University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2003 *Paper* 5

Robust Likelihood-based Analysis of Multivariate Data with Missing Values

Rod Little [*] An Hyonggin [†]

[*]University of Michigan, rlittle@umich.edu

[†]University of Michigan, hyongg@umich.edu

Robust Likelihood-based Analysis of Multivariate Data with Missing Values

Rod Little and An Hyonggin

Abstract

The model-based approach to inference from multivariate data with missing values is reviewed. Regression prediction is most useful when the covariates are predictive of the missing values and the probability of being missing, and in these circumstances predictions are particularly sensitive to model misspecification. The use of penalized splines of the propensity score is proposed to yield robust model-based inference under the missing at random (MAR) assumption, assuming monotone missing data. Simulation comparisons with other methods suggest that the method works well in a wide range of populations, with little loss of efficiency relative to parametric models when the latter are correct. Extensions to more general patterns are outlined.

Draft Dec 24, 2003

Robust likelihood-based analysis of multivariate data with missing values

Roderick Little and Hyonggin An
University of Michigan

Abstract

The model-based approach to inference from multivariate data with missing values is reviewed. Regression prediction is most useful when the covariates are predictive of the missing values and the probability of being missing, and in these circumstances predictions are particularly sensitive to model misspecification. The use of penalized splines of the propensity score is proposed to yield robust model-based inference under the missing at random (MAR) assumption, assuming monotone missing data. Simulation comparisons with other methods suggest that the method works well in a wide range of populations, with little loss of efficiency relative to parametric models when the latter are correct. Extensions to more general patterns are outlined.
KEYWORDS: double robustness, incomplete data, penalized splines, regression imputation, weighting

1. Introduction

Missing values arise in empirical studies for many reasons. For example, in longitudinal studies, data are missing because of *attrition,* when subjects drop out prior to the end of the study. In most surveys, some individuals provide no information because of non-contact or refusal to respond (*unit* nonresponse). Other individuals are contacted and provide some information, but fail to answer some of the questions (*item* nonresponse). Often indices are constructed by summing values of particular items. For example, in economic studies, total net worth is a combination of values of individual assets or liabilities, some of which may be missing. If any of the items that form the index are missing, some procedure is needed to deal with the missing data.

The missing data *pattern* simply indicates which values in the data set are observed and which are missing. Specifically, let $Y = (y_{ij})$ denote an $(n \times p)$ rectangular dataset without missing values, with *i*th row $y_i = (y_{i1}, \ldots, y_{ip})$ where y_{ij} is the value of variable Y_j for subject *i*. With missing values, the pattern of missing data is defined by the *missing-data indicator matrix* $M = (m_{ij})$ with *i*th row $m_i = (m_{i1}, \ldots, m_{ip})$, such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is present. We assume throughout that (y_i, m_i) are independent over *i*.

Some methods for handling missing data apply to any pattern of missing data, whereas other methods assume a special pattern. For simplicity we consider methods for the simple pattern of *univariate* nonresponse, where missingness is confined to a single variable, say Y_p, and $Y_1, \ldots, Y_{p-1}$ are fully observed. In Section 7 we discuss extensions of our methods to more general patterns, such as *monotone* missing data, where the variables can be arranged so that $Y_{j+1}, \ldots, Y_p$ is missing for all cases where Y_j is missing, for all *j* = 1, ..., *p*-1. This pattern arises commonly in longitudinal data subject to attrition.

The performance of alternative missing-data methods depends strongly on the missing-data mechanism, which concerns the reasons why values are missing, and in particular whether missingness depends on the values of variables in the data set. For example, subjects in a longitudinal intervention may more likely to drop out of a study because they feel the treatment was ineffective, which might be related to a poor value of an outcome measure. Rubin (1976) treated *M* as a random matrix, and characterized the missing-data mechanism by the conditional distribution of M given *Y*, say $f(M|Y, f)$, where f denotes unknown parameters. When missingness does not depend on the values of the data *Y*, missing or observed, that is:

$$f(M|Y, f) = f(M|f) \text{ for all } Y, f ,$$

the data are called missing completely at random (MCAR). With the exception of planned missing-data designs, MCAR is a strong assumption, and missingness often does depend on recorded variables. Let Y_{obs} denote the observed values of Y and Y_{mis} the missing values. A less restrictive assumption is that missingness depends only on values Y_{obs} that are observed, and not on values Y_{mis} that are missing. That is:

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

The missing data mechanism is then called missing at random (MAR). Many methods for handling missing data assume the mechanism is MCAR or MAR, and yield biased estimates when the data are not MAR (NMAR).

