# Extensions of the Penalized Spline Propensity Prediction Method of Imputation

Guangyu Zhang and Roderick Little
Department of Biostatistics
University of Michigan

## 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 consistent, but the PSPP method does not yield consistent estimates of other parameters, like conditional means or regression coefficients. We discuss properties of a simplified version of PSPP that does not center the regressors prior to including them in the prediction model. We then extend 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.

## 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 $(Y, X_1, ..., X_{p-1})$ denote a $p$ dimensional vector of variables with $Y$ subject to missing values and $X_1, ..., X_{p-1}$ fully observed covariates. We consider the problem of estimating the mean of $Y$, and the conditional mean of $Y$ in subclasses defined by a categorical variable, and the regression coefficient of $Y$ on a continuous variable.

Many statistical methods have been proposed for this problem. 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 the conditional mean of $Y$ given a covariate $X_1$, if the missing-data mechanism is such that missingness depends on $X_1$, but does not depend on $Y$ or $X_2, ..., X_{p-1}$. Another approach is to impute predictions based on a parametric model $Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} + \varepsilon_i$, where $\varepsilon_i$ is the error term with $\varepsilon_i \sim N(0, \sigma^2)$. One can estimate $(\beta_0, ..., \beta_{p-1})$ 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 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 a single covariate $X$, imputations can be based on the penalized spline $y_i = s(x_i) + \varepsilon_i$ with truncated polynomial basis

$$s(x) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{k=1}^{K} \beta_{pk} (x - \kappa_k)_+^p \qquad (1)$$

where 1, $x, ..., x^p, (x-\kappa_1)_+^p, ..., (x-\kappa_k)_+^p$ is known as the truncated power basis of degree $p$; $\kappa_1 < .... < \kappa_K$ 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 $\hat{\beta} = (\hat{\beta}_0, ..., \hat{\beta}_p, \hat{\beta}_{p1}, ..., \hat{\beta}_{pK})^T$ is obtained by minimizing

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

where $\lambda$ is a smoothing parameter and $D$ = $diag(0_{p+1}, 1_K)$ . The fitted values are $\hat{y}$ = $X(X^T X + \lambda^{2p} D)^{-1} X^T y$. 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 non-parametric form for predictions when $K$ is large.

When there are several covariates in the regression model, such models are subject to the curse of dimensionality, which relates to the difficulty of fitting nonparametric regression functions when the regressor space has high dimension. Little and An (2004) proposed Penalized Spline Propensity Prediction (PSPP), which addresses the curse of dimensionality 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 property, described below in section 2. We also discuss properties of a simplified version of 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 robust estimates of these parameters. This motivates robust extensions of the PSPP method for estimating subgroup means and regression coefficients. These proposed extensions are described in sections 4 and 5. Section 6 presents concluding remarks.

## 2. The Penalized Spline Propensity Prediction (PSPP) Method

Let $(Y, X_1, ..., X_{p-1})$ denote a $p$ dimensional vector of variables with $Y$ subject to missing values and $X_1$, … $X_{p-1}$ fully observed covariates. The missingness of $Y$ depends only on $X_1, ..., X_{p-1}$, 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:

$$Y^*\text{=logit }(\Pr(M = 0 \mid X_1, ..., X_{p-1})) \qquad (2)$$

The key property of the propensity score is that, conditioning on the propensity score and assuming MAR, missingness of $Y$ does not depend on the covariates $X_1$, …, $X_{p-1}$ (Rosenbaum and Rubin, 1983). Thus, the mean of $Y$ can be written as

