Biometrics 65, 911–918
September 2009
DOI: 10.1111/j.1541-0420.2008.01155.x

Extensions of the Penalized Spline of Propensity Prediction Method of Imputation

Guangyu Zhang

Department of Epidemiology and Biostatistics, School of Public Health, University of Maryland, College Park, Maryland 20742, U.S.A.
email: guangyuz@umd.edu

and

Roderick Little

Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, Michigan 48109, U.S.A.
email: rlittle@umich.edu

Summary. Little and An (2004, *Statistica Sinica* **14,** 949–968) 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 fitted that includes the spline of the estimated logit 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.

Key words: Missing at random; Penalized spline; Propensity.

1. Introduction

Missing data problems are common in many applications of statistics. In this article, we consider univariate nonresponse, where the missingness is confined to a single variable. Let (Y, $X_1, \ldots, X_p$) denote a $(p + 1)$-dimensional vector of variables with Y subject to missing values and $X_1, \ldots, X_p$ 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.

A simple approach to this missing-data problem is complete case (CC) analysis, 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 a 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 $X_2, \ldots, X_p$. Another approach is to fit a parametric model relating Y to the X's using the CCs, and impute the missing Y's with predictions from this model. For example, one might fit a linear regression model

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

This approach is very efficient and yields consistent estimates if the model assumptions are correct. However, potential sensitivity to model misspecification motivates the study of robust estimation procedures.

Robins, Rotnitzky, and Zhao (1994, 1995) and Rotnitzky, Robins, and Scharfstein (1998) developed a class of doubly robust (DR) augmented orthogonal inverse probability-weighted estimators in missing data models. Specifically, the marginal mean of Y is estimated by adding the mean of the weighted residuals to the predictions from a parametric or semiparametric model. The DR property lies in the fact that if either (a) the prediction model is correctly specified or (b) the model for the probability of response on which the weight is based is correctly specified, then the estimated marginal mean of Y will be consistent. The DR property fails if both the prediction and response model are misspecified. An alternative way to achieve DR is to include the weight as a linear term in the imputation model (Firth and Bennett, 1998; Scharfstein, Rotnitzky, and Robins, 1999; Sarndal, Swensson, and Wretman, 2003; Bang and Robins, 2005). More information on this class of estimators can be found in Robins and Rotnitzky (2001), Lunceford and Davidian (2004), and Yu and Nan (2006).

Semiparametric and nonparametric modeling of the mean structure is another approach to yield robustness, by weakening assumptions about the 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 a truncated polynomial basis

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

where $1, x, \ldots, x^q, (x - \kappa_1)_+^q, \ldots, (x - \kappa_k)_+^q$ is known as the truncated power basis of degree q; $\kappa_1 < \cdots < \kappa_K$ 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 $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_q, \hat{\beta}_{q1}, \ldots, \hat{\beta}_{qK})^T$ is obtained by minimizing

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

where λ is a smoothing parameter and $D = \text{diag}\,(0_{q+1}, 1_K)$. The fitted values are $\hat{y} = X(X^T X + \lambda^{2q} D)^{-1} X^T y$. 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 PSPP method yields an estimate of the marginal mean of the missing variable with a DR property, described below in Section 2. We propose a simplification of the PSPP that does not center the regressors prior to including them in the prediction model.

Little and An (2004) did not consider whether PSPP yields robust estimates for other parameters, such as conditional means or regression coefficients. In Section 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 4 and 5. We apply the proposed methods to an online weight loss study in Section 6, and Section 7 presents concluding remarks.