The main ideas of this article can be summarized in the following propositions:

(a) When the missing data mechanism is unknown and NMAR, methodological options are limited and not very appealing to the practitioner. Thus, in studies where missing data are likely to arise, efforts should be made to render the MAR assumption plausible, by measuring covariates that characterize nonrespondents (Little and Rubin, 1999).

(b) The most useful covariates for nonresponse adjustment are (i) predictive of the missing values Y_{mis} and (ii) predictive of the missing data indicator *M*. Of the two, criterion (i) is the most important, since conditioning on a covariate that is predictive of *M* but not of Y_{mis} leads to a loss of efficiency without a compensating reduction in bias. Section 3 presents an analysis in support of these statements.

(c) All missing-data adjustments require modeling assumptions relating the missing data to observed covariates. Sensitivity to assumptions is a particularly serious issue for analysis involving covariates that are useful for missing-data adjustments, as described in (b).

(d) Given (a) - (c), missing-data methods based on MAR and models that make relatively weak assumptions relating the covariates to the missing data are useful. Methods of this kind based on propensity splines are proposed in Sections 4 and 5 below, for the special case of univariate nonresponse. These methods are assessed by simulation in Section 6. Some extensions of these methods to more general missing data problems are outlined in Section 7, and Section 8 presents concluding remarks.

2. Limitations of NMAR analyses when the missing data mechanism is unknown.

There is an extensive literature of methods for NMAR missing-data mechanisms; early examples include Heckman's (1976) proposals for handling selectivity bias, and Rubin's (1977) Bayesian analysis. See also Little and Rubin (2002, chapter 15). The difficulty of the problem can be seen by considering the simplest situation of a single variable Y_1 (that is, p = 1), observed for r cases and missing for n - r cases, with no covariate information. Suppose the respondent values of Y_1 are independently distributed with mean m_{1R} and variance s_{11}, and the nonrespondent values are independently distributed with mean m_{1NR} and variance s_{11}. If the observations are independent, then MCAR=MAR, and $m_{1R} = m_{1NR}$. In that case, the sample mean $\bar{y}_1$ based on the r complete cases is unbiased, and in many cases optimal for the mean. If, on the other hand, the data are NMAR, the bias of $\bar{y}_1$ for inference about the overall mean is easily seen to be $f\, l\, s_{11}^{1/2}$, where $f = (n-r)/n$ is the fraction of missing values and $l = (m_{1R} - m_{1NR})/s_{11}^{1/2}$ is the standardized difference in respondent and nonrespondent means. Assuming asymptotic normality and ignoring t corrections, the noncoverage rate of the usual 95% confidence interval $\bar{y}_1 \pm 1.96\sqrt{s_{11}/r}$ based on the complete cases is

$$\Phi(-1.96 + \sqrt{r}\, f\, l) + \Phi(-1.96 - \sqrt{r}\, f\, l),$$

Table 1. Coverage of 95% confidence interval for population mean when the respondent mean has a bias $f\ l = 0.1$**.**

Respondent sample size	20	50	100	200
Coverage rate (%)	7.4	10.9	18.0	29.2

where Φ denotes the normal cumulative density function. Table 1 tabulates this noncoverage rate as a function of the respondent sample size *r*, for a fixed bias of $f\ l = 0.1$. Clearly bias has an increasing distorting effect on the noncoverage as the sample size increases.

Analysis options are clearly limited in the absence of information about the nonrespondents. Other than assuming the bias away, the only alternative is to widen the interval to allow for potential bias. Three approaches to this are:

(a) To develop bounds for the quantity of interest that include all possible values of the missing data. For example, for a binary outcome, one might calculate the sample proportion with all missing values imputed as one, and all missing values imputed as zero (Horowitz and Manski, 2000). This approach tends to be very conservative, and is limited to variables that have finite support.