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

This motivates the Penalized Spline Propensity Prediction Method (PSPP), which is based on the following model:

$$(X_2, ..., X_{p-1} \mid Y^*) \sim N_{p-2}((s_2(Y^*), ..., s_{p-1}(Y^*)), \Sigma)$$

$$(Y \mid Y^*, X_2, ..., X_{p-1}; \beta) \qquad (4)$$

$$\sim N(s_p(Y^*) + g(Y^*, X_2^*, ... X_{p-1}^*; \beta), \sigma^2)$$

where $s_j(Y^*) = E(X_j \mid Y^*)$, $j = 2, ..., p-1$, is a spline for the regression of $X_j$ on $Y^*$ of the form (1); $X_j{}^*$ = $X_j - s(Y^*)$ is the residual of the spline model and represents the part in $X_j$ not explained by the propensity score; $N_{p-2}((s_2(Y^*), ..., s_{p-1}(Y^*)), \Sigma)$ is a multivariate normal distribution with mean $(s_2(Y^*), ..., s_{p-1}(Y^*))$ and variance covariance structure $\sum$; $s_p(Y^*)$ is a spline of $Y$ on $Y^*$ of the form (1) and $g$ is a parametric function indexed by unknown parameter $\beta$ with $g(Y^*, 0, ..., 0; \beta) = 0$ for all $\beta$. The variable $X_1$ is not included in the $g$ function to prevent multicollinearity. The first step of fitting a PSPP model estimates the propensity score, for example by a logistic regression model of $M$ on $X_1, ..., X_{p-1}$; in the second step, the regression of $Y$ on $Y^*$ is fit as a spline model with the other covariates included in the model parametrically in the $g$ function.

PSPP has a double robustness property for predicting the mean of $Y$ based on model (4), formalized in the following theorem:

**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 $\mu_y$ if either (a) the mean of $Y$ given $(Y^*, X_2, ..., X_{p-1})$ in model (4) is correctly specified, or (b1) the propensity $Y^*$ is correctly specified, and (b2) $E(X_j \mid Y^*) = s_j(Y^*)$ for $j = 2, ..., p-1$ and $E(Y \mid Y^*) = s_p(Y^*)$. The robustness feature derives from the fact that the regression function $g$ does not have to be correctly specified.

Little and An demonstrate the robustness property with simulations in which the PSPP method is compared with several other methods. They show the PSPP method yield relatively robust estimates of the population mean under different mean and propensity structures.

The above theorem requires that the covariates $X_2^*$, …, $X_{p-1}^*$ in the PSPP method are centered by regressing $X_2, ..., X_{p-1}$ on splines of $Y^*$ and taking residuals. We now show that this centering is not needed, and covariates can be added directly into the regression, simplifying the method considerably. This property is elucidated in the following theorem:

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

$$(Y \mid Y^*, X_2, ..., X_{p-1}; \beta)$$

$$\sim N(s(Y^*) + g(Y^*, X_2, ... X_{p-1}; \beta), \sigma^2), \qquad (5)$$

that is, the covariates $X_2, ..., X_{p-1}$ 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 property as that derived from model (4) (see appendix for proof).

## 3. The PSPP method for the subgroup means of $Y$ conditional on a categorical covariate.

The PSPP method is robust for estimating the marginal mean of $Y$. An interesting question is whether it also yields robust estimates of other parameters, such as subgroup means. To address this issue, we consider the PSPP method for a subgroup mean when (a) the propensity of response is correctly specified and (b) the regression model for $Y$ given the covariates is incorrectly specified.

**Example 1. Robustness studies of PSPP for estimating conditional means: effect of failing to condition on a subgroup variable.** We simulate 100 datasets with 1000 subjects, with two covariates $X_1, X_2$ and a 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

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

We create missing values of $Y$ from a model for the propensity to respond:

$$\text{logit}\,(P(M = 0 \mid X_1, X_2)) = X_2 .$$

We imputed 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 $Y$ given $X_1$ and $X_2$, namely $[X_1]$.

(c) The PSPP Method with null the $g$ function, which we denote $[s(Y^*)]$.

(d) The PSPP Method with $X_1$ included, namely $[s(Y^*) + X_1]$. This model correctly specifies the mean of $Y$ given the covariates.

For all the penalized spline methods in this paper, we choose 20 equally spaced fixed knots and a truncated linear basis.