2. Penalized Spline of Propensity Prediction

Let $(Y, X_1, \ldots, X_p)$ denote a vector of variables with Y subject to missing values and $X_1, \ldots, X_p$ fully observed covariates. The missingness of Y depends only on $X_1, \ldots, X_p$, so the missing data mechanism is missing at random (MAR; 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:

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

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

$$\mu_y = E[(1 - M)Y] + E[M \times E(Y \mid X^*)]. \quad (3)$$

Because 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 PSPP, which is based on the following model:

$$\begin{aligned} &(X_1, \ldots, X_p \mid X^*) \sim N_p((s_1(X^*), \ldots, s_p(X^*)), \Sigma), \\ &(Y \mid X^*, X_1, \ldots, X_p; \beta) \sim N\big(s(X^*) \\ &\quad + g(X^*, X_1^*, X_2^*, \ldots X_p^*; \beta), \sigma^2\big), \end{aligned} \quad (4)$$

where $N_k(\mu, \Sigma)$ denotes the k-variate normal distribution with mean μ and covariance matrix Σ, $s_j(Y^*) = E(X_j \mid Y^*)$; $j = 1, \ldots, p$, is a spline for the regression of X_j on X^* of the form (equation 1); $X_j^* = X_j - s_j(X^*)$ is the residual of the spline model and represents the part in X_j not explained by the propensity score; $s(X^*)$ is a spline of Y on X^* of the form (equation 1); and g is a parametric function indexed by unknown parameter β with $g(X^*, 0, \ldots, 0; \beta) = 0$ for all β. In practice, variables included in the g-function should be predictive of the outcome. In specifying the g-function in equation (4), care is needed to avoid multicollinearity, because X^* is itself a function of $(X_1^*, X_2^*, \ldots, X_p^*)$; a simple way of doing this is to leave out of the g-function the covariate that is most highly correlated with X^*. Also note that in practice X^* involves unknown parameters, which need to be estimated from the data, yielding an estimated logit propensity $\hat{X}^*$.

The predicted mean of Y from model (4) has the following DR property (Little and An, 2004). Let $\hat{\mu}_y$ be the prediction estimator for equation (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 $(X^*, X_1, \ldots, X_p)$ in model (4) is correctly specified, or (b1) the logit propensity X^* is correctly specified, and (b2) $E(X_j \mid X^*) = s_j(X^*)$ for $j = 1, \ldots, p$ and $E(Y \mid X^*) = s(X^*)$. The robustness feature derives from the fact that the regression function g does not have to be correctly specified.

The covariates $X_1^*, \ldots, X_p^*$ from model (4) are centered by regressing $X_1, \ldots, X_p$ on splines of X^* and taking residuals. We now propose a simpler method that adds $X_1, \ldots, X_p$ directly to the regression without centering and we show this method also has the DR property:

THEOREM 1. *The PSPP method based on model* (4) *can be simplified as follows:*

$$(Y \mid X^*, X_1, \ldots, X_p; \beta) \sim N\big(s(X^*) + g(X^*, X_1, \ldots X_p; \beta), \sigma^2\big), \quad (5)$$

that is, the covariates $X_1, \ldots, X_p$ enter the parametric function g without centering. Let $\hat{\mu}_y$ be the prediction estimator for equation (3) based on model (5), and assume MAR, then $\hat{\mu}_y$ has the same DR property as that derived from model (4) (see Web Appendix A for proof). For this reason, we focus on the uncentered version of the PSPP method for the remainder of the article.

The first step of fitting a PSPP model (5) estimates the parameters in the logit propensity score X^*, for example by a logistic regression model of M on $X_1, \ldots, X_p$, yielding

the estimated logit propensity score $\hat{X}^*$; in the second step, the regression of Y on $\hat{X}^*$ is fit as a spline model with the other covariates included in the model parametrically in the g-function. We use equally spaced fixed knots in this article when fitting model (5). Let κ_0 = the minimal value of the data and d = (maximal value – minimal value)/$(K+1)$, we derive the knots by the following formula: $\kappa_i = \kappa_{i-1} + d$, $i = 1, \ldots, K$, where K is the total number of knots. We implement the spline model using the `Proc Mixed` procedure in `SAS` with a truncated linear basis and treat $(\hat{X}^* - \kappa_1)_+, \ldots, (\hat{X}^* - \kappa_K)_+$ as random effects and 1, $\hat{X}^*$ and covariates in the g-function as fixed effects. More information on implementation of penalized spline models can be found in SAS (1992), Ngo and Wand (2004), and Ruppert et al. (2003).