(b) Conduct a sensitivity analysis for alternative models for nonignorable nonresponse (Rubin, 1977; Little and Wang, 1996; Scharfstein, Rotnitsky and Robins, 1999).

(c) Add a prior distribution for the nonrespondent values and apply the Bayesian paradigm. For example, Rubin (1977) considers the model:

$$m_{1R} \sim \text{ const.}; \ m_{1NR} \mid m_{1R} \sim N(m_{1R}, l\, s_{11})$$

An alternative approach is to attempt to measure covariates that capture differences between respondents and nonrespondents, so that the missing-data mechanism can be considered MAR. For the remainder of this paper we consider models under the assumption that the missing data are MAR, while recognizing that residual dependence of the missing data indicators on missing values of the data may require one of the approaches (a) - (c) delineated above.

3. Covariates to the rescue?

Suppose now that fully observed covariates are available, and let $Y_1,...Y_{p-1}$ denote the variables observed for all *n* cases, and Y_p the variable with missing values, observed for the first *r* cases. The mean of Y_p can be written as

$$m_p = E[(1-M)Y_p] + E[ME(Y_p \mid X)],$$

and $E[(1-M)Y_p]$ can be estimated from the complete cases. To estimate the second term $E[ME(Y_p \mid X)]$, note that under MAR, $E(Y_p \mid X) = E(Y_p \mid X, M = 0) = E(Y_p \mid X, M = 1)$. Hence for incomplete cases ($M = 1$) one can estimate $E(Y_p \mid X)$ from the complete cases and predict the *Y* for each incomplete case by substituting the *X* for that case into the regression formula. If the regression is linear, this leads to the regression estimate:

$$\hat{m}_p = n^{-1}\left(\sum_{i=1}^{r} y_{ip} + \sum_{i=r+1}^{n} \hat{y}_{ip}\right) \tag{1}$$

where $\{y_{ip}, i = 1,..., r\}$ are the observed values of Y_p, and $\hat{y}_{ip} = \hat{b}_0 + \sum_{j=1}^{p-1} \hat{b}_j y_{ij}$ is the prediction from the regression of Y_p on $(Y_1,...,Y_{p-1})$, computed on the *r* complete cases. Eq. (1) is the maximum likelihood (ML) estimate of m_p for a variety of models, including multivariate normality for $(Y_1,...,Y_p)$ (e.g., see Little and Rubin, 2002).

The impact of regressing on covariates for inference about m_p can be assessed by comparing the mean squared error of $\hat{m}_p$ relative to the estimate based on the complete cases,

$\bar{y}_p = \sum_{i=1}^{r} y_{ip} / r$. Consider this comparison for a single covariate (p = 2), where Y_1 and Y_2 are bivariate normal, and the missing data are MAR. The regression estimate (1) is then unbiased for m_2 with mean squared error (e.g. Little and Rubin, 2002)

$$mse(\hat{m}_2) = (s_{22} / r)\left((1-r^2) + (r/n) r^2 + (1-r^2)(1-r/n)^2 \Delta^2\right),$$

ignoring $O(1/r^2)$ terms, where r is the correlation between respondent values of Y_1 and Y_2, and Δ is the difference in the nonrespondent and respondent mean of Y_1, divided by the respondent variance of Y_1. Note the r^2 measures the association between Y_1 and Y_2 and Δ^2 measures the association between Y_1 and M. The mean squared error of $\bar{y}_2$ is

$$mse(\bar{y}_2) = (s_{22} / r) + (1-r/n)^2 \Delta^2 r^2 s_{22},$$

where the first term on the right side is the variance and the second term is the bias. Subtracting and simplifying yields

$$mse(\bar{y}_2) - mse(\hat{m}_2) = (1-r/n) s_{22}\left[(1-r/n) r^2 \Delta^2 + r^2 / r - (1-r/n)(1-r^2)\Delta^2 / r\right]. \qquad (2)$$

The first term in the square parenthesis in Eq. (2) is $O(1)$ and is the bias that has been eliminated by the regression of Y_2 on Y_1 (more generally under NMAR one expects the regression to reduce bias, although it could increase). Both r^2 and Δ^2 must be large for this term to be substantial. The second and third terms in the square parenthesis represent variance reduction from the regression on Y_1. This variance reduction is substantial when r^2 is large. In fact, if r^2 is small and Δ^2 is large, as when Y_1 is predictive of M but not predictive of Y_2, the net value of these terms may be negative, reflecting an increase in variance from the regression on Y_1. These results are summarized in Table 2. Eq. (2) generalizes to a multivariate set of predictors, with the