Table 1. Marginal mean of $Y$ (simulation 1)

| Methods | Bias | STD | RMSE |
|---|---|---|---|
| BD | 5 | 39 | 31 |
| CC | 420 | 43 | 420 |
| (a)Correct ANCOVA $[X_1, X_2]$ | 6 | 38 | 31 |
| (b)Wrong ANCOVA $[X_1]$ | 419 | 43 | 419 |
| (c)PSPP $[s(Y^*)]$ | 6 | 39 | 31 |
| (d)PSPP $[s(Y^*) + X_1]$ | 6 | 38 | 31 |

Table 2. Conditional mean of $Y$ given $X_1$ (Simulation 1)

| Methods | $X_1 = 1$ | | | $X_1 = 2$ | | | $X_1 = 3$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Bias | STD | RMSE | Bias | STD | RMSE | Bias | STD | RMSE |
| BD | 3 | 52 | 41 | 10 | 62 | 51 | 0 | 75 | 60 |
| CC | 417 | 58 | 417 | 422 | 73 | 422 | 415 | 90 | 415 |
| (a)Correct ANCOVA $[X_1, X_2]$ | 3 | 52 | 41 | 10 | 62 | 52 | 0 | 75 | 60 |
| (b)Wrong ANCOVA $[X_1]$ | 417 | 58 | 417 | 422 | 73 | 422 | 415 | 90 | 415 |
| (c)PSPP $[s(Y^*)]$ | 74 | 51 | 75 | -20 | 64 | 54 | -127 | 78 | 130 |
| (d)PSPP $[s(Y^*) + X_1]$ | 3 | 52 | 41 | 10 | 62 | 52 | 1 | 75 | 59 |

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 analyze estimates from the data before deletion (BD) and estimates based on the complete cases (CC). Bias, empirical standard deviation (STD) and root mean square error (RMSE) are summarized in Tables 1 and 2. CC analysis yields estimates with largest 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 yields unbiased marginal mean estimate but biased conditional mean estimates; adding $X_1$ in the PSPP method yields unbiased estimates of the marginal mean of $Y$ and the conditional means of $Y$ given $X_1$, with biases, STDs and RMSEs very close to those of BD.

**Example 2. Robustness studies of PSPP for estimating conditional means: impact of misspecification of the mean structure.** In the second simulation study, we generate $X_1$ and $X_2$ as in Example 1; but the mean of $Y$ given $X_1$ and $X_2$ is simulated as:

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

The missingness of $Y$ depends on both $X_1$ and $X_2$, with

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

Again, we simulate 100 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) An incorrectly specified regression model for $Y$, namely $[X_1 + X_2 + X_1 \times X_2]$.
(c) The PSPP model without the $g$ function, namely $[s(Y^*)]$.
(d) Model (c) with $X_1$ included, that is, $[s(Y^*) + X_1]$.
(e) Model (c) with the $g$ function with centered $X_2$, namely, $Y = s(Y^*) + g(X_2^*)$.
(f) Model (c) with the $g$ function with uncentered $X_2$, namely, $Y = s(Y^*) + g(X_2)$.

The correctly-specified ANCOVA model yields unbiased estimates that are close to those of BD (Table 3 and 4). CC analysis and the wrongly specified ANCOVA model yield biased estimates. The PSPP methods yield consistent estimates for the marginal mean of $Y$ but not the conditional means of $Y$ given $X_1$. Unlike Example 1, adding $X_1$ or the $g$ function does not correct the bias in estimating the conditional means. Biases and RMSEs from PSPP remain large compared to those of BD.

For the first example, adding $X_1$ to the PSPP model correctly specifies the mean of $Y$ given $X_1$ and yields unbiased subgroup means of $Y$ given $X_1$; but for the second example, adding $X_1$ into the PSPP methods but misspecifying the regression on $X_2$ yields biased subgroup mean estimates. Adding $X_1$ into the PSPP model implies an underlying assumption: for different levels of $X_1$, the spline curves follow the same trend. That is not a correct assumption for the second example. We need a model that relaxes that assumption. One solution is to include the interaction of propensity score and $X_1$ into the model, yielding a stratified PSPP method.

