complete cases, $\overline{Y}$, is an unbiased estimate of $E(Y)$ if is the missing data mechanism is missing complete at random (MCAR), which means the missingness of Y does not depends on the covariates $X_1,...,X_p$ or Y. MCAR is a strong assumption and is usually not realistic. In practice it is very common to assume the missingness of Y depends only on the observed covariates $X_1,...,X_p$ but not on Y, which is called missing at random (MAR) (Rubin, 1976; Little and Rubin, 2002).

When the missing data are MAR, many methods can be applied to derive the marginal mean of Y. One of the standard methods is to fit a parametric model. For example, we can derive the marginal mean of Y based on a linear regression model $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i$, where ε_i is the error term, with $\varepsilon_i \sim N(0,\sigma^2)$. We can solve this model by maximum likelihood (ML) approach (Little and Rubin, 2002; Anderson, 1957; Rubin, 1974). The marginal mean of Y can be derived as $\hat{Y} = n^{-1}(\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \hat{y}_i)$, with $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ij}$, where $\hat{\beta}_0,...,\hat{\beta}_p$ are the maximum likelihood estimators based on the complete cases. An alternative to the ML estimators is to add prior distributions for the parameters $\beta_0,...,\beta_p$ and σ^2 and derive the posterior distributions of Y given the covariates and the unknown parameters $\beta_0,...,\beta_p$ and σ^2 , we denote it as $p(Y \mid X_1,...,X_p,\beta_0,...,\beta_p,\sigma^2)$ (Gelman, Carlin, Stern and Rubin, 1995). Missing values

of *Y* and the unknown parameters $\beta_0,...,\beta_p$ and σ^2 are drawn iteratively by Gibbs' sampler or by Markov Chain Monte Carlo (MCMC) method (Casella and George, 1992; Geman and Geman, 1984). When the posterior distribution reaches stationary condition after *N*th iteration, *M* sets of data are created such that within each data every missing Y_i is substituted by an independent draw from the posterior distribution. For each dataset a posterior mean of *Y*, $\bar{Y}^{(l)}$, $l = 1,....,M$, is derived as the average of the observed values and the posterior draws. The marginal mean of *Y* is the average the posterior means over the *M* datasets. Usually *M* needs to be a large number. However, if we can assume approximate normality for the posterior distribution of $\beta_0,...,\beta_p$ and σ^2 given the observed data, $p(\beta_0,...,\beta_p,\sigma^2 \mid X_1,...,X_p)$, we only need to create a small number of datasets to estimate the marginal mean of *Y*, which is the idea of multiple imputation (Little and Rubin, 2002; Rubin, 1978). For each dataset the missing values are replaced by independent posterior draws and the complete-data analysis technique is applied to each imputed dataset. The marginal mean of *Y* can be derived using Rubin's combination rules (Rubin 1978, 1987, 1996; Rubin and Schenker, 1986; Barnard and Rubin, 1999). Let $\hat{\mu}_d$ be the estimated marginal mean of the *d*th dataset, $d = 1,...,D$, where D is the total number of imputed datasets. The marginal mean of *Y* is derived as $\hat{\mu} = \sum_{d=1}^{D} \hat{\mu}_d / D$.

The parametric approach described above is very efficient and yields consistent estimates if the model assumptions are correct. But the drawback is that it is very sensitive to model misspecification, particularly when data are not MCAR. In recent years, researchers have developed robust imputation methods, given concerns with effects of model misspecification and a growth of interest in nonparametric and semiparametric methods (Robins, Rotnitzky and Zhao, 1994; Rotnitzky, Robins and Scharfstein, 1998; Little and An, 2004; Bang and Robins, 2005). An estimator is doubly robust (DR) if either the joint distribution of the complete data is correctly specified or a model for the missing data mechanism is correctly specified. The Penalized Spline Propensity Prediction (PSPP) model is an imputation model with a double robustness property. This method includes the propensity score into the imputation model

nonparametrically and yields robust estimators for the marginal mean. The DR property can also be achieved by including the propensity score into the model parametrically, as in the linear in the weight prediction (Firth and Bennett, 1998; Scharfstein, Rotnitzky and Robins, 1999; Bang and Robins, 2005) and the calibration method (Robins, Rotnitzky and Zhao, 1994; Scharfstein, Rotnitzky and Robins, 1999). We describe these three methods in section 3.2. All these methods show consistency property based on asymptotic theory, but their efficiency and confidence interval coverage in small and moderate sized samples has received little attention. In section 3.3 we compare root mean square error (RMSE), width of confidence interval and non-coverage rate of these estimators under different mean functions and propensity score functions and varying degree of model misspecification. We also include weighted complete-case analysis for comparison due to its close relationship with the calibration estimator. We apply these methods to an online weight loss study in section 3.4. Section 3.5 contains concluding remarks.

3.2 Doubly robust estimators

In this chapter we describe three doubly robust estimators, namely the PSPP method, the linear in the weight method and the calibration method. We also describe the weighted mean of complete cases as an important special case of the calibration estimator.

3.2.1 Penalized Spline of Propensity Prediction (PSPP)

Let $(Y, X_1, ..., X_p)$ denote a vector of variables with Y subject to missing values and $X_1, ..., X_p$ fully-observed covariates. We assume that the missingness of Y depends only on $X_1, ..., X_p$, so the missing data mechanism is missing at random (MAR). Let M be an indicator variable with $M = 1$ when Y is missing and $M = 0$ when Y is observed. Define the logit of the propensity score for Y to be observed as:

$$P^* = \text{logit}\left(\Pr(M = 0 \mid X_1, ..., X_p)\right). \quad (8)$$

The PSPP method is based on the balancing property of the propensity score, which means, conditioning on the propensity score and assuming MAR, missingness of *Y* does not depend on $X_1,...,X_p$ (Rosenbaum and Rubin, 1983). The mean of *Y* can be written as

$$\mu_y = E[(1-M)Y] + E[M \times E(Y \mid P^*)]. \quad (9)$$

Thus the missing data can be imputed based on the observed values conditioning on the propensity score. This leads to the Penalized Spline of Propensity Prediction Method (PSPP) (Little and An, 2004; Zhang and Little, 2005), described in the following model:

$$(Y \mid P^*, X_1,..., X_p; \beta) \sim N(s(P^*) + g(P^*, X_2,...X_p; \beta), \sigma^2), \quad (10)$$

where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and constant variance σ^2. There are two components in the mean function. The first part, $s(P^*)$, consists of the propensity score P^*. Since the true relationship of *Y* and P^* is usually unknown, the PSPP method includes propensity score in the mean function nonparametrically. The second part of mean function is a parametric function $g(P^*, X_2,...X_p; \beta)$, which includes covariates that predict the mean of the *Y* and the propensity score. One of the predictors, here X_1, is omitted from the *g* - function to avoid multicollinearity.

We can implement this model by a number of spline models with different choices of bases (Eilers and Marx, 1996; Ruppert, Wand and Carroll, 2003; Ngo and Wand, 2004; Eubank, 1998; Wahba, 1990). In this paper we choose the penalized spline with truncated linear basis with the form:

$$s(P^*) = \beta_0 + \beta_1 P^* + \sum_{k=1}^{K} \gamma_k (P^* - \kappa_k)_+, \quad (11)$$

where 1, $P^*, (P^* - \kappa_1)_+, ..., (P^* - \kappa_k)_+$ is the truncated linear basis; $\kappa_1 < .... < \kappa_K$ are selected fixed knots and *K* is the total number of knots. This model can be fitted using a number of existing software packages, such as PROC MIXED in SAS (SAS, 1992; Ngo and Wand, 2004, Littell, Milliken, Stroup, and Wolnger. 1996; Ruppert, 2002) and lme() in S-plus (Pinheiro and Bates, 2000). The first step of fitting a PSPP model estimates the propensity score, for example by a logistic regression model or probit model of *M* on $X_1,..., X_p$; in the second step, the regression of *Y* on P^* is fit as a spline model with the

other covariates included in the model parametrically in the g - function. When Y is a continuous variable we choose a normal distribution with constant variance. For other types of data, extensions of the PSPP are straightforward by using the generalized linear models with different link functions.

The predicted mean of Y has a doubly robust property meaning that the predicted mean of Y is consistent if either (a) the mean of Y given $(P^*, X_1, ..., X_p)$ in model (3) is correctly specified, or (b1) the propensity P^* is correctly specified, and (b2) $E(Y \mid P^*) = s(P^*)$. The robustness feature derives from the fact that the regression function g does not have to be correctly specified (An and Little, 2004; Zhang and Little, 2005).

3.2.2 Linear in the weight prediction (LWP)

The Linear in the weight prediction method includes the weight as a linear term in the imputation model (Scharfstein, Rotnitzky and Robins, 1999; Bang and Robins, 2005). For continuous Y it can be written as,

$$(Y \mid X_1, ..., X_p; \beta) \sim N(g(X_1, ...X_p; \beta) + \alpha * \hat{W}, \sigma^2)$$

The first component of the mean function, $g(X_1, ...X_p; \beta)$, is a parametric function with covariates that predict the mean of *Y*. It is the same as a linear regression model. The second part of the mean function includes the estimated weight, $\hat{W}$, as a linear term in the model, where $\hat{W} = 1/\Pr(R = 1 \mid X_1, ..., X_P)$ is the inverse of the estimated propensity score of respondents. Similar approach has been applied in the sample survey setting, where the weights are due to sampling rather than nonresponse (Sarndal, Swensson and Wretman, 2003 ; Firth D. and Bennett, 1998). The LWP has a similar double robustness property as the PSPP method meaning that if either the mean function of *Y* given the covariates are correctly specified or the weight is correctly estimated, then the marginal mean of missing variable *Y* will be consistent. Like the PSPP method, the first step of fitting a Linear in the weight model estimates the propensity score, for example by a

logistic regression model or probit model of M on $X_1,...,X_p$; in the second step, the regression of Y on the weight and the other covariates is fit parametrically.

3.2.3 Calibration method (CAL)

The calibration method calibrates the predictions from a parametric model by adding the mean of the weighted residuals, with weights equal to the inverse of the propensity scores (Robins, Rotnitzky and Zhao, 1994; Scharfstein, Rotnitzky and Robins, 1999). The calibration method consists of three steps. Firstly a parametric model is fit to the complete cases and predictions are derived for all the subjects based on the regression model. Secondly, the propensity score is estimated by a logistic regression model or a probit model of M on $X_1,...,X_p$. Then the marginal mean of Y can be estimated by combining mean of the predictions with mean of the weighted residuals, where the residual of a complete case is the difference of the observed and predicted values for the complete case. The estimator of the mean is

$$\hat{\mu} = n^{-1}(\sum_{i=1}^{n} \hat{y}_i) + n^{-1}(\sum_{i=1}^{r} \hat{w}_i (y_i - \hat{y}_i)),$$

where $\hat{w}_i = 1/\Pr(R_i = 1 \mid X_1,...,X_P)$ is the estimated weight for the *i*th subject, and $\hat{y}_i$ is the regression prediction from a parametric model for the *i*th subject . This method has a double robustness property meaning that if either the prediction model is correctly specified or the weight is correctly estimated, then the marginal mean of Y is consistent.

3.2.4 Weighted complete-case analysis (WCC)

If we set the predictions equal to zero we obtain the weighted complete-case estimator from the calibration method as follows:

$$\hat{\mu} = (\sum_{i=1}^{r} \hat{w}_i y_i) / (\sum_{i=1}^{r} \hat{w}_i),$$

where $\hat{w}_i = 1/\Pr(R_i = 1 \mid X_1,...,X_P)$ is the estimated weight for the *i*th subject. For example if a subject in complete cases has a selection probability of 0.1, it will have a

weight of 10, which means this subject will represent 10 subjects when estimate the mean of missing variable. This estimator is commonly used to handle unit non-response in surveys (Little, 1983, 1986, 1991; Little and Rubin, 2002; Horvitz, and Thompson, 1952; Cochran, 1968).

3.3 Simulation studies

In this section we conduct simulation studies to compare root mean square error, average width of confidence interval and non-coverage rate of the estimators described above. In 3.3.1 we assume we have a correctly specified propensity model and we study the performance of the estimators when the mean function of the missing variable given the covariates is not correctly specified. For the propensity function we have the missingness of Y depend on the fully observe covariates, but with different degree of dependency. We also study how the properties of the estimators depend on the sample size.

In 3.3.2, we study the performance of the various estimators when we have wrongly specified propensity scores. When the propensity score is wrong but the mean function is correct, we will have consistent estimates of the marginal mean. But when both are wrong, none of the estimators yields consistent marginal mean. We conduct simulation study to compare these estimators with correctly or wrongly specified mean functions.

3.3.1 Performance of the DR estimators when the propensity score is correct specified

We conduct three simulation studies in 3.3.1 with different mean and propensity functions. Simulation 1 and 2 concern a simple mean function with a single covariate and simulation 3 concerns more complex mean functions with 2 covariates. We vary the degree of dependency of the missingness of Y on the covariates, the degree to which the regression is misspecified, and the sample size.

Simulation 1. A misspecified quadratic mean function. We simulate 500 datasets with sample size of 50, 100, 200, 400, 800 and 1,500 respectively, with a fully-observed covariate X_1 from standard normal distribution and a continuous response variable *Y*. Let *M* be an indicator variable with $M = 1$ when *Y* is missing and $M = 0$ when *Y* is observed. We create missing values of *Y* from the following response propensity model:

$$\text{logit}\,(P(M = 0 \mid X_1)) = \delta_1 * X_1,$$

with two choices of δ_1, δ_1 = 0.1 and δ_1 = 0.5. The larger δ_1 models a stronger dependency of the response propensity and X_1. For both values of δ_1, the overall probability of missing values has an expected value of 0.5. The distribution of *Y* given X_1 is:

$$Y \mid X_1 \sim N\left(\mu(X_1), 1\right),$$
$$\mu(X_1) = 1 + X_1 + \delta_2 * X_1^2,$$

with two choices for the coefficient of X_1^2, namely 0.8 and 4. For the parametric predictions of the CAL method we assume the regression is linear in X_1, that is δ_2 = 0. Hence larger values of δ_2 imply more serious misspecification of the prediction model in the CAL method. We thus have four different combinations of simulated mean and propensity functions based on the values of δ_1 and δ_2, low low (LL) and low high (LH), high low (HL), high high (HH) (See Table 3.1). For example for the high low cell, the high corresponds to the larger coefficient of the quadratic term in the mean function, which leads to greater misspecification of a imputation model without X_1^2; the low corresponds to the smaller coefficient of δ_1 in the propensity model. This simulation is expected to be favorable to the PSPP, since the spline on the propensity score closely approximates the true regression function on the covariates.

Table 3.1 Simulation 1 classified by degree of misspecification in the mean function and the degree of diversity of the propensity function

	Propensity function	
Mean Function	logit $(P(M=0\mid X_1))$ $=0.1X_1$	logit $(P(M=0\mid X_1))$ $=0.5X_1$
$Y\mid X_1 \sim N\left(1+X_1+0.8X_1^2,1\right),$	Low low (LL)	Low high (LH)
$Y\mid X_1 \sim N\left(1+X_1+4X_1^2,1\right),$	High low (HL)	High high (HH)

We estimate the propensity score by a correctly specified logistic regression including the intercept and a linear term in X_1, namely ,

$$\hat{p}(M=0\mid X_1)=e^{(\hat{\delta}_0+\hat{\delta}_1 X_1)}/(1+e^{(\hat{\delta}_0+\hat{\delta}_1 X_1)}).$$

We then estimate the marginal mean of missing variable by the following versions of PSPP, LWP, CAL and WCC:

(a) The PSPP Method with null *g* function, which we denote $[s(P^*)]$. We use the logit of the propensity score in the spline function instead of the propensity score directly. The marginal mean of *Y* is estimated as the average of the observed data and imputed data. For the penalized spline method in this paper, we choose 5 equally spaced fixed knots when sample size is 50, 10 equally spaced fixed knots when sample size is 100, 20 equally spaced fixed knots when sample size is 200 or more. A truncated linear basis is applied for the spline model, that is, $s(P^*)=\beta_0+\beta_1 P^*+\sum_{k=1}^{K}\gamma_k(P^*-\kappa_k)_+$. We fit this model using PROC MIXED in SAS with $(P^*-\kappa_1)_+,...,(P^*-\kappa_k)_+$ treated as random effects and the intercept and P^* treated as the fixed effects.