Table 2. Effect on bias and variance of the estimated mean of Y_2 of regression on a fully-observed covariate Y_1, for combinations of the association between Y_1 and Y_2 (r^2) and the association between Y_1 and M (Δ^2).

	r^2 Low	r^2 High
Δ^2 Low	bias change: ≈ 0 variance change: ≈ 0	bias change : ≈ 0 variance change: $\downarrow$
Δ^2 High	bias change: ≈ 0 variance change: $\uparrow$	bias change : $\downarrow$ variance change: $\downarrow$

obvious generalizations of r^2 and Δ^2. Clearly, the key for both bias and variance reduction is that Y_1 is a good predictor of Y_2.

4. Robust MAR inference with a single covariate

4.1. Robust Prediction

In the previous section we noted that the key to reduce mean squared error for inference about the mean of Y_p is to find predictors that are predictive of Y_p and the missing data indicator *M*. These are the circumstances under which inference are most sensitive to misspecification of the regression of Y_p on $Y_1,...,Y_{p-1}$, since the bias reduction is dependent on an appropriate specification of the model relating Y_p to $Y_1,...,Y_{p-1}$. Thus we now consider robust alternatives to the linear additive model (1). We first consider the case of a single covariate, $p = 2$. Extensions to more than one covariate are discussed in Sections 5 and 7.

Standard regression modeling methods, such as adding polynomial terms and interactions to the regression in (1), are useful strategies. Perhaps the simplest way to weaken assumptions about the relationship between Y_2 and a continuous covariate Y_1 is to group the covariate into categories and regress on dummy variables for the categories. The resulting regression estimate,

$$\hat{m}_2 = \sum_{c=1}^{C} p_c \overline{y}_{c2} \,, \tag{3}$$

is the average of the respondent mean $\overline{y}_{c2}$ in each category weighted by the sample proportion $p_c = n_c / n$ in that category.

An attractive alternative to categorization of continuous covariates is to fit a smooth but relatively nonparametric relationship between Y_2 and the covariate (Cheng, 1994). For example, one might model the regression of Y_2 on Y_1 via a penalized spline (Eilers and Marx, 1996, Ruppert and Carroll, 2000) with a power-truncated spline basis:

$$\begin{aligned} &(Y_2 \mid Y_1, f) \sim N\left(s_2(Y_1, f), s^{\,2}\right) \\ &s_2\left(Y_1, f\right) = f_0 + \sum_{j=1}^{q} f_j y_{i1}^{j} + \sum_{k=1}^{K} f_{q+k}\left(Y_1 - t_k\right)_+^q, \end{aligned} \quad , \tag{4}$$

where q is the degree of polynomial, $\left(x\right)_+^q = x^q I\left(x \ge 0\right)$, $t_1 < \cdots < t_K$ are selected fixed knots, and K is the total number of knots. Then, the penalized least-squares estimator $\hat{f} = \left(\hat{f}_0, \ldots, \hat{f}_{q+K}\right)^{\mathrm{T}}$ can be obtained by minimizing the penalized sum of squared errors

$$\sum_{i=1}^{n} \left\{ y_{i2} - f_0 - \sum_{j=1}^{q} f_j y_{i1}^{j} - \sum_{k=1}^{K} f_{q+k}\left(y_{i1} - t_k\right)_+^q \right\}^2 + l \sum_{k=1}^{K} z\left(f_{q+k}\right),$$

where z is a suitable nonnegative function, and l is a smoothing parameter. The smoothing parameter can be estimated by generalized cross validation or by ML for a linear mixed model, treating $\left(f_0, \ldots, f_q\right)^{\mathrm{T}}$ as a fixed parameter vector and $\left(f_{q+1}, \ldots, f_{q+K}\right)^{\mathrm{T}}$ as a random vector. Cheng (1994) achieves nonparametric smoothing by another method, kernel regression; an attractive feature of the ML version of penalized splines is that they are easily implemented with widely available software such as PROC MIXED in SAS (SAS, 1992) and lme( ) in S-plus (Pinheiro and Bates, 2000).