3. PSPP is Not Doubly Robust for Subgroup Means

The DR property of PSPP for estimating the marginal mean of Y does not apply 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

$$\begin{aligned} &Y \mid X_1, X_2 \sim N(\mu(X_1, X_2), 1), \\ &\mu(X_1, X_2) = I[X_1 = 1] + 3 \times I[X_1 = 2] \\ &\quad + 5 \times I[X_1 = 3] + 10X_2, \end{aligned}$$

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

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

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 $[s(X^*_{\text{correct}})]$. The logit propensity score X^*_{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 $[s(X^*_{\text{correct}}) + X_1]$. This model correctly specifies the mean of Y given the covariates, because 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 $[s(X^*_{\text{wrong}})]$.
(f) Model (e) with X_1 included, namely $[s(X^*_{\text{wrong}}) + X_1]$. This model correctly specifies the mean of Y given the covariates, because it includes the main effects of X_1 and X_2.

We choose 20 equally spaced fixed knots with a truncated linear basis for the PSPP methods. 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), obtained from the original simulated data without deleting any values, and estimates based on the CCs. Empirical bias, empirical standard error (SE), and root mean square error (RMSE) are summarized in Table 1. CC analysis yields estimates with large biases and RMSEs. The correctly specified ANCOVA model (a) yields unbiased estimates close to the BD estimates. The incorrectly 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

Table 1

Example 1: Empirical bias, empirical SE, and RMSE for various estimates of (A) marginal mean of Y and (B) conditional mean of Y given X_1. Entries are multiplied by 100.

	(A) Marginal mean			(B) Conditional mean of Y given X_1								
	$E(Y)$			$E(Y \mid X_1 = 1)$			$E(Y \mid X_1 = 2)$			$E(Y \mid X_1 = 3)$		
Methods	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RSE
BD	0	45	35	−3	63	51	6	82	67	−1	98	77
CC Analysis	368	58	368	328	78	328	505	118	505	416	124	416
(a) Correct ANCOVA $[X_1, X_2]$	0	45	36	−3	64	51	7	83	68	−1	98	78
(b) Wrong ANCOVA $[X_1]$	398	58	398	328	78	328	505	118	505	416	124	416
(c) PSPP $[s(X^*_{\text{correct}})]$	−2	50	40	213	69	214	−271	111	271	−139	141	162
(d) PSPP $[s(X^*_{\text{correct}}) + X_1]$	0	45	36	−3	64	51	7	84	69	−1	99	78
(e) PSPP $[s(X^*_{\text{wrong}})]$	−20	47	41	43	65	63	−44	84	76	−145	107	154
(f) PSPP $[s(X^*_{\text{wrong}}) + X_1]$	0	45	36	−3	64	52	7	84	68	−1	99	78

Bias: empirical bias, the average of the deviation of estimates from the true value over the 500 replications.
SE: empirical standard error, the standard deviation of the estimates over the 500 replications.
RMSE: root mean square error, the average of the square root of the squared deviation of the estimates from the true value over the 500 replications.

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 SEs 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, because 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:

$$\begin{aligned} Y \mid X_1, X_2 &\sim N(\mu(X_1, X_2), 1), \\ \mu(X_1, X_2) &= I[X_1 = 1] + 3 \times I[X_1 = 2] \\ &\quad + 5 \times I[X_1 = 3] + 10X_2 + X_2^2 \\ &\quad - 1 + 4 \times I[X_1 = 1] \times X_2 - 10 \times I[X_1 = 2] \times X_2. \end{aligned}$$

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

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

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 $[s(X^*)]$. The logit propensity score X^* is modeled as an additive function of X_1, X_2, and X_2^2 and hence is correctly specified.
(c) Model (b) with X_1 included, that is, $[s(X^*) + X_1]$.
(d) Model (b) with X_2 and X_2^2 included, that is, $[s(X^*) + X_2 + X_2^2]$.