## 4. Stratified Penalized Spline Propensity Prediction for subgroup means

Let $I_c = 1$ if $X_1 = c$; $I_c = 0$ if $X_1 \neq c$, $c = 1, ..., C$, and form the propensity in each category of $X_1$: $Y^{*c} = Y^* \times I_c$, $c = 1, ..., C$. The stratified PSPP method is based on the following model:

$$(Y \mid Y^{*1}, ..., Y^{*C}, X_2, ..., X_{p-1}; \beta) \sim N(\sum_{c=1}^{C} s_{pc}(Y^{*c}) + g(Y^*, X_2, ..., X_{p-1}; \beta), \sigma^2), \tag{6}$$

where $s_{pc}(Y^{*c})$, $c = 1, ..., C$, is a spline for the regression of $Y$ on $Y^{*c}$; $g$ is a parametric function indexed by unknown parameter $\beta$. Within each category of $X_1$,

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

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 consistency property.

Table 3. Marginal mean of $Y$ (Simulation 2)

| Methods | Bias | STD | RMSE |
|---|---|---|---|
| BD | 5 | 33 | 27 |
| CC | 233 | 53 | 233 |
| (a)Correct ANCOVA [ $X_1, X_2, X_1 * X_2, X_2^2$ ] | 5 | 32 | 27 |
| (b)Wrong ANCOVA [ $X_1, X_2, X_1 * X_2$ ] | 17 | 33 | 30 |
| (c)PSPP [ $s(Y^*)$ ] | 0 | 42 | 35 |
| (d)PSPP [ $s(Y^*) + X_1$ ] | 1 | 43 | 35 |
| (e)PSPP [ $s(Y^*) + g(X_2^*)$ ] | 2 | 41 | 33 |
| (f) PSPP [ $s(Y^*) + g(X_2)$ ] | 1 | 41 | 33 |
| Stratified PSPP [ $\sum(s_c(Y^{*c}))$ ] | 7 | 41 | 34 |

**<u>Example 2 continued</u>**
We apply stratified PSPP to the data in the second simulation study, and the results indicate that the method yields consistent estimates of the marginal mean of $Y$ and the subgroup means of $Y$ given $X_1$ (Table 3-4). For the marginal mean estimation, bias, STD and RMSE from stratified PSPP are close to those of BD analysis; for the conditional means, stratified PSPP yields estimates with smaller biases and RMSEs than those of PSPP.

## 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 $X_1$, based on a regression model for $Y$ given $X_1$. To yield consistent parameter estimates for the regression coefficients, we again include the interaction of propensity score and $X_1$ in the model for predicting the missing values of $Y$. Specifically, we propose the following bivariate PSPP method, based on the model:

$$(Y \mid Y^*, X_1, X_2, ..., X_{p-1}; \beta) \sim N(s(Y^*, X_1) + g(Y^*, X_2, X_3 ..., X_{p-1}; \beta), \sigma^2), \tag{7}$$

where $g$ is a parametric function; $s(Y^*, X_1)$ is a bivariate smoothing spline of $Y^*$ and $X_1$.

**Example 2 Continued**

We apply the bivariate PSPP method to the second simulation study in the section 3. We switch $X_1$ and $X_2$, which means $X_1$ is a standard normal and $X_2$ is a multinomial variable. The mean of $Y$ given $X_1$ and $X_2$ is simulated as:

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

The missingness of $Y$ depends on both $X_1$ and $X_2$, with

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

We impute the missing $Y$ by the following methods:

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

(b) A wrongly specified regression model of $Y$ given $X_1$ and $X_2$, namely, $[X_1 + X_2 + X_1 \times X_2]$.

(c) A spline prediction model for $Y$ given $X_1$, namely, $[s(X_1)]$.

(d) Model (c) with $X_2$ included as a predictor, namely, $[s(X_1) + X_2]$.