4.2. Weighting the complete cases.

An alternative to prediction, commonly used for unit nonresponse adjustments in sample surveys, is to weight the complete cases by the inverse of an estimate of the probability of response (e.g. Little and Rubin, 2003, Section 3.3). The mean of Y_2 can be written as:

$$m_2 = E\left[\frac{(1-M)Y_2}{p(Y_1)}\right] / E\left[\frac{1-M}{p(Y_1)}\right],$$

where $p(Y_1) = \Pr(M = 0 \mid Y_1)$ is the probability that Y_2 is observed given Y_1. The denominator in this equation can be ignored under correct specification of the $p(Y_1)$, since it then equals one. Replacing population quantities by sample estimates yields the weighted complete-case estimate:

$$\hat{m}_2 \equiv \bar{y}_{2w} = \left(\sum_{i=1}^{r} w_i y_{i2}\right) / \left(\sum_{i=1}^{r} w_i\right), \tag{5}$$

or

$$\hat{m}_2 \equiv \bar{y}_{2w} = \left(\sum_{i=1}^{r} w_i y_{i2}\right) / n, \tag{5A}$$

where the weight w_i for respondents is a reciprocal of an estimate of $p(y_{i1})$. If Y_1 is grouped into categories, and respondents in category *c* are weighted by the inverse of the estimated response rate r_c / n_c in category *c*, then the resulting estimator (5) or (5A) is identical to the regression estimate (3). Note that if the true response rate is the same for all the categories *c*, as when the data are MCAR, then weighting by the true response rate yields the unweighted sample mean $\bar{y}_2$ based on the complete cases, which is less efficient if the categorized covariate is predictive of response. This is a simple and instructive illustration of increased efficiency when weights are estimated from the sample rather than from population parameters (e.g. Robins, Rotnitsky and Zhao, 1994).

In section 4.1 we used splines to smooth the predictions from a regression model. A different use of smoothing is to smooth the weights in the weighted estimator (5). That is, the weights are replaced by the inverse of the estimated propensity to respond, computed by fitting a spline to the logistic regression of the missing-data indicator M on Y_1.

$$\begin{aligned} &w_i = 1/p_i(\hat{f}), p_i(f) = \Pr(M_i = 1 \mid y_{i1}, f), \\ &\operatorname{logit}\left(p_i(f)\right) = s_M(y_{i1}; f), \end{aligned} \tag{6}$$

where $s_M(y_{i1}; f)$ is a spline for the binary outcome M_i analogous to (5), with unknown parameters f, and $\hat{f}$ is an estimate of f. The latter can be obtained by fitting a generalized linear mixed model for the spline regression of M on Y_1 (Breslow and Clayton, 1993). The utility of splines for prediction, as in Eq. (4), and for weighting, as in Eq. (6) is compared in the simulations in Section 6.

4.3. Calibration estimators

The mean of Y_2 can be written in a way that combines the features of prediction and weighting:

$$m_2 = E\left[\frac{(1-M)}{p(Y_1)}\left(Y_2 - E(Y_2 \mid Y_1)\right)\right] + E\left(E(Y_2 \mid Y_1)\right).$$

Estimating quantities in this expression leads to a "calibration" estimator of the form

$$\hat{m}_2 = n^{-1}\left(\sum_{i=1}^{r} w_i(y_{i2} - \hat{y}_{i2})\right) + n^{-1}\left(\sum_{i=1}^{n} \hat{y}_{i2}\right). \tag{7}$$

where the predictions $\hat{y}_{i2}$ from the model are calibrated by adding a term consisting of weighted residuals from the model. The estimator (7) has properties of semi-parametric efficiency and "double-robustness" (Robins, Rotnitsky and Zhao, 1994; Robins and Rotnitsky, 2001), in the sense that the estimate is consistent if just one of the models for prediction and weighting are correctly specified. However, since the calibration of the predictions is to correct effects of model misspecification, we believe that the calibration of the predictions in Eq. (7) is unnecessary if the prediction model does not make strong parametric assumptions, as in (4). This conjecture is supported by the simulation studies in Section 6.