We compute empirical bias and RMSE as for Example 1. In addition, we apply the above methods to 200 bootstrap samples for each dataset and derive bootstrap SEs $\hat{\sigma}_{\text{BS}}$ and 95% confidence intervals for the mean of the form $\hat{\mu} \pm 1.96\hat{\sigma}_{\text{BS}}$, based on a normal approximation for the bootstrap distributions. The coverage rate of this interval is estimated as the percentage of the 500 samples with the 95% confidence intervals covering the true value. The correctly specified ANCOVA model yields estimates with small empirical bias and RMSE and the coverage rate close to the 95% nominal level (Table 2). CC analysis and the wrongly specified ANCOVA model yield biased estimates and poor confidence coverage. 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, because the estimates of the conditional means are still biased.

In the second example, the PSPP method $[s(X^*) + X_1]$ assumes that for different levels of X_1, the splines of Y on X^* have the same shape; because 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 estimated logit propensity score and X_1 into the model, yielding a stratified PSPP method discussed in the next section.

4. Stratified Penalized Spline Propensity Prediction for Subgroup Means

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

$$\begin{aligned} &(Y \mid X^*, X_1, \ldots, X_p; \beta) \\ &\quad \sim N\left(\sum_{c=1}^{C} I_c s_c(X^*) + g(X^*, X_1, X_2, \ldots, X_p; \beta), \sigma^2\right), \quad (6) \end{aligned}$$

where g is a parametric function indexed by unknown parameter β as before; $I_c s_c(X^*) = I_c(\gamma_{0c} + \sum_{j=1}^{q} \gamma_{jc}(X^*)^j + \sum_{k=1}^{K} \gamma_{qkc}(X^* - \kappa_k)_+^q)$ is the fitted curve for the cth level of X_1. Within each level of X_1,

$$\begin{aligned} E(Y \mid X^*, X_1 = c, X_2, \ldots, X_p; \beta) &= s_c(X^*) \\ &\quad + g(X^*, X_1 = c, X_2, \ldots, X_p; \beta). \end{aligned}$$

Note that this method is not the same as applying PSPP within strata defined by X_1, because 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 (see Web Appendix B for proof).

Example 2 continued. Row (e) in Table 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, the RMSE for these parameters is only slightly larger than for BD and the coverage rate is close to the 95% nominal level. Thus the stratified PSPP has fixed the bias for the subgroup means in the PSPP method. 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.

5. A Bivariate PSPP 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 logit 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:

$$\begin{aligned} &(Y \mid X^*, X_1, X_2, \ldots, X_p; \beta) \\ &\quad \sim N\big(s(X^*, X_2) + g(X^*, X_1, X_2, \ldots, X_p; \beta), \sigma^2\big), \quad (7) \end{aligned}$$

Table 2

Example 2: Empirical bias, RMSE, and coverage rate (Cov) for various estimates of (A) marginal mean of Y, (B) conditional mean of Y given X_1, and intercept and slopes for regression of Y on X_2, X_2^2. Entries are multiplied by 100.

Methods	Overall mean		
	Bias	RMSE	Cov
(A) Marginal mean of Y			
BD	−1	30	93
CC	199	199	7
(a) Correct Model $[X_1 + X_2 + X_1 \times X_2 + X_2^2]$	−1	30	93
(b) PSPP $[s(X^*)]$	−3	34	95
(c) PSPP $[s(X^*) + X_1]$	−3	33	94
(d) PSPP $([s(X^*) + X_2 + X_2^2])$	−3	33	95
(e) Stratified PSPP $(Y = \sum I_c s_c(X^*))$	−1	31	94
(f) Bivariate PSPP $(Y = s(X^*, X_2))$	−1	30	94
(g) Stratified bivariate PSPP $(Y = \sum I_c s_c(X^*) + s(X^*, X_2))$	−2	30	94