(b) The linear in the weight model, namely $[\alpha_0+\alpha_1 * \hat{w}]$, where $\hat{w}$ is the inverse of the estimated propensity score. Missing values are predicted from the regression of *Y* on the

inverse of the estimated probability to respond, that is $E(Y \mid X_1) = \alpha_0 + \alpha_1 \hat{W}$, where $\hat{W} = 1/\hat{p}(M = 0 \mid X_1)$. The marginal mean of Y is estimated as the average of the observed data and imputed data.

(c) The calibration method, [$\hat{\mu} = n^{-1}(\sum_{i=1}^{n} \hat{y}_i) + n^{-1}(\sum_{i=1}^{r} \hat{w}_i (y_i - \hat{y}_i))$], where $\hat{w}_i$ is the estimated weight of the *i*th subject and $\hat{y}_i$ is the predicted value of the *i*th subject from a prediction model of *Y* regressing on X_1 .

(d) Weighted complete-case analysis, [$\hat{\mu} = (\sum_{i=1}^{r} \hat{w}_i y_i)/(\sum_{i=1}^{r} \hat{w}_i)$], with $\hat{w}_i$ the inverse of the estimated propensity score.

We also include a correctly specified regression model of *Y* regressing on X_1 and X_1^2 and a wrongly specified regression model of *Y* regressing on X_1 for comparison.

To derive confidence intervals we apply the above methods to 200 bootstrap samples for each dataset. The variance (V_{boot}) of the marginal mean of *Y* for each dataset is estimated as

$$\hat{V}_{boot} = \frac{1}{199} \sum_{b=1}^{200} (\hat{\mu}^{(b)} - \hat{\mu}_{boot})$$

where $\hat{\mu}^{(b)}$ is the estimated marginal mean of Y for the *b*th bootstrap sample, *b* = 1,…, 200; $\hat{\mu}_{boot}$ is the average the estimated marginal mean of *Y* over the 200 bootstrap samples. We construct the 95% confidence interval for each data set as $(\hat{\mu}_{boot} - 2 * \sqrt{\hat{V}_{boot}},\ \hat{\mu}_{boot} + 2 * \sqrt{\hat{V}_{boot}})$. The non-coverage rate is the percentage of the 500 samples with the 95% confidence intervals not covering the true value. The average width of CI's, CIW, is the average of $4 * \sqrt{\hat{V}_{boot}}$ over the 500 samples.

We derive the relative root mean square error (RRMSE) compared with the before deletion analysis as

RRMSE=100*(RMSE(estimator)-RMSE(BD))/RMSE(BD).

where RMSE(estimator) is the average of the estimated RMSE over the 500 samples of the different estimators, RMSE(BD) is the average of the estimated RMSE over the 500 before deletion samples. Similarly, relative width of confidence interval (RCI) compared with the before deletion estimator,

RCI=100*(CIW(estimator)- CIW(BD))/ CIW(BD),

where CIW(estimator) is the average of the estimated CIW of different estimators, CIW(BD) is the CIW of the before deletion analysis, which is the average width of CI over the 500 before deletion datasets.

Values of RRMSE for the above methods are displayed in Figure 1. Among the four methods we are comparing (PSPP, LWP, CAL, WCC), the PSPP method yields smallest RRMSE in all the cases, which is very close to the correctly specified regression model. When the propensity of response is not strongly related to X_1 (LL and HL), the linear in the weight method, calibration method and weighted complete-case analysis yield similar RRMSEs. The gain of the PSPP method over the other methods is sizeable in the LL case, but the difference is even more dramatic in the HL case, where the quadratic term in the mean function is more important. When the missingness of Y strongly depends X_1 (LH and HH) and the complete cases are no longer approximately a random sample of the original data, the RRMSEs of the different methods are more differentiated. The linear in the weight method yields largest RRMSE, follows by the calibration method and weighted complete-case analysis. Again, the gain in RMSE of the PSPP method over the other methods is sustained when the coefficient of the quadratic term is small (LH case), and even greater in the HH case, where the quadratic term in the mean function is more important. The wrong regression model yields much larger RRMSE than the PSPP , LWP, CAL and WCC in the LH and HH cases.

The average widths of confidence interval follow similar pattern as the RRMSEs (Figure 2). For the LL and HL cases, among the four methods we are comparing, the

PSPP method yields the narrowest CI's on average, and the linear in the weight method, calibration method and weighted complete-case analysis yield CI's with similar width. The gain of the PSPP method over the other methods is smaller in the LL case than in the HL case. For the LH and HH cases, the linear in the weight method yield largest width of CI, followed by the calibration method and weighted complete-case analysis. Again the gain of the PSPP method over the other method is smaller in the low misspecication case (LH) than in the high misspecification case (HH). The correctly specified regression model yields smallest width of CI in all cases. The wrong regression model yields width of CI greater than the PSPP, but smaller than the LWP, CAL and WCC in the LH and HH cases.

The linear in the weight method yields width of CI's large with of CI when the sample size is small (sample size =50). One possible reason is that it is very sensitive to extreme propensity scores when the sample size is small. A small propensity score corresponds to a large weight, which leads to extreme predictions under the linear model. The PSPP method, on the other hand, estimates a spline curve through the propensity scores and the curvature prevents these extreme predictions for cases with small propensity scores.

Figure 3 displays non-coverage rates for the four methods. In general, all methods yield non-coverage rates close to the 5% nominal level when the sample size reaches 200 or more. When the sample size is less than 200, the non-coverage rate of the linear in the weight method is smaller than the 5% nominal level, and the other methods yields non-coverage rate greater than the 5% nominal level. The PSPP method has better coverage than the weighted complete-case analysis and calibration prediction in the LH and HH cases; for the LL and HL case, the coverage rates of these three methods are similar. The wrong regression model yields large non-coverage rate in the LH and HH cases due the bias of the estimates.

Simulation 2. Mean function depends linearly on weight. Simulation 1 concerns a situation where the LWP provides a poor fit to the data. Simulation 2 is designed to be

more favorable to the LWP method. The PSPP is also expected to do well in this case since it can approximate the true regression function on the covariates. We simulate 500 datasets with the same sample sizes as simulation 1, with one complete covariates X_1 from standard normal distribution and a continuous response variable *Y*. As for simulation 1, we create missing values of *Y* from the response propensity model:

$$\text{logit}\,(P(M=0\mid X_1))=\delta_1 * X_1,$$

with two values of δ_1, δ_1 = 0.1 and δ_1 = 0.5. We derive the weight (W) as the inverse of the propensity score and the mean structure of *Y* depends on the weight as follows:

$$Y\mid W \sim N\left(\mu(W),1\right),$$
$$\mu(W)=1+W+\delta_2 * W^2,$$

where δ_2 is the coefficient for the quadratic term, chosen to equal to 0.8 or 4. The larger δ_2 means the greater misspecification of an imputation model without W^2. We thus have four different simulation conditions depending on the values of δ_1 and δ_2 (Table 3.2).

Table 3.2 Simulation 2 classified by degree of misspecification in the mean function and the degree of diversity of the propensity score

	Propensity function	
Mean Function	$\text{logit}\,(P(M=0\mid X_1))=0.1X_1$	$\text{logit}\,(P(M=0\mid X_1))=0.5X_1$
$Y\mid W \sim N\left(1+W+0.8W^2,1\right)$	Low low (LL)	Low high (LH)
$Y\mid W \sim N\left(1+W+4W^2,1\right)$	High low (HL)	High high (HH)

Like simulation 1, we estimate the propensity score by a correctly specified logistic regression including the intercept and a linear term in X_1. We then estimate the marginal mean of *Y* using the following versions of PSPP, LWP, CAL and WCC. .

(a) The PSPP method with null the *g* function, which we denote $[s(P^*)]$. We use the same procedure as simulation 1. The marginal mean of *Y* is estimated as the average of the observed data and imputed data.

(b) The linear in the weight prediction model, namely $[\alpha_0 + \alpha_1 * \hat{w}]$, where $\hat{w}$ is the inverse of the propensity score. Like simulation 1, we fit the linear regression model of *Y* on $\hat{w}$ to the complete cases and the missing values are predicted by the regression model. The marginal mean of *Y* is the average of the observed data and the predicted data.

(c) The calibration estimator, $[\hat{\mu} = n^{-1}(\sum_{i=1}^{n} \hat{y}_i) + n^{-1}(\sum_{i=1}^{r} \hat{w}_i (y_i - \hat{y}_i))]$, where $\hat{w}_i$ is the weight for the *i*th subject and $\hat{y}_i$ is the prediction from a linear regression model of *Y* on X_1.

(d) The weighted complete-case analysis, $[\hat{\mu} = (\sum_{i=1}^{r} \hat{w}_i y_i)/(\sum_{i=1}^{r} \hat{w}_i)]$ with $\hat{w}_i$ the inverse of the estimated propensity score.

We also include a correctly specified regression model of *Y* regressing on $\hat{w}$ and $\hat{w}^2$ and a wrongly specified regression model of *Y* regressing on X_1 for comparison.

We derive RRMSE, relative width of confidence interval (RCI) and non-coverage rate as simulation 1 and the results are shown in Figures 4-6. For the LL and HL cases, where the complete cases resemble a random sample of the original data, all of the method yields very similar RRMSE. For the LH and HH cases, the PSPP method and the LWP yield smaller RRMSE than the CAL and the WCC. For the LH case, the LWP model is very close to the correctly specified mean model, and consequently the RRMSE of the LWP model is small; while in the HH case, the quadratic term is more important and the LWP model no longer close to the true mean model and thus yields slightly larger RRMSE than the PSPP method. The CAL has better RRMSE than WCC, but is considerably less efficient than the PSPP for the LH and HH situations. The correct regression model yields large RRMSE when the sample size is small because the mean

functions depend on $\hat{w}$ and $\hat{w}^2$ and extreme small propensity score corresponds to large weight, which leads to large predictions.

All methods yield similar width of confidence interval in the LL and HL situation. For the LH and HH situation, the weighted complete-case analysis yields largest width of confidence interval, followed by the calibration method and the PSPP method yields smallest width of confidence interval. The linear in the weight method, the calibration method and the correct regression model yield very large width of confidence intervals at sample size of 50.

For the non-coverage rate, all estimators except the wrong regression model yield non-coverage rate close to 5% nominal level in the LL, LH and HL situations. For the HH situation, the PSPP method, calibration method and the weighted complete-case analysis yield non-coverage rate above the 5% nominal level at small sample sizes and the wrong regression model yields much larger non-coverage rate compared with other methods.

Simulation 3. Mean function include the interaction of the covariates. This simulation concerns more complex mean function and is designed to be more favorable to the CAL method. We simulate 500 datasets with sample size of 50, 100, 200, 400, 800 and 1,500 respectively, with independent complete covariates X_1 and X_2 from standard normal distribution and a continuous response variable *Y*. We create missing values of *Y* from the response propensity model:

$$\text{logit}\,(P(M=0 \mid X_1, X_2)) = 0.25 * X_1 - \delta_1 * X_2$$

where δ_1 equals to 0.1 or 0.5. The mean structure of *Y* depends on X_1 and X_2 as follows,

$$Y \mid X_1, X_2 \sim N\left(\mu(X_1, X_2), 1\right),$$
$$\mu(X_1, X_2) = X_1 + X_2 + \delta_2 * X_1 * X_2,$$

where δ_2 is the coefficient from the interaction term, which is 0.8 or 4. Like simulations 1 and 2, we thus have four different simulation conditions depending on the values of δ_1 and δ_2 (Table 3.3).

Table 3.3 Simulation 3 classified by degree of misspecification in the mean function and the degree of diversity of the propensity score

	Propensity function	
Mean Function	logit $(P(M=0 \mid X_1, X_2))$ $=0.25*X_1-0.1*X_2$	logit $(P(M=0 \mid X_1, X_2))$ $=0.25*X_1-0.5*X_2$
$Y \mid X_1, X_2$ $\sim N\left(X_1+X_2+0.8*X_1*X_2,1\right)$	Low low (LL)	Low high (LH)
$Y \mid X_1, X_2$ $\sim N\left(X_1+X_2+4*X_1*X_2,1\right)$	High low (HL)	High high (HH)

We estimate the propensity score by a correctly specified logistic regression, which is modeled as an additive function of X_1 and X_2 as follows,

$$\hat{p}(M=0 \mid X_1, X_2)=e^{(\hat{\delta}_0+\hat{\delta}_1 X_1+X_2)}/(1+e^{(\hat{\delta}_0+\hat{\delta}_1 X_1+X_2)}).$$

We then estimate the marginal mean of missing variable by the following methods.

(a) The PSPP Method with null g function, which we denote $[s(P^*)]$. We follow the same procedure as simulation 1. The marginal mean of *Y* is derived as the average of the observed data and imputed data.

(b) Linear in the weight method namely $[\alpha_0+\alpha_1*\hat{w}]$, where $\hat{w}$ is the inverse of the estimated propensity score. The marginal mean of *Y* is the average of the observed data and the imputed data.

(c) The calibration method $[\hat{\mu}=n^{-1}(\sum_{i=1}^{n}\hat{y}_i)+n^{-1}(\sum_{i=1}^{r}\hat{w}_i(y_i-\hat{y}_i))]$, where $\hat{w}_i$ is the weight for the *i*th subject and $\hat{y}_i$ is the prediction for the *i*th subject from a prediction model by regressing *Y* on X_1 and X_2.

(d) Weighted complete-case analysis $[\hat{\mu}=(\sum_{i=1}^{r}\hat{w}_i y_i)/(\sum_{i=1}^{r}\hat{w}_i)]$, is the same as simulation 1.

We also include a correctly specified regression model of Y regressing on X_1, X_2 and $X_1 * X_2$ and a wrongly specified regression model which does not include the interaction term $X_1 * X_2$.

We derive RRMSE, relative width of confidence interval (RCI) and non-coverage rate as simulations 1 and the results are shown in Figures 7-9. The correct regression model yields much smallest RRMSE at all cases. Among the four methods we are comparing, the PSPP method yields smallest RRMSE in the LL, HL and HH situations. The gain of the PSPP method over the other methods is minimal in the LL case, but the difference is more dramatic in the HL and HH situations. For LH case, the CAL model is very close to the correctly specified mean model, and consequently the RRMSE of the CAL model is the smallest. In the HL and HH cases, the quadratic term is more important for the mean model and thus the CAL model is no longer close to the true mean model and yields larger RRMSE than the PSPP method.

The PSPP, LWP, CAL and WCC yield similar widths of confidence interval in the LL, HL and HH situations (Figure 8). For the LH situation, the CAL yields smallest width of CI's, which is consistent with the results of RRMSE. The LWP method yields estimators with large confidence intervals at the sample size of 50. The wrong regression model yields smaller width of CI's than the PSPP, LWP, CAL and WCC in general and the correct regression model yields much smaller widths of confidence interval under all situations.

The PSPP method and the linear in the weight method yield estimators with non-coverage rate below the 5% nominal level at sample size of 50. Other than that, the PSPP method yields non-coverage rates closer to the 5% nominal level than the other methods (Figure 9). The wrong regression model yields large non-coverage rate for LH and HH cases.

3.3.2 Misspecification of propensity score

In this section we study the performance of different estimators when we misspecify the propensity scores. The DR properties of the PSPP, LWP and CAL protect against model misspecification of either the regression prediction model or the propensity model; but when both models are incorrect, these methods do not yield consistent mean estimates. We use the follow simulation study to show the DR property of the above estimators.

Simulation 4. Misspecification of the propensity score. We simulate 500 datasets with sample size of 500 each, with independent complete covariates X_1 and X_2 from standard normal distribution and a continuous response variable *Y*. We create missing values of *Y* from the response propensity model:

$$\text{logit}\,(P(M=0\mid X_1,X_2))=0.75*X_1-0.5*X_2$$

The mean structure of *Y* depends on X_1 and X_2,

$$Y\mid X_1,X_2\sim N\left(\mu(X_1,X_2),1\right),$$
$$\mu(X_1,X_2)=1+X_1+X_2+X_1*X_2,$$

The marginal mean of *Y* is 1 in both settings and the missing percentages are about 50%.