5. Robust MAR Inferences with More than One Covariate

With sufficient sample size, a penalized spline provides a useful model for predictions based on a single covariate. With several covariates an additive model might be fitted with splines on the continuous covariates. In particular, Scharfstein and Izzary (2003) consider a flexible class of estimators that includes (7) as a special case where the propensity score model and mean model follow generalized additive regressions. We propose here a prediction model that addresses the "curse of dimensionality" by focusing the spline on a particular function of the covariates most sensitive to model misspecification, namely the propensity score. Suppose that Y_p is subject to missing values and $Y_1,\ldots,Y_{p-1}$ are fully observed covariates, and $p \geq 3$ so that there are at least 2 covariates. We first define the logit of the propensity score for Y_p to be observed, given the covariates $Y_1,\ldots,Y_{p-1}$:

$$Y_1^* = \text{logit}\left(\Pr(M=0 \mid Y_1,\ldots,Y_{p-1})\right). \tag{8}$$

The key property of the propensity score is that conditional on the propensity score and assuming MAR, missingness of Y_p does not depend on $Y_1,\ldots,Y_{p-1}$ (Rosenbaum and Rubin, 1983). Thus the mean of Y_p can be written as:

$$m_p = E[(1-M)Y_p] + E[M \times E(Y_p \mid Y_1^*)].$$

This motivates the following method:

(a) Estimate Y_1^* by a logistic regression of M on $(Y_1,...,Y_{p-1})$, yielding estimated propensity $\hat{Y}_1^*$; this regression can include nonlinear terms and interactions in $(Y_1,...,Y_{p-1})$, and with sufficient data could be modeled by a spline as in (6).

(b) Predict the missing values of Y_p by a spline regression of Y_p on $\hat{Y}_1^*$. Since variables other than $\hat{Y}_1^*$ may be good predictors of Y_p, the other covariates are entered in the regression parametrically, for example as linear additive terms.

More specifically, we replace one of the predictor variables, say Y_1, by Y_1^*, to avoid multicollinearity; we then predict the missing values of Y_p using the following model for the distribution of $Y_2,...Y_p$ given Y_1^*:

$$\begin{aligned}&(Y_2,...,Y_{p-1} \mid Y_1^*) \sim N\left(\left(s_2(Y_1^*),...,s_{p-1}(Y_1^*)\right),\Sigma\right)\\&(Y_p \mid Y_1^*, Y_2,...,Y_{p-1}, b) \sim N\left(s_p(Y_1^*)+g(Y_1^*,Y_2^*,...,Y_{p-1}^*,b), s^2\right),\end{aligned} \tag{9}$$

where $s_j(Y_1^*)=E(Y_j \mid Y_1^*)$ is a spline for the regression of Y_j on Y_1^*, $Y_j^* = Y_j - s_j(Y_1^*)$, $j=2,...,p-1$, and g is a parametric function indexed by unknown parameters b, with the property that

$$g(Y_1^*,0,\ldots,0,b)=0 \text{ for all } b\ . \tag{10}$$

Examples of functions $g()$ that satisfy (10) include a linear additive model for $(Y_2^*,...,Y_{p-1}^*)$,

$$g(Y_1^*,Y_2^*,...,Y_{p-1}^*,b)=\sum_{j=2}^{p-1} b_j Y_j^*\ , \tag{11}$$

and a model that includes first order interactions between Y_1^* and $(Y_2^*,...,Y_{p-1}^*)$:

$$g(Y_1^*,Y_2^*,...,Y_{p-1}^*,b)=\sum_{j=2}^{p-1} b_j Y_j^* + \sum_{j=2}^{p-1} b_{j+p-2} Y_1^* Y_j^*\ .$$

We call Eq. (9) a propensity spline prediction model. The idea of explicitly including the propensity score as a covariate in the prediction model was previously proposed by David et al. (1983) and in a more general context in Robins (1999). The use of a spline on one regressor variable is an application of Yu and Ruppert's (2002) partially linear single-index model in the missing-data setting.

The following theorem defines a double robustness property of prediction estimates of the mean of Y_p based on the model (9).