(e) The PSPP method without the $g$ function, that is $[s(Y^*)]$.

(f) Model (e) with $X_2$ included, that is, $[s(Y^*) + X_2]$.

(g) Model (e) with a $g$ function of centered $X_1$, namely, $[s(Y^*) + g(X_1^*)]$.

(h) Model (e) with a $g$ function of uncentered $X_1$, namely, $[s(Y^*) + g(X_1)]$.

(i) Model (e) with a smoothing spline of $X_1$, namely, $[s(Y^*) + s(X_1)]$.

(j) The bivariate PSPP , namely, $[s(Y^*, X_1)]$.

We fit a regression model of $Y$ given $X_1$ on the imputed datasets derived from the methods listed above, as well as the before deletion datasets and the complete cases. We are interested in the estimates of the intercept, regression coefficients of $X_1$ and $X_1^2$. The correctly specified ANCOVA model yields unbiased regression coefficients with biases and RMSEs very close to the before deletion analysis (Table 5). CC analysis and the wrongly specified ANCOVA model yield biased parameter estimates, with relatively large biases. A spline of $X_1$ does not help in estimating the conditional mean of $Y$ given $X_1$. The spline regression models (c) and (d) and the PSPP methods (e), (f), (g), (h) and the PSPP model with a spline of $X_1$ (i) do not yield consistent estimates of the conditional mean of $Y$ given $X_1$; the parameters have larger biases compared to BD analysis. On the other hand, the bivariate PSPP method does yield consistent parameter estimates of the regression of $Y$ given $X_1$; empirical biases and RMSEs are similar to those of BD analysis.

## 6. Conclusion

We have shown that that the PSPP method yields consistent 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 does not have 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 the interaction of the propensity score and the covariate of interest in the prediction model. Simulations are presented to support the robustness properties of these extensions.

These methods extend in obvious ways to inference for the conditional mean of $Y$ given a subset of the covariates $(X_1, ..., X_s)$, $s < p-1$ although the curse of dimensionality comes 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, $(X_1, ..., X_{p-1})$. 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.

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

Table 4. Conditional mean of $Y$ given categorical $X_1$

| Methods | $X_1 = 1$ | | | $X_1 = 2$ | | | $X_1 = 3$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Bias | STD | RMSE | Bias | STD | RMSE | Bias | STD | RMSE |
| BD | 8 | 55 | 44 | 2 | 12 | 10 | 3 | 69 | 52 |
| CC | 296 | 76 | 296 | 45 | 22 | 45 | 410 | 104 | 410 |
| (a)Correct ANCOVA [ $X_1, X_2, X_1 * X_2, X_2^2$ ] | 8 | 55 | 44 | 3 | 14 | 11 | 3 | 70 | 54 |
| (b)Wrong ANCOVA [ $X_1, X_2, X_1 * X_2$ ] | 18 | 55 | 46 | 19 | 33 | 30 | 14 | 72 | 56 |
| (c)PSPP [ $s(Y^*)$ ] | 99 | 62 | 101 | -107 | 48 | 108 | -81 | 83 | 97 |
| (d)PSPP [ $s(Y^*) + X_1$ ] | 57 | 64 | 71 | -172 | 55 | 172 | 123 | 88 | 127 |
| (e)PSPP [ $s(Y^*) + g(X_2^*)$ ] | 29 | 61 | 56 | 24 | 47 | 43 | -100 | 85 | 111 |
| (f) PSPP [ $s(Y^*) + g(X_2)$ ] | 30 | 61 | 56 | 21 | 47 | 42 | -101 | 84 | 111 |
| (g) Stratified PSPP [ $\sum(s_c(Y^{*c}))$ ] | 16 | 60 | 51 | -1 | 37 | 27 | 0 | 85 | 68 |