We estimated the propensity score by the following three logistic regressions:

(A) A correctly specified logistic regression model includes X_1 and X_2 additively, namely, $\hat{p}(M=0\mid X_1,X_2)=e^{(\hat{\delta}_0+\hat{\delta}_1X_1+\hat{\delta}_1X_2)}/(1+e^{(\hat{\delta}_0+\hat{\delta}_1X_1+\hat{\delta}_1X_2)})$. We denote the logit of the propensity score as $P^*_{correct}$.

(B) A wrongly specified logistic regression including X_1 only, namely, $\hat{p}(M=0\mid X_1,X_2)=e^{(\hat{\delta}_0+\hat{\delta}_1X_1)}/(1+e^{(\hat{\delta}_0+\hat{\delta}_1X_1)})$. We denote the logit of the propensity score as $P^*_{wrong\_x_1}$

(C) A wrongly specified logistic regression including X_2 only, that is, $\hat{p}(M=0\mid X_1,X_2)=e^{(\hat{\delta}_0+\hat{\delta}_1X_2)}/(1+e^{(\hat{\delta}_0+\hat{\delta}_1X_2)})$. We denote the logit of the propensity score as $P^*_{wrong\_x_2}$.

For each of the propensity score estimated above, we derive the marginal mean of missing variable by the following versions of the PSPP, LWP, CAL and WCC.

(I) The PSPP method

(a) The PSPP method with null *g* function, which we denote $[s(P^*)]$, where P^* is the logit of the propensity score described above. This model does not specify the mean function of *Y* given the covariates correctly since none of three propensity scores $P^*_{correct}$, $P^*_{wrong\_x_1}$ or $P^*_{wrong\_x_2}$ contains the interaction term of X_1 and X_2. The marginal mean of *Y* is derived as the average of the observed data and imputed data.

(b) Model (a) with X_1 included, which we denote $[s(P^*) + X_1]$. We do not fit this model when we estimate propensity score by model B to prevent multicolinearity.

(c) Model (a) with X_2 included, which we denote $[s(P^*) + X_2]$. We do not fit this model when we estimate propensity score by model C to prevent multicolinearity.

(d) Model (a) with X_2 and $X_1 * X_2$ included, which we denote $[s(P^*) + X_2 + X_1 * X_2]$. This model correctly specified the mean function of *Y* given the covariates when we estimate the propensity score by mode A and B. We do not fit this model when we estimate the propensity score by model C to prevent multicolinearity.

(II) LWP method

(e) Linear in the weight method namely $[\alpha_0 + \alpha_1 * \hat{w}]$, where $\hat{w}$ is the inverse of the propensity score estimated by model A, B, C. The marginal mean of *Y* is the average of the observed data and the predicted data.

(f) Model (e) with X_1 included, namely $[\alpha_0 + \alpha_1 * \hat{w} + X_1]$.

(g) Model (e) with X_2 in included, namely $[\alpha_0 + \alpha_1 * \hat{w} + X_2]$.

(h) Model (e) with X_2 and $X_1 * X_2$ included, namely $[\alpha_0 + \alpha_1 * \hat{w} + X_2 + X_1 * X_2]$.

(i) Model (e) with X_1 , X_2 , $X_1 * X_2$ included, namely, $[\alpha_0 + \alpha_1 * \hat{w} + X_1 + X_2 + X_1 * X_2]$. This model correctly specified the mean function of *Y* given the covariates no matter which methods we use to estimate the propensity score.

(III) CAL estimator

(j) The calibration method, denoted as $[\hat{\mu} = \bar{y} + n^{-1}(\sum_{i=1}^{r} \hat{w}_i (y_i - \bar{y}))]$, where $\hat{w}_i$ is the weight for the *i*th subject and derived as the inverse of the estimated propensity score, $\bar{y}$ is the mean of the complete cases.

(k) The calibration method, denoted as $[\hat{\mu} = n^{-1}(\sum_{i=1}^{n} \hat{y}_{i\_x_1}) + n^{-1}(\sum_{i=1}^{r} \hat{w}_i (y_i - \hat{y}_{i\_x_1}))]$, $\hat{y}_{i\_x_1}$ is the prediction of the *i*th subject from a regression model of *Y* on X_1.

(l) The calibration method, denoted as $[\hat{\mu} = n^{-1}(\sum_{i=1}^{n} \hat{y}_{i\_x_2}) + n^{-1}(\sum_{i=1}^{r} \hat{w}_i (y_i - \hat{y}_{i\_x_2}))]$, $\hat{y}_{i\_x_2}$ is the prediction of the *i*th subject from a regression model of *Y* on X_2.

(m) The calibration method, $[\hat{\mu} = n^{-1}(\sum_{i=1}^{n} \hat{y}_{i\_x_2, x_1 * x_2}) + n^{-1}(\sum_{i=1}^{r} \hat{w}_i (y_i - \hat{y}_{i\_x_2, x_1 * x_2}))]$, $\hat{y}_{i\_x_2, x_1 * x_2}$ is the prediction of the *i*th subject from a prediction model of *Y* on X_2 and $X_1 * X_2$.

(n) The calibration method, $[\hat{\mu} = n^{-1}(\sum_{i=1}^{n} \hat{y}_{i\_x_1, x_2, x_1 * x_2}) + n^{-1}(\sum_{i=1}^{r} \hat{w}_i (y_i - \hat{y}_{i\_x_1, x_2, x_1 * x_2}))]$, $\hat{y}_{i\_x_1, x_2, x_1 * x_2}$ is the prediction of the *i*th subject from a prediction model by regressing *Y* on X_1, X_2 and $X_1 * X_2$. This prediction model is correctly specified.

(IV) WCC method

(o) Weighted complete-case analysis $[\hat{\mu} = (\sum_{i=1}^{r} \hat{w}_i y_i) / (\sum_{i=1}^{r} \hat{w}_i)]$, with $\hat{w}_i$ the inverse of the estimated propensity score.

We estimate bias, which is the average of the deviations of the estimates from the true mean over the 500 simulated data sets, and empirical standard error (SE) that is the standard deviation of the estimates over the 500 simulated data sets. We also calculate the root mean square error, the mean of the squared difference of the estimates from the true value over the 500 data sets (Table 3.4).

When the propensity score is correctly specified all methods yields consistent estimates with small bias, empirical standard error and RMSE (Table 3.4, column A). When the propensity score is wrongly specified but the mean function is correctly specified, the PSPP method (model d), the LWP (model i) and the CAL method (model n) yields estimates with small bias, empirical standard error and RMSE (Table 3.4, column B&C). When neither the propensity score nor the mean function is correctly specified, the PSPP method (model a, b, c), the LWP method (model e, f, g) and the CAL method (model j, k, l, m) yield biased estimates. For model (h) in column B, where the propensity score is wrongly specified but the mean function is close to the correctly specified form, we find the linear in the weight method yields estimates with small bias, empirical standard error and RMSE; but it will yields biased result if the mean function is not close to the correct mean function (Table 3.4, column C). The weighted complete-case analysis (model o) yields biased results (as expected) when the propensity score is not correctly specified.

3.4 An Example: Online Weight Loss Study

We apply the different estimators in this paper to a data from an online weight loss study conducted by Kaiser Permanente (Couper et al., 2005). The study randomized approximately 4,000 subjects to the treatment or the control group. Subjects in the treatment group received tailored weight loss information online. The tailoring information was based on their answers to an initial survey, which contained baseline measurements such as baseline weight, motivation to weight loss, etc; for the control group, information provided online was the same for all the subjects. A follow-up survey was sent to the subjects at month 3, which collected follow-up measurements such as current weight. Our goal is to compare the short-term treatment effects; in particular, we

compare the reduction of the body mass index (BMI), defined as difference of 3-month BMI and baseline BMI.

There were 2059 subjects in the treatment group and 1956 subjects in the control group at the baseline. At 3 month 623 subjects in the treatment group and 611 subjects in the control group responded to the second survey. We assume the data are missing at random. Subjects in the treatment group who remained in the study have much lower baseline BMI than those who dropped out (33.75 vs 35.57; $P<0.001$), but this differences is not seen in the control group (35.22 vs 35.50; $P=0.47$); On the other hand, for the control group subjects who remained in the study have better baseline health, as measured by the number of previous diseases, than those who dropped out of the study ($P<0.01$); this differences was not seen in the treatment group ($P=0.56$). These differences suggest that interactions between treatment and baseline covariates need to be included when estimating the propensity scores.

We estimate the propensity scores by a logistic regression, with the inclusive criterion of retaining all variables with P-values less than 0.20. The final model includes the following covariates: baseline BMI; number of previous disease; baseline self care; which is harder–eating less or being active; baseline exercise support; baseline activity level; baseline eating topology; education; ethnic identity; treatment; interaction of treatment and baseline BMI; interaction of treatment and baseline eating topology; interaction of treatment and baseline activity level; interaction of treatment and number of previous disease; interaction of treatment and which is harder–eating less or being active.

To derive the BMI reduction within each group, we apply the stratified PSPP method, which is a straight extension of the PSPP method by fitting different spline curves to the different subgroups (Zhang and Little, 2005). For the linear in the weight prediction method, calibration method and weighted complete-case analysis we apply the method to the treatment or control group separately.

The baseline covariates in the stratified PSPP method, linear in the weight prediction method and the baseline covariates for the prediction model of the calibration estimator include: ethnic identity; baseline medical advice; baseline eating topology; baseline cardio exercise; baseline activity level; baseline BMI; number of previous disease; number of weigh loss methods tried; motivation of weigh loss; which is harder– eating less or being active.

We include the result of the complete-case analysis for comparison. Results are summarized in Table 3.5. Empirical Standard errors (SE) and the corresponding confidence intervals are obtained from 200 bootstrap samples. The treatment group has a larger reduction of BMI after 3 month (-0.91 (0.09)) compared to the control group (-0.45 (0.10)) based on the complete case analysis. The treatment effect is stronger based on the DR estimators than that of the complete-case analysis. The stratified PSPP method, the linear the weight prediction method, the calibration estimator and the weighted complete-case analysis yield similar results, with the reduction of BMI ranging from -1.01 to -1.04 for the treatment group and -0.40 to -0.42 in the control group. The 95% confidence intervals for the treatment group do not overlap with the control group suggesting a treatment effect on the weight loss.

It is not surprising that the DR estimators yield similar results for this online weight study. We estimate the propensity scores by conditioning on a large number of baseline covariates, which characterize the respondents and the nonrepondents well and as a results we have a well-defined propensity score. Second, we include a large number of baseline covariates in the parametric part of the DR estimators and the same parameterization should not lead to results with large discrepancies. More over in this study we do not find much different from result of complete-case analysis and those of the DR estimators, which suggest that the data is more like the LL case in the simulation studies in this paper and as a results the different methods do not differentiate significantly.

3.5 Conclusion

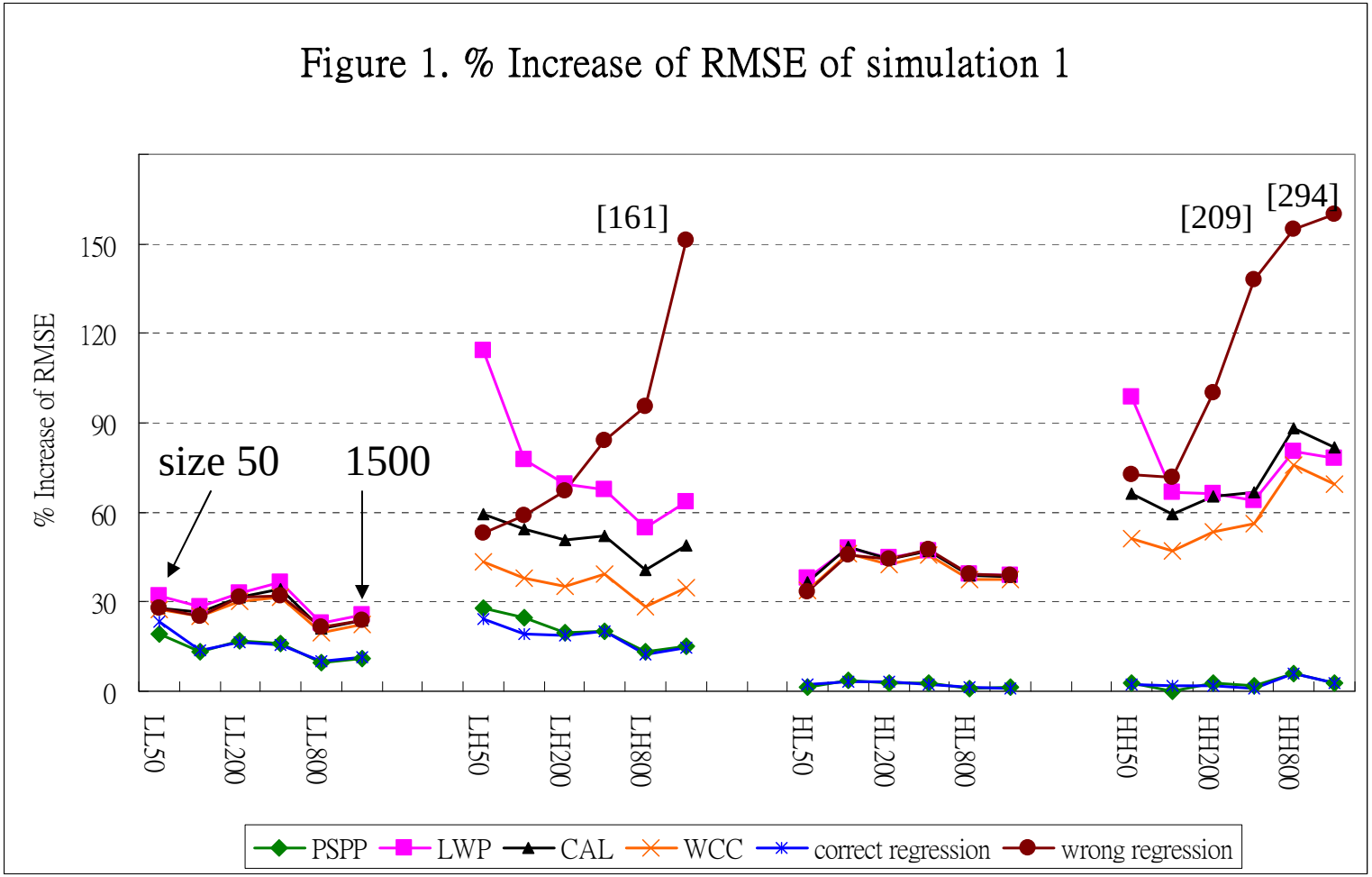
We have compared properties of four methods for estimating the mean from incomplete data that have double robustness properties, in that they yield consistent estimates if either the prediction model or the model for the propensity to respond is correctly specified. For the problems simulated, we find a clear advantage for the PSPP method over the calibration, linear in the weight and response weighting methods when the propensity model is correctly specified. In particular, the PSPP method yields estimates with lower mean squared error and narrower confidence intervals based on bootstrap standard error and coverage that were similar or closer to the nominal level than coverage based on the other methods.

Confidence intervals for the PSPP method did not achieve their nominal coverage when the sample sizes were small, particularly when n = 50. However, the alternative methods, including calibration, were generally no better than PSPP in this situation. When the propensity model was incorrectly specified there was less difference between the methods, which (as expected) tended to do well when the prediction model is well specified and poorly when it is not.

The results from any simulation need to be interpreted with caution, since they are limited to the conditions simulated. We attempted to design our simulations in a way that varied the key elements of the problem at hand -- the extent of misspecification of prediction and propensity models, and the sample size. We fixed the fraction of missing data, which is clearly an important element, at 50%. However, in our experience the effect of this factor is predictable, with the performance of all the methods converging to the complete-data inference as the fraction of missing cases goes down. We chose a relatively high fraction of missing data to accentuate differences between the methods. We simulated normal models with constant variance, and hence did not assess the effect of alternative variance structures and error distributions. There are many ways in which models can be misspecified, and no single simulation can cover all the possibilities.