Theorem 1. Let $\hat{m}_p$ be the prediction estimator (1) based on the model (9), and assume MAR. Then $\hat{m}_p$ is a consistent estimator of m_p if either (a) the mean of Y_p given $(Y_1^*, Y_2, ..., Y_{p-1})$ in the model (9) is correctly specified, or (b1) the propensity Y_1^* is correctly specified, and (b2) $E(Y_j^* \mid Y_1^*) = s_j(Y_1^*)$ for $j = 2,..., p$, that is, the splines $s_j(Y_1^*)$ correctly model the regressions of Y_j on Y_1^* for $j = 2,..., p$. The robustness feature of condition (b2) is that the regression function g does not have to be correctly specified.

Outline proof of Theorem 1. Consistency under (a) follows from the usual properties of prediction under a well-specified regression model. For consistency under (b1) and (b2), we need to show that

$$\left(\sum_{i=r+1}^{n} \hat{y}_{ip} / (n-r)\right) \rightarrow E(Y_p \mid M = 1) \text{ as } (n-r) \rightarrow \infty . \qquad (12)$$

Let $\hat{y}_{ip}$ be a prediction for a nonrespondent $(i = r+1,..., n)$. Note that

$$\hat{y}_{ip} = \hat{s}_p(\hat{y}_{i1}^*) + g(\hat{y}_{i1}^*, \hat{y}_{i2}^*, ..., \hat{y}_{ip-1}^*; \widehat{b}),$$

where $\hat{y}_{ij}^{*} = y_{ij} - \hat{s}_j(\hat{y}_{i1}^{*})$ and $\hat{s}_j$ denotes the sample estimate of the spline s_j. Since by assumption the propensity model and the splines are correctly specified,

$$\hat{y}_{ip} \to \tilde{y}_{ip} = s_p(y_{i1}^{*}) + g(y_{i1}^{*}, y_{i2}^{*}, ..., y_{ip-1}^{*}; b^{*}) ,$$

where b^{*} is the limiting value of $\hat{b}$, and the estimates $\hat{y}_{i1}^{*}$ and $\hat{s}_j$ have been replaced by limiting values y_{i1}^{*} and s_j. Now

$$\begin{aligned} E(\tilde{y}_{ip} \mid y_{i1}^{*}) &= s_p(y_{i1}^{*}) + E\left(g(y_{i1}^{*}, y_{i2}^{*}, ..., y_{ip-1}^{*}; b^{*}) \mid y_{i1}^{*}\right) \\ &\simeq s_p(y_{i1}^{*}) + g\left(y_{i1}^{*}, E(y_{i2}^{*} \mid y_{i1}^{*}), ... E(y_{ip-1}^{*} \mid y_{i1}^{*}); b^{*}\right) \\ &= s_p(y_{i1}^{*}) + g\left(y_{i1}^{*}, 0, ... 0; b^{*}\right) = s_p(y_{i1}^{*}) = E(y_{ip} \mid y_{i1}^{*}). \end{aligned}$$

Hence the conditional expectation of $\hat{y}_{ip}$ given y_{i1}^{*} converges to $E(y_{ip} \mid y_{ip}^{*})$, which equals $E(y_{ip} \mid y^{*}_{ip}, m_i = 1)$ by the balancing property of propensity scores. Hence $\sum_{i=r+1}^{n} \hat{y}_{ip} / (n-r)$ converges to $E(Y_p \mid M = 1)$, as required for consistency.

The double robustness property in this theorem is more restrictive than the double robustness property for the calibration estimator (7), which requires only (a) or (b1) and does not require (b2) in addition to (b1). On the other hand, (b2) is weak, since the univariate regressions on Y_1^{*} are modeled nonparametrically in (9) with splines $\{s_j(Y_1^{*})\}$. The inclusion of g in (9) does not affect consistency even if it is misspecified, and it has the potential of improving efficiency, as for example when the variables $(Y_2, ..., Y_{p-1})$ are predictive of Y_p but the propensity score Y_1^{*} is not. The properties of propensity spline prediction are further explored in the simulation study in the next section.