	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		
Methods	Bias	RMSE	Cov	Bias	RMSE	Cov	Bias	RMSE	Cov	Bias	RMSE	Cov	Bias	RMSE	Cov	Bias	RMSE	Cov
(B) Conditional mean of Y given X_1, and intercept and slopes for regression of Y on X_2, X_2^2																		
BD	−1	53	93	0	8	94	0	56	95	0	20	95	−2	26	95	1	25	95
CC Analysis	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	53	93	0	10	94	1	58	95	0	19	95	−2	26	95	1	24	95
(b) PSPP $[s(X^*)]$	112	116	66	−91	93	43	−154	156	53	−94	94	20	−163	163	1	93	93	20
(c) PSPP $[s(X^*) + X_1]$	47	69	89	−121	122	28	53	90	94	−100	100	14	−139	139	17	98	99	19
(d) PSPP $([s(X^*) + X_2 + X_2^2])$	50	70	89	−1	47	97	−137	146	72	37	43	78	17	40	92	−39	48	78
(e) Stratified PSPP $(Y = \sum I_c s_c(X^*))$	0	54	95	−4	13	92	4	60	96	−81	81	41	−86	87	60	81	81	43
(f) Bivariate PSPP $(Y = s(X^*, X_2))$	10	54	94	18	30	98	−60	87	93	2	22	97	2	29	96	−3	26	96
(g) Stratified bivariate PSPP $(Y = \sum I_c s_c(X^*) + s(X^*, X_2))$	−2	53	93	−7	16	98	5	59	96	4	23	96	4	30	95	−5	29	95

where g is a parametric function and $s(X^*, X_2)$ is a bivariate P-spline of X^* and X_2. Estimation of the smoothing function $s(X^*, X_2)$ 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 pairwise products of the basis functions. The resulting bivariate basis is called the tensor product basis (Ruppert et al., 2003). In this article, we choose five equally spaced knots for each variable when fitting the bivariate splines using a tensor product basis.

Example 2 continued. Row (f) in Table 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 two-way interactions between the logit propensity score and X_1, and the logit propensity score and X_2. This yields the following combination of the stratified PSPP and bivariate PSPP models:

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

where $I_c s_c(X^*)$ and $s(X^*, X_2)$ 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, row (g)).

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 4000 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 months, 623 subjects in the treatment group and 611 subjects in the control group responded to the second survey. 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 difference 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 difference 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 score.

We assume MAR in our analysis, and estimate the propensity score by a logistic regression, with the inclusive criterion of retaining all variables with P-values less than 0.20. The MAR assumption is supported by the fact that we have rich baseline information for characterizing the respondents and nonrespondents, although it would be advisable to conduct a sensitivity analysis to assess the impact of violations from MAR, for example treating drop-outs as treatment failures.

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

(a) PSPP method with null g-function, denoted as $[s(X^*)]$, where X^* is the logit propensity score defined in Section 2.
(b) Model (a) with treatment as a covariate, denoted as $[s(X^*) + \text{treatment}]$.
(c) Model (b) with baseline covariates, denoted as $[s(X^*) + \text{treatment} + g(\text{baseline vars})]$.
(d) Stratified PSPP method with null g-function, denoted as $[\sum_{c=1}^{2} I_c s_c(X^*)]$.
(e) Model (d) with baseline covariates, denoted as $\sum_{c=1}^{2} I_c s_c(X^*) + g(\text{baselinevars})]$.

Results are summarized in Table 3. SEs and the corresponding confidence intervals are obtained from 200 bootstrap samples. The treatment group has a larger reduction of BMI after 3 months (−0.91 (0.09)) compared to the control group (−0.45 (0.10)) based on the CC analysis. The stratified PSPP method (models (d) and (e)) and the PSPP method with the treatment as a covariate (models (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 (models (b)–(e)). On the other hand, the PSPP method without treatment as a covariate (model (a)) does not show 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 (models (c) and (e)).