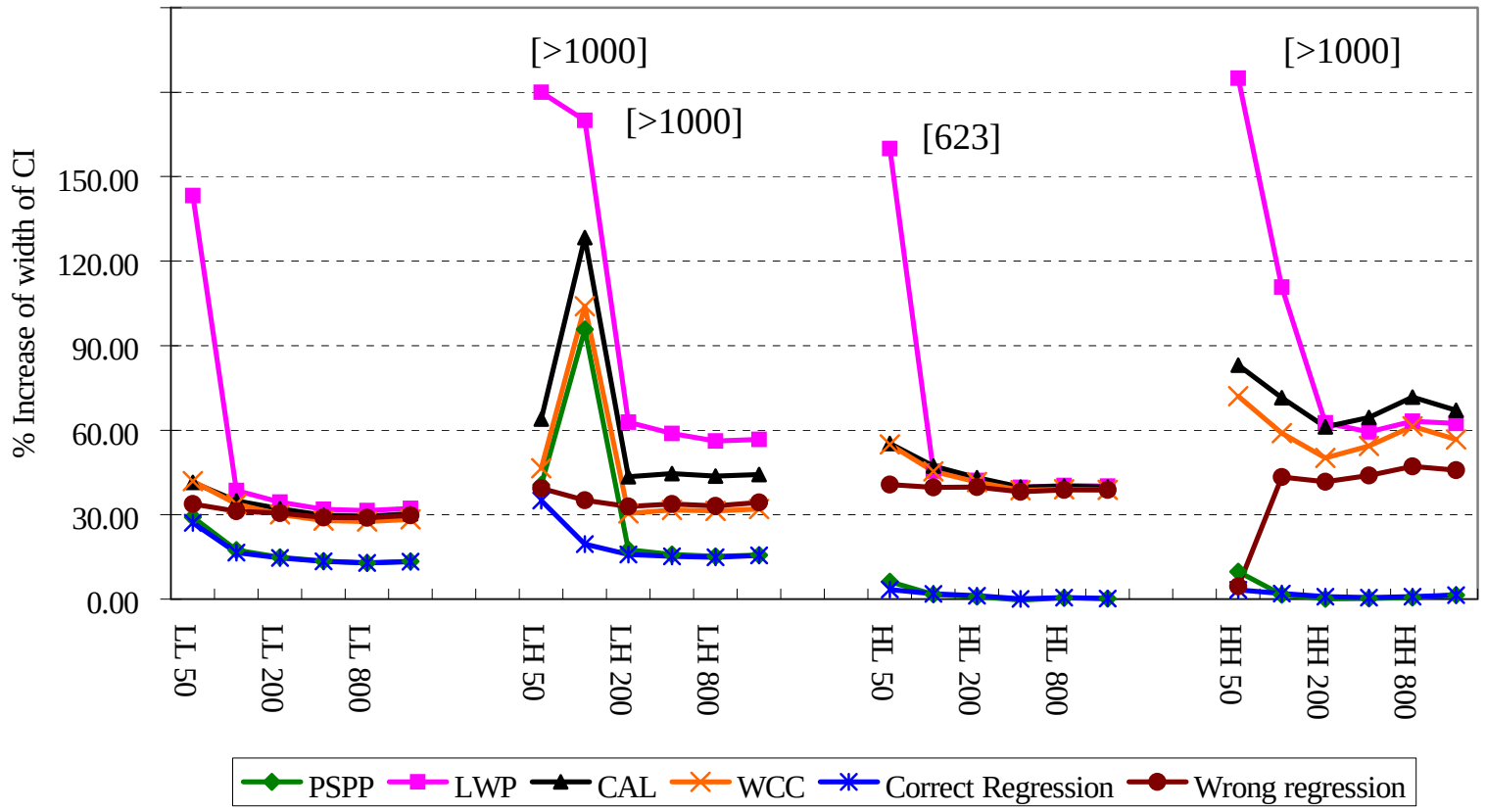
It should also be noted that we restricted attention to inference about the unconditional mean. The calibration approach can also be applied to achieve double robustness for inferences about regression coefficients (Robins, Rotnitzky and Zhao, 1994; Rotnitzky, Robins and Scharfstein, 1998; Robins and Rotnitzky, 2001; Lunceford and Davidian, 2004; Yu and Nan , 2006), a problem where the PSPP method seems less useful. Specifically, Zhang and Little (2005) consider extensions of PSPP to handle conditional means and simple regression coefficients, but extensions to multiple linear regression seem less readily available and appealing. So this is an area where calibration methods appear to have the edge.

With these caveats, our conclusion based on these simulations is that if robustness to restricted parametric models is desired for inferences about means, then it is best achieved by a method like PSPP that specifies a flexible mean structure in directions of the model space that are vulnerable to model misspecification, here the propensity to respond given the covariates. Other methods like calibration do provide robustness through the double robustness property, but they were inferior to the robust PSPP modeling approach in terms of efficiency and confidence coverage, and they did not correct the under-coverage of bootstrap confidence intervals for PSPP in small sample sizes. The latter may be better addressed by "biting the bullet" and adopting more parsimonious parametric models. An alternative method of computing PSPP confidence intervals that might improve on the bootstrap for small samples is to compute Bayesian credibility intervals based on Bayesian version of the PSPP model, with non-informative priors for the parameters. We did not assess this option in our simulations, to keep the number of comparisons manageable.



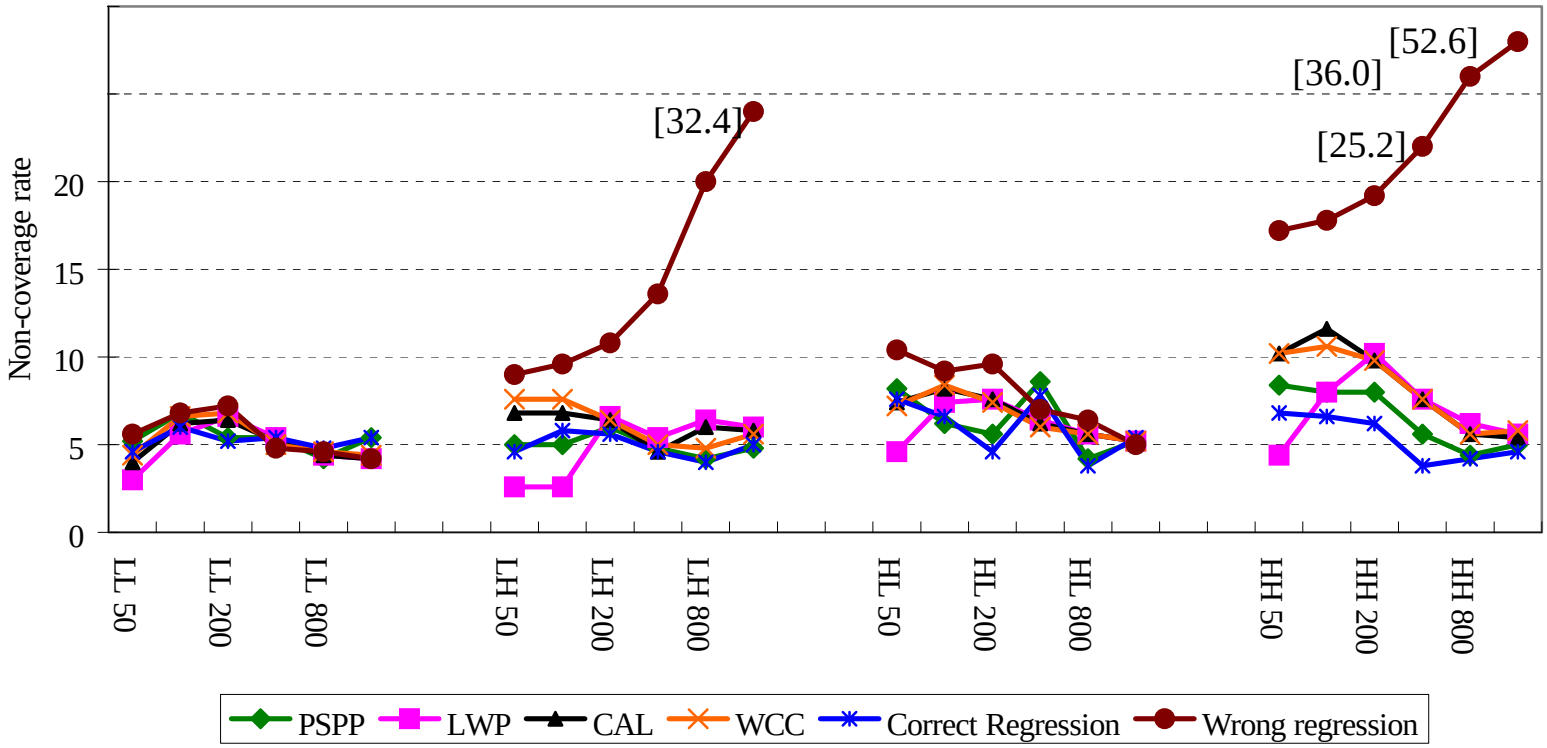
Note: (1) LL: low low ; LH: Low High; HL: High Low; HH: High High.
(2) LL 50: Low low cases, sample size 50.
(3) Points beyond 150 are not in the real scales, numbers in the parenthesis show the real values.

Figure 2. % Increase of Width of CI of Simulation 1

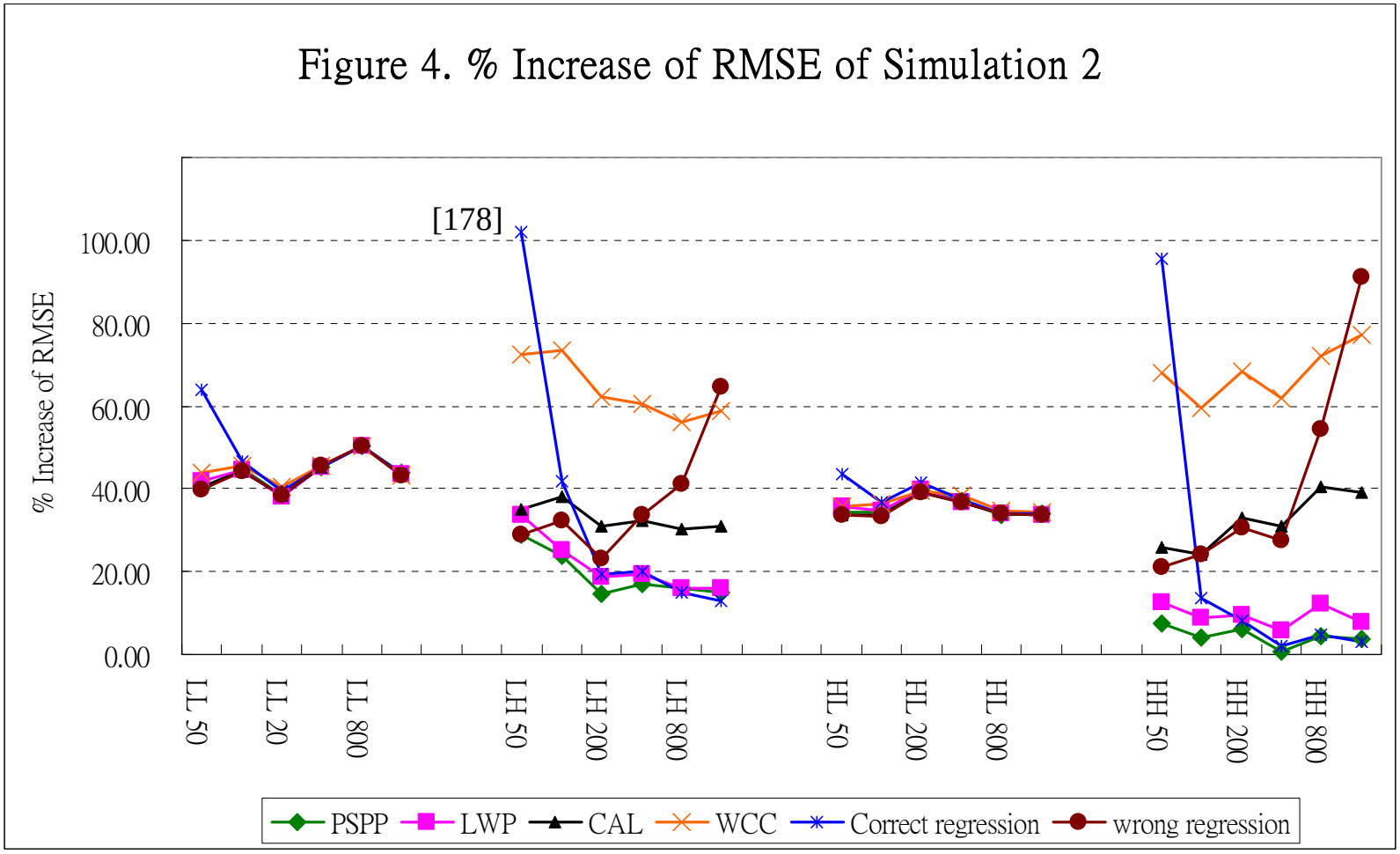


Note: (1) LL: low low ; LH: Low High; HL: High Low; HH: High High.
(2) Points beyond 150 are not in the real scales, numbers in the parenthesis show the real values.

Figure 3. Non-coverage rate of Simulation 1

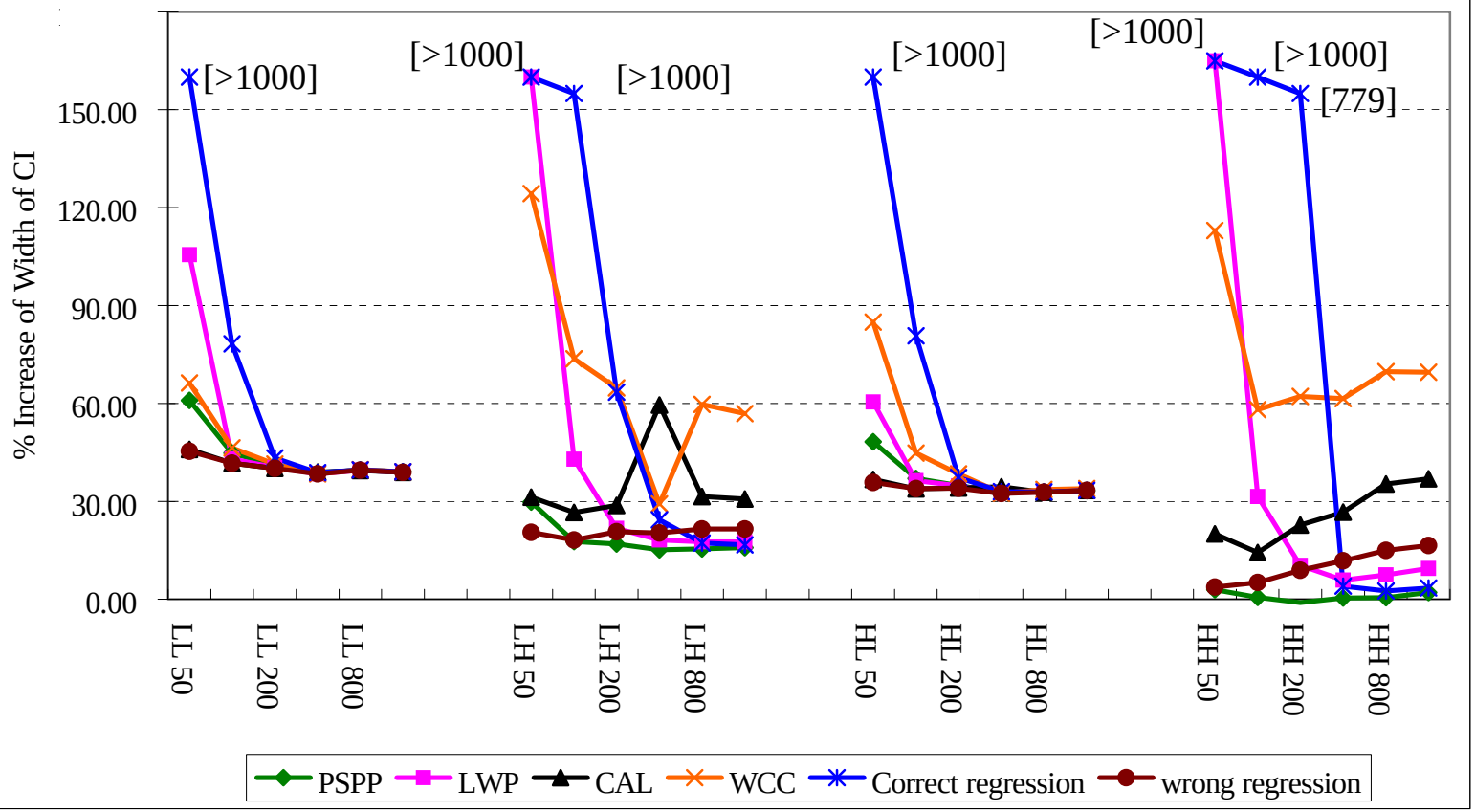


Note: (1) LL: low low ; LH: Low High; HL: High Low; HH: High High.
(2) LL 50: Low low cases, sample size 50.
(3) Points beyond 20 are not in the real scales, numbers in the parenthesis show the real values.