Table 5. Regression of $Y$ given continuous $X_1$

| Methods | Intercept | | | $X_1$ | | | $X_1^2$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Bias | STD | RMSE | Bias | STD | RMSE | Bias | STD | RMSE |
| BD | 0 | 25 | 20 | -3 | 36 | 29 | -2 | 30 | 24 |
| CC | -49 | 31 | 50 | 124 | 51 | 124 | -35 | 37 | 42 |
| (a)Correct ANCOVA [ $X_1, X_2, X_1 * X_2, X_1^2$ ] | 1 | 24 | 20 | -2 | 36 | 30 | -2 | 30 | 24 |
| (b)Wrong ANCOVA [ $X_1, X_2, X_1 * X_2$ ] | 44 | 23 | 44 | 27 | 38 | 40 | -34 | 30 | 38 |
| (c)Spline on $X_1$ [$s(X_1)$] | -15 | 32 | 28 | 61 | 39 | 64 | -18 | 33 | 31 |
| (d)Spline on $X_1$ [$s(X_1) + X_2$] | 2 | 31 | 24 | 58 | 38 | 61 | -19 | 33 | 32 |
| (e)PSPP [ $s(Y^*)$ ] | -69 | 33 | 69 | -129 | 37 | 129 | 63 | 32 | 63 |
| (f)PSPP [ $s(Y^*) + X_2$ ] | -71 | 33 | 72 | -119 | 40 | 119 | 66 | 31 | 66 |
| (g) PSPP [ $s(Y^*) + g(X_1^*)$ ] | 24 | 28 | 31 | 21 | 46 | 42 | -29 | 34 | 37 |
| (h)PSPP [ $s(Y^*) + g(X_1)$ ] | 24 | 28 | 31 | 23 | 46 | 43 | -30 | 34 | 38 |
| (i)PSPP +Spline on $X_1$ [$s(Y^*) + s(X_1)$] | 38 | 26 | 41 | 50 | 40 | 54 | -45 | 35 | 48 |
| (j)Bivariate PSPP [ $s(Y^*, X_1)$ ] | 1 | 25 | 20 | -1 | 38 | 32 | -4 | 31 | 26 |

## References

Eilers, P.H.C. and Marx, B.D. (1996). "Flexible Smoothing with B-Splines and Penalties." *Statist. Sci.* 11 89-121.

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

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

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*." Spinger, New York.

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.

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 University Press.

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.

## Appendix: (1) Proof of Theorem 2

**Lemma 1**:

Let $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $X_1 = \begin{pmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{1n} \end{pmatrix}$,

$$X_2 = \begin{pmatrix} x_{21} & \cdots & x_{M1} & (x_1 * x_2)_1 & \cdots & (x_1 * x_M)_1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{2n} & \cdots & x_{Mn} & (x_1 * x_2)_n & \cdots & (x_1 * x_M)_n \end{pmatrix}$$

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

Let $X_{(1)} = \begin{pmatrix} f_1((X_1)_1) & \dots & f_{N1}((X_1)_1) \\ \vdots & \ddots & \vdots \\ f_1((X_1)_n) & \cdots & f_{N1}((X_1)_n) \end{pmatrix}$ and $X_{(2)} = \begin{pmatrix} g_1((X_2)_1) & \dots & g_{N2}((X_2)_1) \\ \vdots & \ddots & \vdots \\ g_1((X_2)_n) & \cdots & g_{N2}((X_2)_n) \end{pmatrix}$ 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 $E_A(Y \mid X_1) = X_{(1)} * \underline{\beta_1}$. 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 $E_A(g_i(X_2) \mid X_1) = X_{(1)} * \underline{\delta_i}$, $i = 1, \dots, N2$. 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 $\hat{X}_{(2)} = [\hat{g}_1(X_2), \dots, \hat{g}_{N2}(X_2)]$, $\underline{\hat{\delta}} = [\underline{\hat{\delta}_1}, \dots, \underline{\hat{\delta}_{N2}}]$. 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 $\underline{\hat{\beta}_1} = \underline{\hat{\gamma}_1} + \underline{\hat{\delta}} * \underline{\hat{\gamma}_2}$.