In this study the stratified PSPP method (models (d) and (e)) and the PSPP method with treatment as a covariate (models (b) and (c)) yield similar results, a reflection of the fact that the spline curves are almost parallel for the treatment and control group. The slight departure from parallelism is shown by the relatively larger differences of BMI reduction between the two groups with the stratified PSPP method. In practice, we recommend the stratified PSPP method for subgroup means because it does not constrain the spline curves to be parallel across the groups, and it retains the DR property.

Table 3
BMI reduction within groups estimated by various methods

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

SE and 95% CI are based on 200 bootstrap samples.

7. Discussion

We have shown that the PSPP method yields an estimate of the marginal mean of Y with a DR 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 DR 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 logit 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 $s(X^*, X_2)$ 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, 2003). It can be fit using the `tpspline` procedure from `SAS` (SAS, 1992; Wand, 2003; Ngo and Wand, 2004). 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 `tpspline` procedure are comparable to those from a P-spline with a tensor product basis.

More generally, a PSPP method that yields DR estimates of the conditional mean of Y given a subset of the covariates $(X_1, \ldots, X_s)$, $s < p$, requires inclusion of the interactions between the logit propensity score and $(X_1, \ldots, X_s)$; 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 $(X_1, \ldots, X_p)$. 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 logit propensity score, a method that has a DR property. The DR property can also be achieved by modeling the relationship parametrically using the augmented orthogonal inverse probability-weighted estimators described in the introduction. We are currently conducting extensive simulations to compare the performance of these methods with the PSPP method. In general, we find that the PSPP method yields estimates with smaller RMSE and comparable or better confidence coverages. These results will be reported in a future paper.

To be reliable, PSPP depends on some degree of overlap in the distributions of response propensities for respondent and nonrespondents. In the extreme case where these distributions do not overlap at all, the spline on the propensity is being fitted to the range of propensities of respondents, and then interpolated outside this range to the propensities for nonrespondents. Such an interpolation is highly questionable, as in any application of regression where the predictions are to values of independent variables not seen in the data. No method can be expected to work well in such situations, including the inverse-probability weighted methods, for which this situation leads to extreme or undefined weights. PSPP may have the virtue of yielding more conservative (that is larger) estimates of uncertainty than parametric models in this setting, but sensitivity analysis may be a better option.

We apply the PSPP to data with univariate nonresponse in this article. Extensions to monotone and general patterns of missing data have been explored. We propose a stepwise PSPP approach that preserves the DR property for the missing data in a monotone pattern and the part with least missing information is imputed first. More information for the stepwise procedure can be found at Little and Zhang (2008). For a general pattern of missing data, the sequential imputation methods of Raghunathan et al. (2001) can be extended to provide PSPP imputations that condition on the spline of the logit propensity that each variable is missing. An (2004) discusses these extensions in detail.

8. Supplementary Materials

Web Appendices referenced in Sections 2 and 4 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

This research is supported by CECCR Center grant P50 CA101451. We thank two referees for useful comments on an earlier draft, and Trivellore Raghunathan for assistance with Theorem 1.

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.

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

Couper, M. P., Peytchev, A., Little, R. J. A., Strecher, V. J., and Rothert, K. (2005). Combining information from multiple modes to reduce nonresponse bias. Contributed Paper in *Proceedings of Joint Statistical Meetings,* Survey Research Methods Section. Alexandria, Virginia: American Statistical Association, 2910–2917.

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

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

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

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. and Zhang G. (2008, in press). Robust likelihood-based analysis of longitudinal data with missing values. Invited Chapter. In *Methodology in Longitudinal Surveys*, P. Lynn (ed). 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* **9**, 1–54.

Raghunathan, T. E., Lebkowski, 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.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association* **93,** 1321–1339.

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

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

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

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

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.

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.

Received March 2007. *Revised June* 2008.
Accepted June 2008.