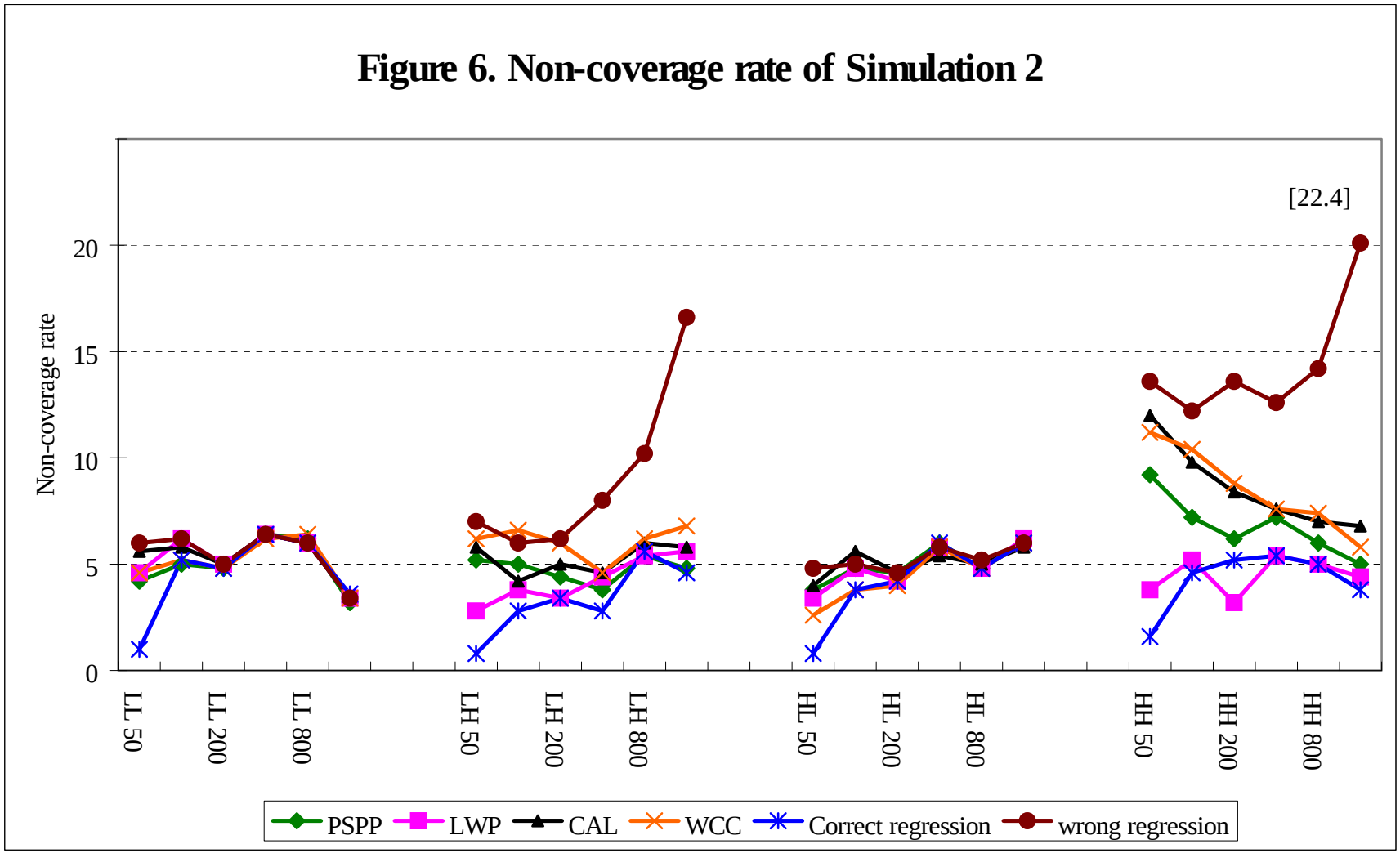
Note: (1) LL: low low ; LH: Low High; HL: High Low; HH: High High.
(2) LL 50: Low low cases, sample size 50.
(3) Points beyond 100 are not in the real scales, numbers in the parenthesis show the real values.

Figure 5. % Increase of Width of CI of Simulation 2

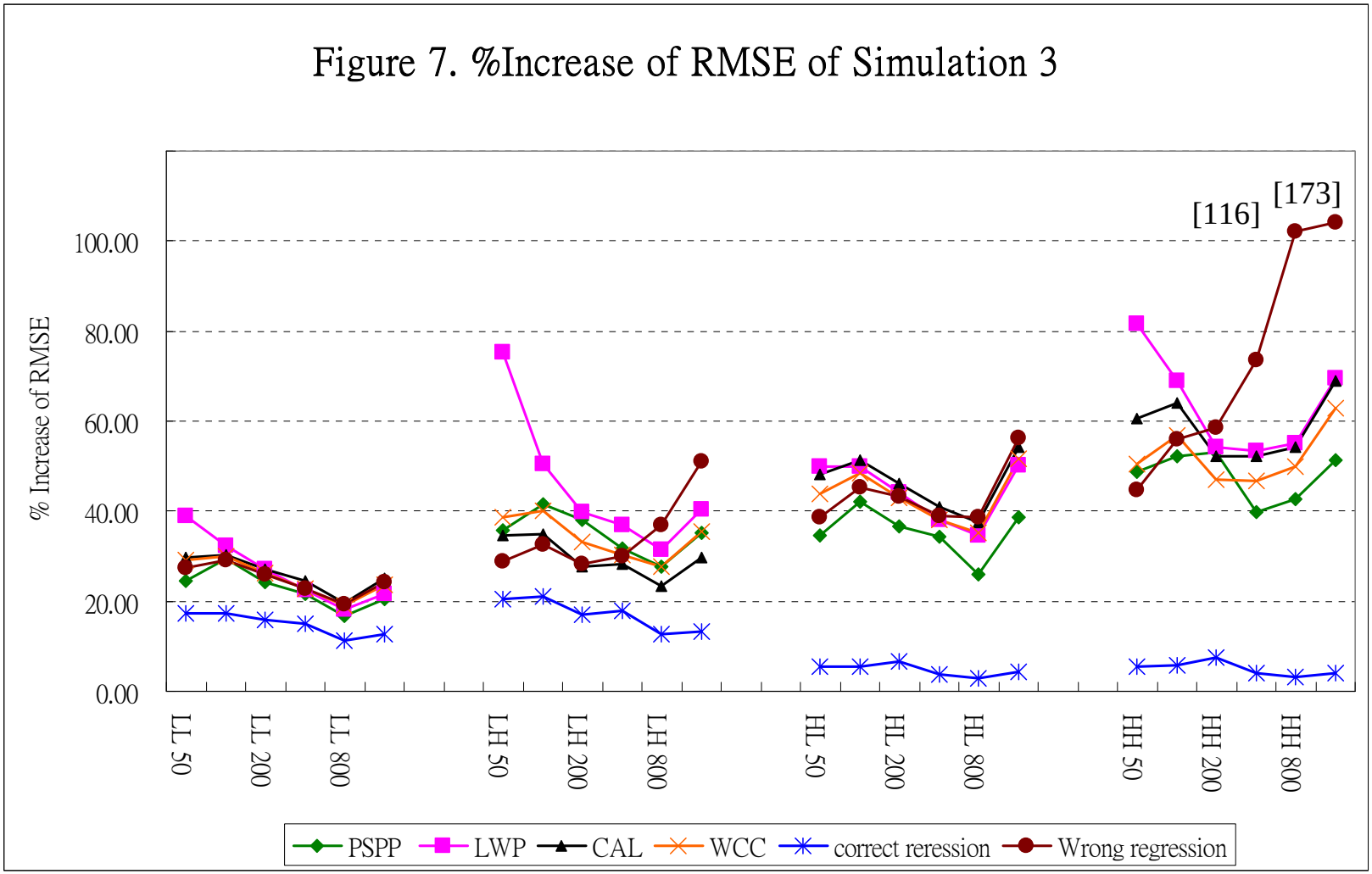


Note: (1) LL: low low ; LH: Low High; HL: High Low; HH: High High.
(2) Points beyond 150 are not in the real scales, numbers in the parenthesis show the real values.

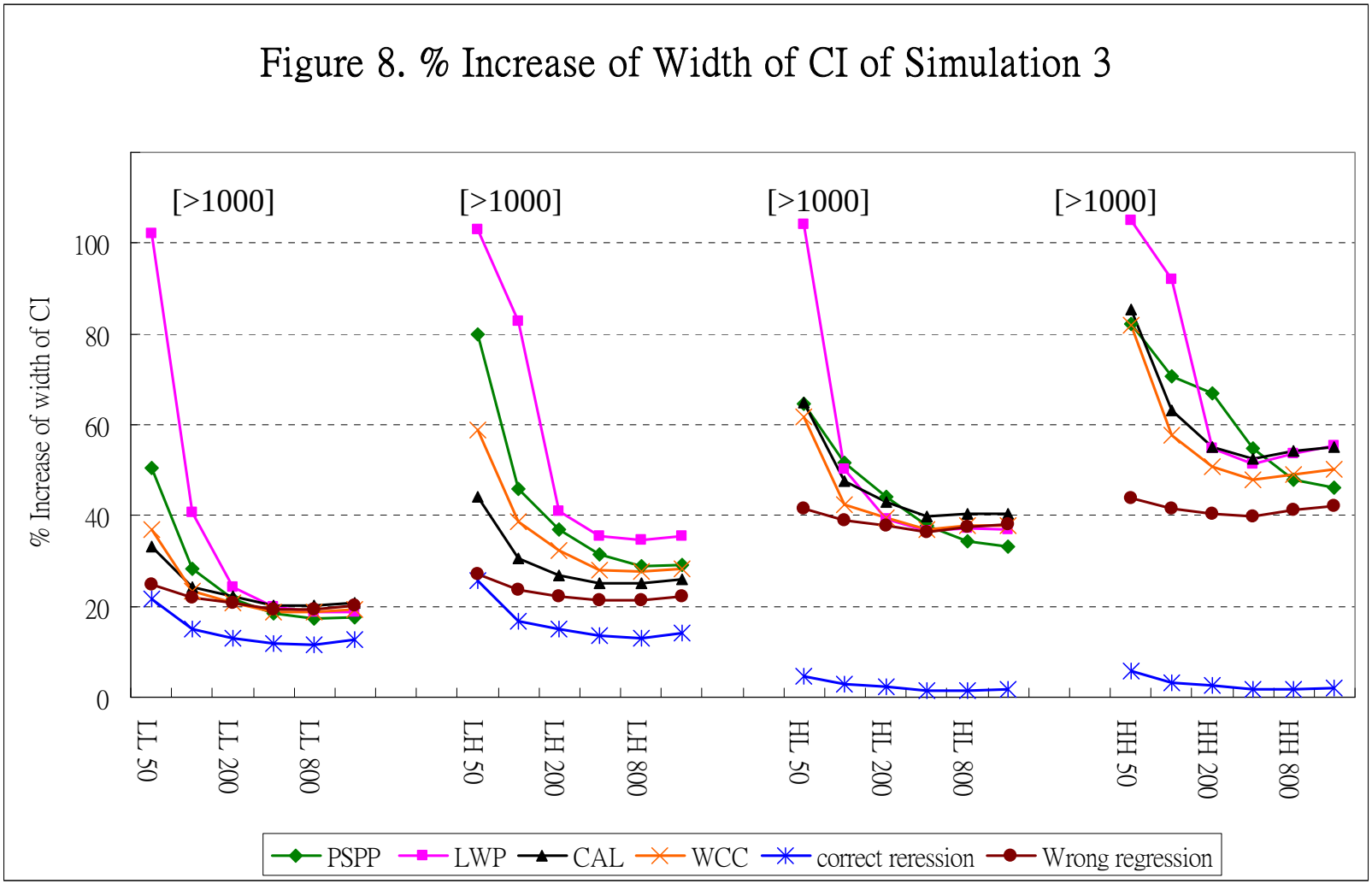
Figure 6. Non-coverage rate of Simulation 2



Note: (1) LL: low low ; LH: Low High; HL: High Low; HH: High High.
(2) Points beyond 200 are not in the real scales, numbers in the parenthesis show the real values

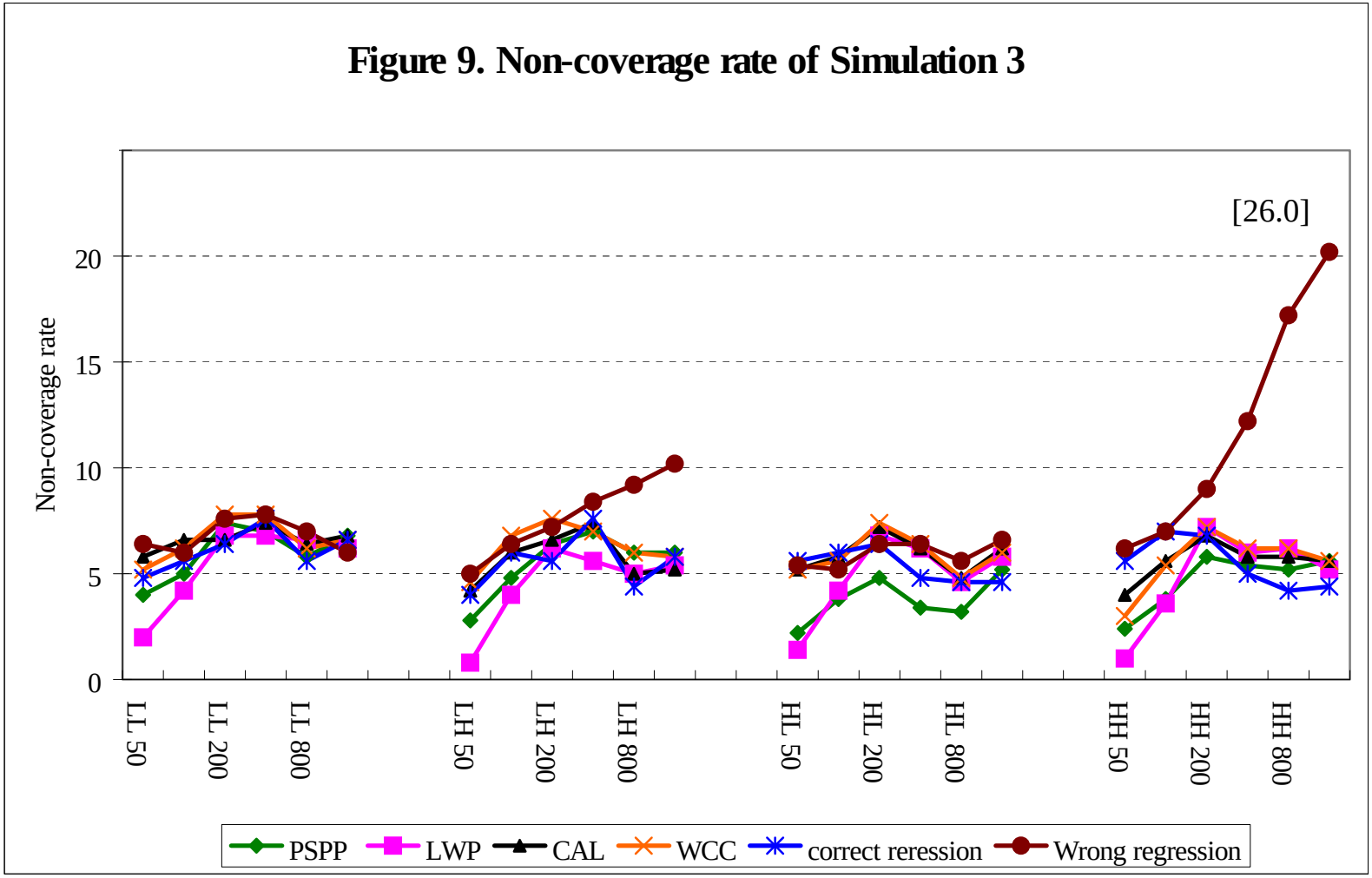


Note: (1) LL: low low ; LH: Low High; HL: High Low; HH: High High.
(2) Points beyond 100 are not in the real scales, numbers in the parenthesis show the real values.