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

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 $X^T(I - H)Y = X^T Y - X^T(X(X^T X)^{-1} X^T)Y = 0$

We have

$Y - X_{(1)}\underline{\hat{\beta}_1} = (X_{(2)} - X_{(1)}\underline{\hat{\delta}})\underline{\hat{\gamma}_2} + (I - H)Y$

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

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 $E_A(Y \mid X_1) = s_1(X_1; \underline{\beta_1})$, $s_1(X_1; \underline{\beta_1})$ is a spline of $X_1$ indexed by the parameter $\underline{\beta_1}$. Let $\underline{\hat{\beta}_1}$ be the restricted maximum likelihood estimate 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 $E_A(g_i(X_2) \mid X_1) = s_1(X_1; \underline{\delta_i})$, $i = 1, \dots, N2$; $s_1(X_1; \underline{\delta_i})$ is a spline of $X_1$ indexed by the parameter $\underline{\delta_i}$. Let $\underline{\hat{\delta}_i}$ be the restricted maximum likelihood estimates of $\underline{\delta_i}$, the predicted value of $g_i(X_2)$ is $\hat{g}_i(X_2) = s_1(X_1; \underline{\hat{\delta}_i})$. Let $\hat{X}_{(2)} = [\hat{g}_1(X_2), \dots, \hat{g}_{N2}(X_2)]$, $\underline{\hat{\delta}} = [\underline{\hat{\delta}_1}, \dots, \underline{\hat{\delta}_{N2}}]$. Substitute $\hat{X}_{(2)}$ into $\hat{Y}_1(X_1, X_2)$ of model (a) and obtain $\hat{Y}_2^*(X_1) = s(X_1; \underline{\hat{\gamma}_1}) + \hat{X}_{(2)} * \underline{\hat{\gamma}_2}$.

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

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:

$$E_A(Y \mid X_1, X_2) = s_1(X_1; \underline{\gamma_1}) + g(X_2; \underline{\gamma_2})$$
$$= \gamma_0 + \gamma_1 * x_1 + \sum\nolimits_{k=1}^{K} \gamma_{1k}(x_1 - k_k)_+ + X_{(2)}\underline{\gamma_2}$$

the fitting criterion is to minimize $\left\| y - X\gamma \right\|^2 + \lambda^2 \underline{\gamma}^T D \underline{\gamma}$, with $\gamma = (\gamma_0, \gamma_1, \underline{\gamma_2}, \gamma_{11}, ..., \gamma_{1K})^T$, $D = \text{diag}(0_{2+N2}, 1_K)$ . Using the mixed model presentation and restricted maximum likelihood, the fitted values are

$\hat{Y}_1(X_1, X_2; \hat{\lambda}) = X(X^T X + \hat{\lambda}^2 D)^{-1} X^T Y$,

$\hat{\lambda}$ is the estimated penalty. When $n \to \infty$, $\hat{\lambda} \to 0$, thus $\hat{Y}_1(X_1, X_2; \hat{\lambda}) \to \hat{Y}_1(X_1, X_2; 0) = X(X^T X)^{-1} X^T Y$, the least squares estimates of model (a).

Similarly, for model (b),

$E_A(Y \mid X_1) = s_1(X_1; \beta_1) = \beta_0 + \beta_1 x_{11} + \sum\nolimits_{k=1}^{K} \beta_{1k}(x_{11} - k_k)_+$

as $n \to \infty$,

$\hat{Y}_2(X_1; \hat{\lambda}) \to \hat{Y}_2(X_1; 0) = X_{(1)}(X_{(1)}^T X_{(1)})^{-1} X_{(1)}^T Y$ .