Note: (1) LL: low low ; LH: Low High; HL: High Low; HH: High High.
(2) Points beyond 100 are not in the real scales, numbers in the parenthesis show the real values.

Figure 9. Non-coverage rate of Simulation 3



Note: (1) LL: low low ; LH: Low High; HL: High Low; HH: High High.
(2) Points beyond 20 are not in the real scales, numbers in the parenthesis show the real values.

Table 3.4. Empirical bias, empirical standard error and RMSE when propensity function is wrong specified

	A: logit (PS)=$X_1 + X_2$			B: logit (PS)=X_1			C: logit (PS)=X_2		
	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
BD	-1	9	7	-1	9	7	-1	9	7
CC	10	13	13	10	13	13	10	13	13
(a) $[s(Y^*)]$	1	12	10	-17	11	18	40	14	40
(b) $[s(Y^*) + X_1]$	0	12	10	--	--	--	14	13	15
(c) $[s(Y^*) + X_2]$	0	12	10	14	13	16	--	--	--
(d) $[s(Y^*) + X_2 + X_1 * X_2]$	-1	11	9	-1	11	9	--	--	--
(e) $[\alpha_0 + \alpha_1 * w]$	0	12	10	-18	12	19	40	14	40
(f) $[\alpha_0 + \alpha_1 * w + X_1]$	5	18	14	-17	12	17	14	13	16
(g) $[\alpha_0 + \alpha_1 * w + X_2]$	-3	15	12	14	14	16	40	14	40
(h) $[\alpha_0 + \alpha_1 * w + X_2 + X_1 * X_2]$	-4	15	12	-2	13	10	34	12	34
(i) $[\alpha_0 + \alpha_1 * w + X_1 + X_2 + X_1 * X_2]$	-1	11	9	-1	11	9	-1	11	9
(j) $\hat{\mu} = \bar{y} + n^{-1}(\sum_{i=1}^{r} w_i (y_i - \bar{y}))$	0	13	10	-17	12	18	40	14	40
(k) $\hat{\mu} = n^{-1}(\sum_{i=1}^{n} \hat{y}_{i\_x_1}) + n^{-1}(\sum_{i=1}^{r} w_i (y_i - \hat{y}_{i\_x_1}))$	0	12	9	-17	11	18	10	11	13
(l) $\hat{\mu} = n^{-1}(\sum_{i=1}^{n} \hat{y}_{i\_x_2}) + n^{-1}(\sum_{i=1}^{r} w_i (y_i - \hat{y}_{i\_x_2}))$	0	15	12	16	14	18	40	14	40
(m) $\hat{\mu} = n^{-1}(\sum_{i=1}^{n} \hat{y}_{i\_x_2, x_1 * x_2}) + n^{-1}(\sum_{i=1}^{r} w_i (y_i - \hat{y}_{i\_x_2, x_1 * x_2}))$	0	12	9	5	12	10	34	12	34
(n) $\hat{\mu} = n^{-1}(\sum_{i=1}^{n} \hat{y}_{i\_x_1, x_2, x_1 * x_2}) + n^{-1}(\sum_{i=1}^{r} w_i (y_i - \hat{y}_{i\_x_1, x_2, x_1 * x_2}))$	-1	11	9	-1	11	9	-1	11	9
(o) $\hat{\mu} = (\sum_{i=1}^{r} w_i y_i) / \sum_{i=1}^{r} w_i)$	0	13	10	-17	11	18	40	15	40

Table 3.5 BMI reduction within groups

Method	Treatment		Control	
	Mean (SE)	95% CI	Mean (SE)	95% CI
Complete Case Analysis	-0.91 (0.09)	(-1.09, -0.73)	-0.45 (0.10)	(-0.65, -0.25)
(a) Stratified PSPP	-1.00 (0.10)	(-1.21, -0.80)	-0.42 (0.09)	(-0.61, -0.23)
(b) Linear in the weight prediction	-1.01 (0.10)	(-1.21, -0.81)	-0.42 (0.09)	(-0.60, -0.23)
(c) Calibration estimator	-1.02 (0.10)	(-1.22, -0.81)	-0.42 (0.09)	(-0.61, -0.24)
(d) Weighted complete-case analysis	-1.04 (0.11)	(-1.27, -0.82)	-0.40 (0.09)	(-0.59, -0.21)

*SE and CI denote empirical standard error and confidence interval. SE and 95% CI are based on 200 bootstrap samples.

CHAPTER IV
THE PSPP METHOD FOR THE MONOTONE PATTERN MISSING DATA

4.1 Introduction

In applications of statistics complete data may not be available for every subject. Missing data may arise by experimental design or by happenstance. For example in some two stage studies only a subset of the subjects are selected for expensive tests thus subjects who have not been chosen will have missing measurements; on the other hand some subjects may drop out of the study and make the data collection impossible.

A naïve way to deal with missing data is to discard cases with missing values, and analyze the cases that are complete (Complete-Case Analysis, CC). CC analysis is simple and yields unbiased estimated if missing values are missing completely at random (MCAR), that is, the missingness does not depend on the values of variables in the data set (Little and Rubin, 2002). Weighted complete-case analysis is an alternative to CC analysis and it can reduce bias when the missing data is not MCAR (Little and Rubin, 2002; Horvitz and Thompson, 1952). Like CC analysis, weighted complete-case analysis discards cases with incomplete information, thus there is a potential loss of information.

Imputation is one common approach to make use of the partial information in an incomplete case. For example, consider a data set with variables $X_1,...,X_p,Y_1$, where $X_1,...,X_p$ are fully observed covariates and Y_1 has missing values. One might impute the missing Y_1 by a parametric regression of Y_1 on $X_1,...,X_p$, for example by a linear model

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \varepsilon_i \sim N(0,\sigma^2). \quad (1)$$

The parameters $\beta_0,...,\beta_p$ in (1) can be estimated using the complete cases and the missing values of Y_1 is substituted by the predicted values from the regression model. Uncertainty in the imputations can be reflected by multiple imputation, where multiple sets of draws are imputed from the predictive distribution of the missing values (Rubin, 1998; Little and Rubin, 2002). This approach is implemented in the SAS software PROC MI (1999).

The parametric approach is very efficient if one can model the relationship of Y_1 and the covariates $X_1,...,X_p$ correctly; but it is potentially sensitive to model misspecification, particularly when the data deviate from MCAR (Little and An, 2004). Nonparametric and semiparametric methods weaken the model assumptions. The Penalized Spline Propensity Prediction method of imputation is one such method (Little and An, 2004; Little and Zhang, 2007; Zhang and Little, 2005). Let M be an indicator variable with $M=1$ when Y_1 is missing and $M=0$ when Y_1 is observed. Define the logit of the propensity for Y_1 to be observed as:

$$P^* = \text{logit}\,(\Pr(M=0 \mid X_1,...,X_p;\phi_1)).$$

The key property of the propensity score is that, conditioning on the propensity score and assuming MAR, missingness of Y does not depend on the covariates $X_1,...,X_p$ (Rosenbaum and Rubin, 1983). This motives the Penalized Spline Propensity Prediction Method (PSPP), which is based on the following model:

$$(Y_1 \mid P^*, X_1,...,X_p;\beta) \sim N(s(P^*)+g(P^*,X_2,...,X_p;\beta),\sigma^2) \qquad (2)$$

where $s(P^*)$ is a spline of Y_1 on P^* and g is a parametric function indexed by unknown parameter β. One variable, here X_1, is not included in the g function to prevent multicollinearity. A variety of spline fitting methods are possible (Eilers and Marx, 1996; Ruppert, Wand and Carroll, 2003; Ngo and Wand, 2004; Eubank, 1998; Wahba, 1990). In this paper we choose the penalized spline with truncated linear basis with the form:

$$s(P^*) = \beta_0 + \beta_1 P^* + \sum_{k=1}^{K} \gamma_k (P^* - \kappa_k)_+ , \qquad (3)$$

where 1, $P^*,(P^*-\kappa_1)_+,...,(P^*-\kappa_k)_+$ is the truncated linear basis; $\kappa_1 < .... < \kappa_K$ are selected fixed knots and *K* is the total number of knots. This model can be fitted using a number of existing software packages, such as PROC MIXED in SAS (SAS, 1992; Ngo and Wand, 2004, Littell, Milliken, Stroup, and Wolnger. 1996; Ruppert, 2002) and lme() in S-plus (Pinheiro and Bates, 2000). We fit this model using PROC MIXED in SAS with $(P^*-\kappa_1)_+,...,(P^*-\kappa_k)_+$ treated as random effects and the intercept, P^* and the parametric function $g(P^*, X_2,..., X_p;\beta)$ treated as the fixed effects.

The predicted mean of Y_1 has a doubly robust property meaning that the predicted mean of Y is consistent if either (a) the mean of *Y* given $(P^*, X_1,..., X_p)$ in model (2) is correctly specified, or (b1) the propensity P^* is correctly specified, and (b2) $E(Y \mid P^*) = s(P^*)$. The robustness feature derives from the fact that the regression function g does not have to be correctly specified (An and Little, 2004; Zhang and Little, 2005).

The PSPP method can be extend to derive the conditional means of a missing variable give a covariate. For subgroup means of a missing variable given a categorical covariate, Zhang and Little (2005) proposed the stratified PSPP method which fits different spline curves for the different categories of the covariate. Let X_1 be a categorical variable with *C* levels. Let $I_c = 1$ if $X_1 = c$; $I_c = 0$ if $X_1 \neq c, \quad c = 1,...,C$. The stratified PSPP method is based on the following model:

$$(Y_1 \mid P^*, X_1,..., X_p;\beta) \sim N(\sum_{c=1}^{C} I_c s_c(P^*) + g(P^*, X_1, X_2,..., X_{p-1};\beta), \sigma^2) \qquad (4)$$

Where g is a parametric function indexed by unknown parameter β as before, with X_p dropped to avoid multicollinearity; $I_c s_c(P^*) = I_c(\gamma_{0c} + \sum_{j=1}^{q}\gamma_{jc}(P^*)^j + \sum_{k=1}^{K}\gamma_{qkc}(P^*-\kappa_k)_+^q)$ is the fitted curves for the *c*th level of X_1. Within each level of X_1,

$$E(Y \mid P^*, X_1 = c, X_2,..., X_{p-1};\beta) = s_c(P^*) + g(P^*, X_1 = c, X_2,..., X_{p-1};\beta).$$

This method yields consistent estimates for the conditional means of Y given X_1. The marginal mean of Y is a weighted average of conditional means, which again has the double robustness property.

In this paper we extend the PSPP method and the stratified PSPP methods for the monotone pattern of missing data, where missing data can be arranged in a way that if Y_j is missing in a unit then Y_{j+1}, Y_{j+2}, …, Y_p are missing as well. Let $(X_1,...,X_P,Y_1,Y_2)$ denote a $(P+2)$-dimensional vector of variables with $\underline{X}=(X_1,...,X_P)$ fully observed covariates and Y_1,Y_2 with missing values in a monotone pattern (Figure 10).

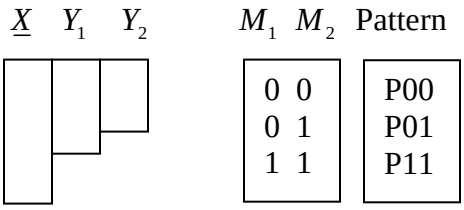


Figure 10 Example of monotone missing data structure

We assume the missing data mechanism is missing at random through out this paper. A standard way to drive the means of Y_1 and Y_2 is to model the joint distribution of Y_1 and Y_2 given the covariates $X_1,...,X_P$, such as $f(Y_1,Y_2 \mid X_1,...,X_P;\theta)$, which can be solved by the factored likelihood approach, that is,

$$f(Y_1,Y_2 \mid X_1,...,X_P;\theta) = f_1(Y_1 \mid X_1,...,X_P;\theta_1) * f_2(Y_2 \mid X_1,...,X_P,Y_1;\theta_2),$$

where θ_1 and θ_2 are unknown parameters (Little and Rubin, 2002). The estimated value of θ_1, $\hat{\theta}_1$, can be derived based on the cases with Y_1 observed and the estimated value of θ_2, $\hat{\theta}_2$ can be derived based on the cases with both Y_1 and Y_2 observed. This parametric approach is efficient if the joint distribution is modeled correctly, but it is very vulnerable to model misspecification.

We propose a stepwise PSPP procedure for the monotone missing data to drive robust estimates for the means of missing variables, described in section 4.2. We illustrate our method with simulation studies and compare our method with several simple alternatives in section 4.3. We apply the stepwise PSPP method to an online weight loss study in section 4.4 and we end this paper with a discussion in section 4.5.

4.2 PSPP for the monotone pattern of missing data

We first consider the case with two missing variables. Let $(X_1,...,X_P,Y_1,Y_2)$ denote a $(P+2)$-dimensional vector of variables with $\underline{X}=(X_1,...,X_P)$ fully observed covariates and Y_1,Y_2 with missing values in a monotone pattern (Figure 10). Let M_1 and M_2 be indicator variables, with $M_i=0$ when Y_i is present and $M_i=1$ when Y_i is missing, $i=1,2$. We can divide the dataset into three parts based on the missing data pattern. The first part contains subjects with both Y_1 and Y_2 present, denoted as P00; the second part contains cases with Y_1 present and Y_2 missing, denoted as P01; the third part contains cases with both Y_1 and Y_2 missing, denoted as P11.

We assume the missingness of Y_1 depends only on the fully-observed covariates $\underline{X}$, which can be represented by the conditional distribution of M_1 given $\underline{X}$, $f(M_1 \mid \underline{X};\phi_1)$. Define the logit of the propensity for Y_1 to be observed as:

$$P_1^*=\text{logit}\,(\Pr(M_1=0 \mid X_1,...,X_p;\phi_1)),$$

and $\hat{P}_1^*=\text{logit}\,(\Pr(M_1=0 \mid X_1,...,X_p;\hat{\phi}_1))$, where $\hat{\phi}_1$ is estimated from logistic regression of M_1 on $\underline{X}$ from all cases. When Y_1 is present, under MAR, the missing of Y_2 depends on $\underline{X}$ and Y_1, which can be represented by the conditional distribution of M_2 given $\underline{X}$, Y_1 and $M_1=0$, that is $f(M_2 \mid \underline{X},Y_1,M_1=0;\phi_2)$. Define the logit of the propensity for Y_2 to be observed as:

$$P_2^*=\text{logit}\,(\Pr(M_2=0 \mid X_1,...,X_p,Y_1,M_1=0;\phi_2)),$$

and $\hat{P}_2^*=\text{logit}\,(\Pr(M_2=0 \mid X_1,...,X_p,Y_1,M_1=0;\hat{\phi}_2))$, where $\hat{\phi}_2$ is estimated from logistic regression of M_2 given $\underline{X}$, Y_1 from cases in P00 and P01.

The natural way to implement PSPP in this setting (by analogy with parametric methods, as in Little and Rubin 2002, chapter 7) is to (a) impute missing values of Y_1 in P11 by regression of Y_1 on $\underline{X}$ and the spline of $\hat{P}_1^*$ and then (b) impute missing values of Y_2 in P01 and P11 by a regression of Y_2 on $\underline{X}$, Y_1 and a spline of $\hat{P}_2^*$, where missing values of Y_1 in P11 are replaced by estimates from (a). This methods yields DR estimate of the mean of Y_1, but does not yield DR estimates of the mean of Y_2, since (a) the estimates of the imputed values of Y_1 in P11 need to be based on a correct prediction model, and (b) the propensity score $\hat{P}_2^*$ only applies to cases in P01 and not to cases in P11, since conditional on $M_1 = 1$, the probability that Y_2 observed is 0, given the monotone pattern. We propose an alternative stepwise PSPP approach that preserves the DR property for the mean of Y_2.

The stepwise PSPP procedure imputes (a) the missing Y_2 in P01 by a regression of Y_2 on $\underline{X}$, Y_1 and a spline of $\hat{P}_2^*$ estimated using cases in P00. After filling in the missing values in this part, we have a two-patterns data structure, where Y_1 and Y_2 are missing for the same set of cases. So we can borrow the propensity score of Y_1 to impute the last part of missing Y_2. This is the general idea of the stepwise PSPP procedure and the key is to derive $\hat{Y}_2$ of P01 to be a random sample of the original data conditioning on the propensity scores.

To derive consistent marginal mean of Y_2 we need to condition on the propensity score $\hat{P}_2^*$ when filling in the missing Y_2 in P01; in addition, we also need to condition on the propensity score of Y_1, $\hat{P}_1^*$, in this step due to the fact that we want to fill in the last part of missing Y_2 conditioning on $\hat{P}_1^*$. There are two ways to impute the missing Y_2 in P01:

(1) Imputation based on the conditional propensity scores. We impute the missing Y_2 in P01 by a bivariate spline $s(\hat{P}_1^*, \hat{P}_2^*)$, where $\hat{P}_1^*$ and $\hat{P}_2^*$ are conditional propensity scores. Estimation of the bivariate smoothing function $s(\hat{P}_1^*, \hat{P}_2^*)$ requires bivariate basis functions, which can be derived in several different ways. We choose the tensor product basis (Ruppert, Wand and Carroll, 2003) to estimate $s(\hat{P}_1^*, \hat{P}_2^*)$ in this paper. With this basis, the bivariate function $s(\hat{P}_1^*, \hat{P}_2^*)$ can be written as

$$s(\hat{P}_1^*, \hat{P}_2^*) = \alpha_0 + \alpha_1 \hat{P}_1^* + \sum_{k=1}^{K_1} \gamma_{1k} (\hat{P}_1^* - \kappa_{1k})_+ + \alpha_2 \hat{P}_2^* + \sum_{k'=1}^{K_2} \gamma_{2k'} (\hat{P}_2^* - \kappa_{2k'})_+ + \alpha_3 \hat{P}_1^* \hat{P}_2^*$$
$$+ \sum_{k=1}^{K_1} \gamma_{3k} \hat{P}_2^* (\hat{P}_1^* - \kappa_{1k})_+ + \sum_{k'=1}^{K_2} \gamma_{4k'} \hat{P}_1^* (\hat{P}_2^* - \kappa_{2k'})_+ + \sum_{k=1}^{K_1} \sum_{k'=1}^{K_2} \gamma_{5kk'} (\hat{P}_1^* - \kappa_{1k})_+ (\hat{P}_2^* - \kappa_{2k'})_+ \quad (4)$$

where $\kappa_{11} < ... < \kappa_{1K_1}$ and $\kappa_{21} < ... < \kappa_{2K_2}$ are selected fixed knots for $\hat{P}_1^*$ and $\hat{P}_2^*$ respectively. We choose 5 equally spaced knots for $\hat{P}_1^*$ and $\hat{P}_2^*$ respectively. When we have more than two missing variables, we can follow the same idea but the high dimensional spline models will be hard to fit. In that case we can impute the missing values based on the marginal propensity function as described below.

(2) Imputation based on the marginal propensity scores. We impute the missing Y_2 of P01 by a penalized spline $s(\hat{P}_{2\_m}^*)$, where

$$\hat{P}_{2\_m}^* = \text{logit}\,(\hat{P}(M_1 = 0 \mid \underline{X}) * \hat{P}(M_2 = 0 \mid \underline{X}, Y_1, M_1 = 0))$$

is the estimated marginal propensity score for Y_2. It is the probability of Y_2 to be observed in the end. This marginal propensity score can be derived easily and we do not need high dimensional spline functions to fill in the missing values.

After fill in the missing Y_2 in P01 we can impute the last part of missing Y_2 by a penalized spline of $\hat{P}_1^*$. We can enrich the imputation model by adding the parametric g function. The above stepwise procedure can be easily extended to derive the subgroup means of a missing variable. We can apply the stratified PSPP method in a stepwise

procedure to fit different spline curves for different subgroups. For the conditional stepwise procedure, we first impute missing Y_2 by the following model: $Y_2 = \sum_{r=1}^{2} I(X_1 = r) * s_r(\hat{P}_2^*, \hat{P}_1^*) + \varepsilon, \varepsilon \sim N(0, \sigma^2)$, where $s_r(\hat{P}_2^*, \hat{P}_1^*)$ is bivariate spline of form (4) for subgroup *r*. For the second step, we apply stratified PSPP as follows: $Y_2 = \sum_{r=1}^{2} I(X_1 = r) * s_r(\hat{P}_1^*) + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$, where $s_r(\hat{P}_1^*)$ is a penalized spline for subgroup *r*. For the marginal stepwise procedure, we first impute missing Y_2 by the following model: $Y_2 = \sum_{r=1}^{2} I(X_1 = r) * s_r(\hat{P}_{2\_m}^*) + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$, where $s_r(\hat{P}_{2\_m}^*)$ is a penalized spline of the marginal propensity scores for subgroup *r*. The second step is the same as the conditional stepwise procedure described above. We illustrate the stepwise PSPP method in a simulation study in section 4.3.

4.3 Simulation study

We report results of a simulation study to evaluate the performance of the stepwise PSPP procedure. We generate 500 datasets with 1000 subjects, with two fully observed covariates as follows: $X_1 \sim Binomial(0.4)$, $X_2 \sim N(0,1)$, where X_1, X_2 are independent. We simulate response variables Y_1 and Y_2 from normal distributions with mean of Y_1, Y_2 as

$$Y_1 \mid \underline{X} \sim N(\ I(X_1 = 1)\ +\ 5 * I(X_1 = 0)\ +\ X_2\ +\ X_2^2,\ 1)$$

$$Y_2 \mid X_1, X_2, Y_1 \sim N(X_2\ +\ Y_1\ -\ I(X_1 = 1),\ 1)$$

where I() is an indicator variable. The missing data mechanism is missing at random with the propensity model:

$$P_1^* = I(X_1 = 1) - 0.5 * I(X_1 = 0)\ +\ X_2 + 1$$

$$P_2^* = 0.5 * I(X_1 = 1)\ - I * (X_1 = 0)\ + 0.5 * Y_1 - 0.7$$

where about 30% of Y_1 and 50% of Y_2 are missing. We can impute the missing Y_1 by applying the stratified PSPP method directly, and we will omit the results for Y_1 in this paper. We derive the marginal and conditional means of Y_2 based on the following methods:

(1) Before deletion analysis (BD)

(2) Complete case analysis (CC)

(3) Correctly specified linear regression models for Y_1, Y_2 respectively, namely, $[Y_1 : X_1, X_2, X_2^2]$, $[Y_2 : X_1, X_2, \hat{Y}_1]$, where $\hat{Y}_1$ is the imputed values of Y_1 from the model $[Y_1 : X_1, X_2, X_2^2]$.

(4) Wrongly specified linear regression models for Y_1, Y_2 respectively, namely, $[Y_1 : X_1, X_2]$, $[Y_2 : X_1, X_2, \hat{Y}_1]$. The first model does not have the quadratic term X_2^2 thus is wrongly specified.

(5) Stratified stepwise PSPP method based on the conditional propensity scores described in part 2 with null g function, namely: $[\text{step 1: } Y_2 = \sum_{r=1}^{2} I(X_1 = r) * s_r(\hat{P}_2^*, \hat{P}_1^*)], [\text{Step 2: } Y_2 = \sum_{r=1}^{2} I(X_1 = r) * s_r(\hat{P}_1^*)]$

(6) Stratified stepwise PSPP method based on the marginal propensity scores described in part 2 with null g function, namely: $[\text{step 1: } Y_2 = \sum_{r=1}^{2} I(X_1 = r) * s_r(\hat{P}_{2\_m}^*)], [\text{Step 2: } Y_2 = \sum_{r=1}^{2} I(X_1 = r) * s_r(\hat{P}_1^*)]$

(7) For imputation of missing Y_2, we compared our method with the weighted complete-case analysis. The marginal mean of Y_2 is calculated based on weighted mean of the complete cases, that is $\hat{\mu}_2 = \frac{\sum_{i=1}^{r} y_{2,i} * \hat{w}_i}{\sum_{i=1}^{r} \hat{w}_i}$, where $\hat{w}_i$ is the weight for the ith subject, which equals to the inverse of the marginal propensity score as follows: $\hat{w}_i = 1/[\hat{P}(M_{1,i}=0|\underline{X}) * \hat{P}(M_{2,i}=0|\underline{X}, Y_1, M_{1,i}=0)]$, where $\hat{P}(M_{1,i}=0|\underline{X})$ and $\hat{P}(M_{2,i}=0|\underline{X}, Y_1, M_{1,i}=0)$ are estimated from correctly specified logistic regression models. For the conditional means we apply the above formula to each subgroup.

We derived the marginal mean and conditional means of Y_2 given X_1. Empirical bias, empirical standard error (SE) and root mean square error (RMSE) are summarized

in Table 4.1. The complete case analysis yields estimates with large bias and RMSE compared to the before deletion analysis. Correctly specified ANCOVA models yields estimates with small bias and RMSE. The wrongly specified ANCOVA model yield biased results for both marginal and conditional means. The two stepwise PSPP methods yield estimates with small bias and RMSE for both marginal and conditional means of Y_2. The results are very close to each other based on the conditional and the marginal propensity scores at the first step. The weighted complete-case analysis yields estimates with small bias and RMSE for both marginal and conditional means.

4.4 Example

We apply the stepwise PSPP method to an online weight loss study conducted by Kaiser Permanente (Couper, Peytchev, Little, Strecher, Rothert, 2005). Approximately 4,000 subjects were randomly assigned to the treatment or the control group. The weight loss information was posted online and the participants were encouraged to read the posted health related topics. Tailored information was available to the treatment group subjects. For the control group, all the subjects have the same untailored information. At 3, 6 and 12 month, follow-up surveys were sent to the participants, which collected measurements such as current weight. Our goal is to compare short-term and long-term treatment effects; in particular, we compare the reduction of the body mass index (BMI), defined as the difference of the follow-up BMI and the baseline BMI.

There were 2059 subjects in the treatment group and 1956 subjects in the control group at the baseline. At 3 month 623 subjects in the treatment group and 611 subjects in the control group responded to the second survey. At 6 month 438 subjects in the treatment group and 397 subjects in the control group remained in the study. At 12 month 277 subjects in the treatment group and 314 subjects in the control group responded to the last survey. We assume the data is missing at random. Comparisons of the baseline measurements between subjects remained in the study and those dropped out at 3 month indicate subjects who remained in the study have much lower baseline BMI than those who dropped out of the study for the treatment group ($P<0.001$), but this differences is not seen in the control group ($P=0.47$); On the other hand, subjects who remained in the

study at 3 month have better baseline health, as shown by the number of previous disease, than those who dropped out of the study for the control group(P<0.01), but this differences was not seen in the treatment group (P=0.56). Similarly, subjects who remained in the study at 6 month have much lower baseline BMI than those who dropped out of the study for the treatment group (P<0.001), but this difference is not seen in the control group (P=0.82). These differences suggest interactions between treatment and baseline covariates are included when estimating the propensity scores.

We estimate the propensity scores by logistic regressions and we keep all the variables with P-value less than 0.20 to get best estimates of the propensity scores. Table 4.2 contains the covariates in the propensity models. We apply the PSPP methods to the data to derive the BMI reduction of 3, 6, and 12 month. For BMI reduction at 3 month, we apply stratified PSPP method to the data as follows.

(a) Stratified PSPP method with null the g function, denoted as $[\sum_{c=1}^{2} I_c s_c(\hat{P}_1^*)]$, where $\hat{P}_1^*$ is the estimated propensity scores of the 3 month.

(b) Stratified PSPP method with baseline covariates in the g function, denoted as $[\sum_{c=1}^{2} I_c s_c(\hat{P}_1^*) + g(\text{baseline covariates})]$.

For BMI reduction at 6 month, we apply stepwise stratified PSPP method as follows.

(c) Stepwise Stratified PSPP method with null the g function, denoted as $[\text{step1}:\sum_{c=1}^{2} I_c s_c(\hat{P}_{2\_m}^*), \text{step2}:\sum_{c=1}^{2} I_c s_c(\hat{P}_1^*),]$, where $\hat{P}_{2\_m}^*$ is the estimated marginal propensity scores of 6m, $\hat{P}_1^*$ is the estimated propensity score of 3m.

(d) Stepwise Stratified PSPP method with the g function, denoted as $[\text{step1}:\sum_{c=1}^{2} I_c s_c(\hat{P}_{2\_m}^*) + g(\text{covariates}), \text{step2}:\sum_{c=1}^{2} I_c s_c(\hat{P}_1^*) + g(\text{covariates})]$.

For BMI reduction at 12 month, we apply stepwise stratified PSPP method as follows.

(e) Stepwise stratified PSPP method with null the g function, denoted as $[\text{step1}:\sum_{c=1}^{2} I_c s_c(\hat{P}^*_{3\_m}); \text{step2}:\sum_{c=1}^{2} I_c s_c(\hat{P}^*_{2\_m}); \text{step3}:\sum_{c=1}^{2} I_c s_c(\hat{P}^*_1)]$, where $\hat{P}^*_{3\_m}$ is the estimated marginal propensity scores of 12m, $\hat{P}^*_{2\_m}$ is the estimated marginal propensity scores of 6m, $\hat{P}^*_1$ is the estimated propensity score of 3m.

(f) Stepwise Stratified PSPP method with null the g function, denoted as $[\text{step1}:\sum_{c=1}^{2} I_c s_c(\hat{P}^*_{3\_m}) + g; \text{step2}:\sum_{c=1}^{2} I_c s_c(\hat{P}^*_{2\_m}) + g; \text{step3}:\sum_{c=1}^{2} I_c s_c(\hat{P}^*_1) + g]$.

For the parametric function g in models (b), (d), (f), we include baseline covariates which predict the outcome, the reduction of BMI. Table 4.3 contains the list of covariates in the g functions. For the variables with missing values, we include that variable in the g function when it is available. For example, to estimate BMI reduction at 12 month, we include 6m BMI in the first step of the stepwise PSPP method, but exclude it in the second and third steps.

In addition to the PSPP methods, we fit regression models to impute the missing BMI reductions for 3, 6 and 12 month sequentially. The covariates in the regression models are the same as the covariates in the g-functions of the PSPP methods, with an extra categorical variable, which indicates the subject is in the treatment or the control group. We also include weighted complete cases analysis for comparison. We estimated the weights based on the same logistic regression model as in the PSPP methods and apply the weighting procedure to the treatment or the control group separately.

We compare our method with complete case analysis. Empirical Standard errors (SE) and the corresponding confidence intervals are obtained from 200 bootstrap samples. Results are summarized in Table 4.4. The treatment group has a larger reduction of BMI after 3 month (-0.91 (0.09)) compared to the control group (-0.45 (0.10)) based on the complete case analysis. The 95% confidence intervals for the treatment group do not overlap with the control group suggesting a treatment effect on the weight loss. At 6 and 12 month, the difference between the two groups is not statistically significant. The PSPP

methods yield the same conclusions, except that treatment effects at 3 and 6 months are stronger after imputation and the BMI reductions for the treatment and control groups increase monotonically. Adding g function into the model does not affect bias but improves efficiency. We did not find much difference for the methods with and without the parametric function in the weight loss study, except for the 12-month BMI reduction, where adding parametric function g reduces variance of the estimates significantly. Weighted complete cases analysis and the sequential regression models yield similar conclusions as the PSPP methods.