For model (c),

$E_A(g_i(X_2) \mid X_1) = \delta_{i0} + \delta_{i1} x_{11} + \sum\nolimits_{k=1}^{K} \delta_{i1k}(x_{11} - k_k)_+$

as $n \to \infty$,

$\hat{g}_i(X_2; \hat{\lambda}) \to \hat{g}_i(X_2; 0) = X_{(1)}(X_{(1)}^T X_{(1)})^{-1} X_{(1)}^T g_i(X_2)$ and

$\hat{X}_{(2)}(\hat{\lambda}) \to \hat{X}_{(2)}(0)$

By lemma 1, $\hat{Y}_2^*(X_1; 0) = \hat{Y}_2(X_1; 0)$, we have,

$$\hat{Y}_2^*(X_1; \hat{\lambda}) = s_1(X_1; \underline{\hat{\gamma}_1}, \hat{\lambda}) + \hat{X}_{(2)}(\hat{\lambda})\underline{\hat{\gamma}_2}$$
$$\to s_1(X_1; \underline{\hat{\gamma}_1}, 0) + \hat{X}_{(2)}(0)\underline{\hat{\gamma}_2} = \hat{Y}_2(X_1; 0)$$

as $n \to \infty$.

From model (b), $\hat{Y}_2(X_1; \hat{\lambda}) \to \hat{Y}_2(X_1; 0)$ 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 specified correctly.

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**

Model:

$$Y \sim N(s_1(Y^{*1}) + ... + s_C(Y^{*c}) + g(Y^*, X_2, ..., X_{p-1}), \sigma^2)$$

Using mixed model presentation and based on the truncated power basis of degree $p$,

$$s_c(Y^{*c}) = \sum\nolimits_{j=1}^{P} \alpha_j \times (Y^{*c})^j + \sum\nolimits_{k=1}^{K} \mu_k (Y^{*c} - \tau_k)_+^P,\ c = 1, ..., C$$

We know that,

$$E(\hat{s}_i(0)) = E(\sum\nolimits_{j=1}^{P} (\hat{\alpha}_j \times 0^j) + \sum\nolimits_{k=1}^{K} \hat{\mu}_k (0 - \tau_k)_+^P)$$
$$= \sum\nolimits_{k=1}^{K} E(\hat{\mu}_k)(0 - \tau_k)_+^P = 0$$

by mixed effects modeling with $\mu_k \sim N(0, \sigma_u^2)$.

Let $X_{i,1} = 1$, we have

$$\hat{Y}_i = \hat{s}_1(Y_i^{*1}) + ... + \hat{s}_c(Y_i^{*c}) + \hat{g}(Y_i^*, X_{i,2}, ..., X_{i,p-1})$$
$$= \hat{s}_1(Y_i^{*1}) + \hat{s}_2(0) + ... + \hat{s}_c(0) + \hat{g}(Y_i^*, X_{i,2}, ..., X_{i,p-1})$$

Take expectation:

$$E(\hat{Y}_i) = E(\hat{s}_1(Y_i^{*1})) + \sum\nolimits_{c=2}^{C} E(\hat{s}_c(0)) + E(\hat{g}(Y_i^*, X_{i,2}, ..., X_{i,p-1}))$$
$$= E(\hat{s}_1(Y_i^{*1}) + \hat{g}(Y_i^*, X_{i,2}, ..., X_{i,p-1}))$$

Then for each level of $X_1 = c$,

$$E(\hat{Y}_i) = E(\hat{s}_c(Y_i^{*c}) + \hat{g}(Y^*, X_{i,2}, ..., X_{i,p-1}))$$
$$= E(\hat{s}_c(Y_i^*) + \hat{g}(Y^*, X_{i,2}, ..., X_{i,p-1}))$$

By PSPP method, within each level of $X_1$,

$Y \sim N(s(Y^*) + g(Y^*, X_2, ..., X_{p-1}), \sigma^2)$.

So for the stratified PSPP,

$E(\hat{Y}_i) = E(\hat{s}_c(Y_i^*) + \hat{g}(Y^*, X_{i,2}, ..., X_{i,p-1}))$ converges to $s(Y_i^*) + g(Y^*, X_{i,2}, ..., X_{i,p-1})$, which completes the proof.