The results show that the treatment group has a fast response in terms of BMI reduction. But later on, subjects in the control group catch up and the two groups show similar levels of weight loss. These results suggest that the tailoring information does help subjects to lose weight, especially in the beginning. Later on, if the subjects continues trying to lose weight, then tailoring and untailoring information does not matter that much. It is very reasonable since people who have a long-lasting motivation to lose weight will practice weight loss method continuously and benefit from it thus tailoring information does not have much gain over the untailored information.

4.5 Conclusion

The Penalized Spline Propensity Prediction method imputes the missing values conditioning on the propensity score of being observed and the fully-observed covariates. It yields robust estimators for the marginal and conditional means of a missing variable. We extend it to the monotone pattern missing data by apply the PSPP method in a stepwise procedure. When the missing variables are low dimensional, for example when we have two missing values in a monotone pattern, we apply bivariate spline in the first step of imputation. But when we have a high dimensional missing data, we will have problems fitting non-parametric models due to the curse of dimensionality. In that case, we propose to derive marginal propensity scores and impute the missing values by the penalized spline functions. The bivariate spline conditioning on the conditional propensity scores yields estimates with smaller bias and RMSE than the penalized spline conditioning on the marginal propensity scores. But since it fit high dimensional spline

functions it requires large sample size. The marginal propensity score approach reduces the high dimensional propensity scores to one dimension and is easier to fit and usually requires smaller sample size than the bivariate spline functions.

This stepwise PSPP method based on the marginal propensity scores is similar to the linear in the weight method (Bang and Robins, 2005), where the weight or the inverse of the propensity scores is included in the imputation model parametrically. For the monotone pattern of missing data, the linear in the weight method can also be applied in a stepwise procedure by including the inverse of the marginal propensity scores in the imputation model. But unlike the PSPP method, the linear in the weight method does not extend to derive the subgroup means easily. We can include the subgroup indicator variable in the imputation model, but it does not guarantee the consistent estimates of the conditional means (Table 4.1, last row). An alternative is to apply the linear in the weight method to each subgroup separately but it will be less efficient due the smaller sample size within the subgroups.

Table 4.1. Bias, STD and RMSE for the marginal and conditional means.

	$E(Y_2)$			$E(Y_2 \mid X_1 = 1)$			$E(Y_2 \mid X_1 = 2)$		
	Bias	STD	RMSE	Bias	STD	RMSE	Bias	STD	RMSE
(1) BD	0	11	9	2	14	11	0	11	9
(2) CC	71	18	71	81	21	81	106	19	106
(3) correct ANCOVA $[Y_1 : X_1, X_2, X_2^2]$, $[Y_2 : X_1, X_2, \hat{Y}_1]$	-1	12	10	1	15	12	0	12	10
(4) wrong ANCOVA $[Y_1 : X_1, X_2]$, $[Y_2 : X_1, X_2, \hat{Y}_1]$	-20	13	21	-13	16	17	-23	14	24
(5) Conditional Stepwise PSPP [step 1: $Y_2 = \sum_{r=1}^{2} I(X_1 = r) * s_r(\hat{P}_2^*, \hat{P}_1^*)$] [Step 2: $Y_2 = \sum_{r=1}^{2} I(X_1 = r) * s_r(\hat{P}_1^*)$]	-2	12	10	1	15	12	-2	13	10
(6) Marginal Stepwise PSPP [step 1: $Y_2 = \sum_{r=1}^{2} I(X_1 = r) * s_r(\hat{P}_{2\_m}^*)$] [Step 2: $Y_2 = \sum_{r=1}^{2} I(X_1 = r) * s_r(\hat{P}_1^*)$]	2	12	10	4	16	13	3	13	11
(7) Weighted complete case analysis	1	13	10	2	17	14	0	14	11
(8) Linear in weight	4	26	21	38	21	38	-18	36	31

Table 4.2. Covariates in the propensity model

3month propensity model	(1) baseline BMI; (2) number of previous disease; (3) baseline self care; (4) which is harder –eating less or being active; baseline exercise support; (5) baseline activity level; (6) baseline eating topology; (7) education; (8) ethnic identity; (9) treatment; (10) interaction of treatment with the following covariates: baseline BMI; baseline eating topology; baseline activity level; number of previous disease; which is harder –eating less or being active.
6month propensity model	(1) ethnic identity; (2) baseline weight past year; (3) which is harder –eating less or being active; (4) baseline eating topology; (5) baseline activity level; (6) baseline TV time; (7) education; (8) baseline motivation; (9) treatment; (10)interaction of treatment with the following covariates:ethnic identity; baseline weight past year; which is harder – eating less or being active; baseline eating topology; education; baseline motivation.
12month propensity model	(1) ethnic identity; (2) baseline activity level; (3) baseline number of meals per day (4) Baseline BMI (5) 3 month BMI

Table 4.3. The baseline covariates in the *g* function of model b, d and f.

	Covariates in the *g* function
3 month BMI reduction	(1) ethnic identity; (2) baseline medical advice; (3) baseline eating topology; (4) baseline cardio exercise; (5) baseline activity level; (6) baseline BMI; (7) number of previous disease; (8) number of weigh loss methods tried; (9) motivation of weigh loss; (10) which is harder –eating less or being active.
6 month BMI reduction	(1) ethnic identity; (2) Education (3) Baseline BMI (4) 3month BMI (5) Age (6) Baseline hip length
12 month BMI reduction	(1) BMI_0m (2) BMI_3m (3) BMI_6m (4) Baseline eating pattern (5) Treatment (6) Interaction of treatment with the following variables: baseline BMI, 3 month BMI, 6 month BMI

Table 4.4. BMI reduction within groups.

		Treatment			Control	
		Mean (SE)	95% CI		Mean (SE)	95% CI
3m	CC	-0.91(0.09)	(-1.09, -0.73)		-0.45(0.10)	(-0.65, -0.25)
	Stratified PSPP, null the *g* function	-1.01 (0.11)	(-1.23, -0.78)		-0.40 (0.10)	(-0.60, -0.20)
	Stratified PSPP, with the *g* function	-1.00 (0.10)	(-1.21, -0.80)		-0.42 (0.09)	(-0.61, -0.23)
	Sequential Regression	-0.97 (0.09)	(-1.16. –0.78)		-0.46 (0.10)	(-0.66, -0.27)
	Weighted complete case analysis	-1.04 (0.11)	(-1.27, -0.82)		-0.40 (0.09)	(-0.59, -0.21)
6m	CC	-0.88 (0.15)	(-1.18, -0.58)		-0.63 (0.15)	(-0.93, -0.33)
	Stepwise Stratified PSPP, null the *g* function	-1.11 (0.22)	(-1.56, -0.67)		-0.57 (0.18)	(-0.93, -0.21)
	Stepwise Stratified PSPP, with the *g* function	-1.18 (0.19)	(-1.57, -0.80)		-0.54 (0.17)	(-0.89, -0.19)
	Sequential Regression	-1.01 (0.13)	(-1.27, -0.76)		-0.61 (0.15)	(-0.92, -0.30)
	Weighted complete case analysis	-1.11 (0.22)	(-1.55, -0.67)		-0.54 (0.17)	(-0.88, -0.19)
12m	CC	-1.24 (0.17)	(-1.58, 0.91)		-0.93 (0.23)	(-1.40, -0.47)
	Stepwise Stratified PSPP, null the *g* function	-1.63 (1.30)	(-4.23, 0.98)		-0.82 (0.50)	(-1.82, 0.18)
	Stepwise Stratified PSPP, with the *g* function	-1.29 (0.25)	(-1.79, -0.79)		-0.97 (0.31)	(-1.58, -0.36)
	Sequential Regression	-1.24 (0.17)	(-1.59, -0.89)		-0.91 (0.22)	(-1.34, -0.48)
	Weighted complete case analysis	-1.42 (0.37)	(-2.15, -0.69)		-0.65 (0.54)	(-1.73, 0.43)

*SE and CI denote empirical standard error and confidence interval. SE and 95% CI is based on 200 bootstrap samples.

CHAPTER V

CONCLUSION AND THE FUTURE WORK

Many methods have been proposed in applications of missing data problems. The parametric methods are efficient when the model assumptions are correct but yield biased results when the assumptions are wrong. Non-parametric and semiparametric methods weaken model assumptions and capture the non-linear relationship between the response and the predictors but they face the curse of dimensionality when the number of covariates increases. The PSPP method addresses the curse of dimensionality by focusing on the propensity score of the missing data. It yields unbiased marginal mean estimates with a double robustness property.

We simplify and extend the PSPP method in Chapter II and IV. In Chapter II, we simplify the PSPP methods by including the covariates into the model without centering. We also describe the extensions of the PSPP methods, namely stratified PSPP method and bivariate PSPP method, which yield unbiased conditional means of the missing variable given a covariate. The key of the extended PSPP methods is to add the interactions of the propensity score and the covariate in the model. We propose to use the PSPP methods in a stepwise procedure for the monotone missing data in Chapter IV. This stepwise procedure yields consistent mean estimation as shown in the simulation study.

We compare the PSPP method with several similar double robustness estimators in Chapter III. All these method yields consistent estimates when the propensity score is correctly specified, but the PSPP method yields estimates with smaller RRMSE and width of confidence interval when the complete cases is not a random sample of the original data. The linear in the weight method and calibration method are very sensitive to extreme propensity scores when the sample size is small and yield large confidence

intervals. This is because that the small propensity score corresponds to the large weights which will lead to out-of-range predictions. The PSPP method, on the other hand, estimates a spline curve through the propensity scores and the curvature prevents extreme predictions with the small propensity scores.

We show the PSPP method yields doubly robust estimates of the conditional mean of a missing variable given a covariate. More generally, a PSPP method that yields doubly robust estimates of the conditional mean of Y given a subset of the covariates $(X_1,...,X_s)$, $s < p$, requires inclusion of the interactions between the propensity score and $(X_1,...,X_s)$; but the curse of dimensionality will be a challenge, especially when the sample size is small. In that case we do not recommend the PSPP method. Parametric approach should be considered instead.

We use bootstrap method to estimate the confidence interval and non-coverage in the study. It is easy to implement but it is computational intensive for large samples. An alternative method is to derive the inference by Bayesian method. Instead of using mean predictions we can impute the missing values by posterior draws. It requires proper specification of the prior distributions for parameters in the model.

We study the case where the response variable has missing values and we assume the missing data mechanism is missing at random. If it is the covariates rather than the outcome that have missing values, the PSPP method may still be useful to increase the robustness of inference. This question deserves future study. When the missing data is not missing at random, the balancing property of the propensity score no longer holds. It remains an open question if the PSPP method can be extended to such setting.

We study the case where the missing data is continuous, extensions to other types data is straightforward by using the generalized linear models.

Another interesting question is whether we can apply the PSPP method to the general pattern of missing data. One possible way is to delete some observed values to

derive the monotone pattern but information contained in the deleted cases is lost thus it is not efficient especially when the missing percentage is high. An alternative approach is to combine the PSPP method with iterative methods of computation, such as EM algorithm or the Bayesian methodology. The feasibility of this approach needs more research.

REFERENCES

An, H. (2004). "Robust Likelihood-Based Inference for Multivariate Data with Missing Values". *Ph.D. Thesis, Department of Biostatistics, University of Michigan, Ann Arbor, MI*.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52, 200-203.

Bang H. and Robins J.M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* 61, 962-972.

Barnard, J. and Rubin, D.B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* 86, 948–955.

Casella, G., and George, E. I. (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167-174.

Couper, M.P., Peytchev, A., Little, R.J.A., Strecher, V.J, Rothert, K. (2005). Combining information from multiple modes to reduce nonresponse bias. *Contributed Paper, Joint Statistical Meetings* 2005.

Cochran, W. G. (1965) The planning of observational studies of human populations. *J. of the Royal Stat. Society, Series A*, 128, 234-256

Cochran, W.G. (1968). The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, 24, 205- 213

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-Splines and penalties. *Statistical Science* 11, 89-121.

Eubank, R. L. (1988) *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker.

Firth D. and Bennett K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society. Series B.* 60, 3-21.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

Geman, S. and Geman, D. (1984). ``Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721--741.

Green, P.J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Horvitz, D. G., and Thompson, D. J. (1952). "A Generalization of Sampling without Replacement From a Finite Universe." *Journal of the American Statistical Association*, 47, 663-685.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolnger, R.D. (1996). *SAS System for Mixed Models*, Cary, NC: SAS Institute Inc., 1996. 633 pp.

Little, R.J.A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association,* 78, 596-604.

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

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

Little, R.J.A and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14, 949-968.

Little, R.J.A. and Rubin, D.B. (1999). Comments on "Adjusting for non-ignorable drop-out using semiparametric models" by Scharfstein, D. O., Rotnitsky, A., and Robins J.M. *Journal of the American Statistical Association,* 94, 1130-1132.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.

Little, R.J.A and Zhang G. (2007, In press). Robust likelihood-based analysis of longitudinal data with missing values. Invited Chapter in *Methodology in Longitudinal Surveys,* P. Lynn (Editor). New York: John Wiley.

Lunceford, J.K. and Davidian, M. (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* 23, 2937-2960.

Ngo, L. and Wand, M.P. (2004) Smoothing with mixed model software. *Journal of Statistical Software*, V9, Issue 1.

Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag, New York.

Raghunathan, T.E., Lepkowski, J.M., VanHoewyk, J. and Solenberger, P. (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 27, 85-95.

Robins J.M and Rotnitzky A. (2001). Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer" *Statistica Sinica*, 11, 920-936.

Robins J.M, Rotnitzky A. and Zhao L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89,846-866.

Robins, J. M., Rotnitzky, A. and Zhao, L.P. (1995). "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data." *Journal of the American Statistical Association*, 90, 106-121.

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

Rosenbaum, P.R., and Rubin, D.B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score**.** *Journal of the American Statistical Association*, 79, 516-524

Rotnitzky A, Robins, JM, and Scharfstein DO (1998), "Semiparametric Regression for Repeated Measures Outcomes with Non-ignorable Non-response," *Journal of the American Statistical Association*, 93, 1321-1339.

Rubin, D.B. (1974). Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association*, 69, 467-474

Rubin, D.B. (1976). Inference and missing data. *Biometrika* 63, 581-592.

Rubin, D.B. (1978). Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 20-34

Rubin D.B. and Schenker N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-74.

Rubin, D.B. (1987). *Multiple imputation for non-response in surveys*. New York: John Wiley

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association*, 91, 473-489

Ruppert, D. (2002). Selecting the number of knots for penalized splines, *Journal of Computational and Graphical. Statistics* , 11, 735-757.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression.* Cambridge University Press.

Sarndal, C-E., Swensson, B. and Wretman, J. (2003). *Model Assisted Survey Sampling.* New York, United states, Springer.

SAS (1992). The Mixed Procedure. Chapter 16 in SAS/STAT software: changes and enhancements, Release 6.07, Technical Report P-229, SAS Institute, Inc., Cary, NC.

SAS (1999). SAS Institute Inc. (1999), *SAS Language Reference: Concepts, Version 8,* Cary, NC: SAS Institute Inc, Cary, NC.

Scharfstein D. O., Rotnitzky A. and Robins J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*, 94, 1096-1146.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

Wand, M.P. (2003). Smoothing and mixed models. *Computational Statistics*, 18, 223-249.

Wood, S.N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society. Series B.* 65, 95-114.

Yu M. and Nan B. (2006), A Revisit of Semiparametric Regression Models with Missing Data. *Statistica Sinica*. 16, 1193-1212.

Zeng, D.L. (2001). "Adjusting for dependent censoring using high-dimensional auxiliary information". *Ph.D. Thesis, Department of Statistics, University of Michigan, Ann Arbor, MI*.

Zhang, G., and Little, R.J. (2005). Extensions of the Penalized Spine Propensity Prediction Method of Imputation. Contributed paper, *Joint Statistical Meetings,* 2005.