6. Simulations

We conducted three simulation studies to examine the performance of the estimators of the mean of Y with missing data under MAR. The first simulation study involved a single fully-observed covariate Y_1, generated from a uniform distribution between –1 and 1, and one variable Y_2 with missing values, generated from a normal distribution with one of four mean structures:

(I) constant : $N\left(6, 2^2\right)$,

(II) linear : $N\left(6+10Y_1, 2^2\right)$,

(III) cubic: $N\left(-4+5\left(1+Y_1\right)^3, 2^2\right)$, and

(IV) sine : $N\left(6+15\sin\left(p\,Y_1\right), 2^2\right)$.

The expected value of Y_2 is 6 for four mean structures, and (II) – (IV) model a strong predictive relationship between Y_1 and Y_2. We created missing values from four different models for the propensity to respond:

(I) constant (MCAR) : $\text{logit}\left(pr\left(M=0\,|\,Y_1\right)\right)=0.5$,

(II) linear : $\text{logit}\left(pr\left(M=0\,|\,Y_1\right)\right)=3Y_1$,

(III) cubic : $\text{logit}\left(pr\left(M=0\,|\,Y_1\right)\right)=3Y_1^3$, and

(IV) sine: $\text{logit}\left(pr\left(M=0\,|\,Y_1\right)\right)=1.5\sin\left(p\,Y_1\right)$.

The response rate for all these propensity structures is 0.5, and (II) – (IV) model a strong predictive relationship between Y_1 and M. The non-MCAR simulations are thus focused on the fourth cell of Table 2, when a well-specified regression adjustment has strong gains. For each combination of mean and propensity structure, 500 simulated data sets with sample size n=100 were generated. Then, six estimators of the mean of Y_2 are compared with the mean of Y_2 before deletion (BD), namely:

(CC) The complete-case estimate, deleting the incomplete cases.

(LP) The prediction estimator (1) based on linear regression.

(SP) The prediction estimator (1) based on a penalized spline regression model (4).

(LW) The weighted estimator (5) with weights computed as the inverse of the response propensity estimated by linear logistic regression of the missing-data indicator on Y_1.

(SW) The weighted estimator (5) with weights computed as the inverse of the response propensity estimated by spline logistic regression of the missing-data indicator on Y_1, as in (6).

(SPW) The calibration estimator (7) with predictions computed as for SP and weights computed as for SW.

For the penalized splines methods, we chose 20 equally spaced fixed knots over Y_1 and a truncated linear basis.

The results from this simulation study are summarized in Table 3 and Figure 1. For each combination of mean structure and response propensity structure and for each estimator, the standardized bias

STDBIAS = 100×(bias/empirical standard error),

is tabulated, where bias is the deviation of the average estimate over the 500 simulated data sets from the true parameter value, and the empirical standard error is standard deviation of the estimates over the 500 simulated data sets. Also the relative root mean squared error compared with the BD estimator

RRMSE = 100×(RMSE(estimator) – RMSE(BD))/RMSE(BD)

is tabulated, where RMSE is the square root of the average squared deviation of the estimate from the true value over the 500 simulated data sets. The RRMSE values for methods other than

CC are displayed in Figure 1, with means and medians as summaries of overall performance. We conclude the following from Table 3 and Figure 1:

(1) In terms of median or mean RRMSE over problems, SP and SPW are the best methods, SW, LW and LP are intermediate, and CC is much worse than the other methods. In particular, methods that include predictions based a spline (SP or SPW) do better than the method that uses weights based on a spline (SW).

(2) The bias of SP is minor and comparable to that of SPW, and the biases and RRMSE's of SP and SPW are very similar. This suggests that calibration is not needed for SP, although it does not hurt the method.

(3) For the constant mean model, Y_1 and Y_2 are independent; in this case the missing data mechanism is always MCAR since missingness depends on a variable Y_1 that is independent of Y_2. None of the methods display bias in this situation, as theory would predict. For data from the constant propensity model, the RRMSE's of all the methods are very similar; for data from the other propensity models, CC analysis is best. For non-constant mean models and missing-data mechanisms other than MCAR, CC analysis has a large bias and is not competitive with other methods.

(4) When Y_1 and Y_2 are linearly related, LP is the best method, as predicted by theory, but SP is nearly as good, showing little loss in efficiency. LW and SW are noticeably inferior in this case.

(5) When Y_1 and Y_2 are not linearly related and data are not MCAR, LP predictably suffers from bias from model misspecification; SP does much better in these cases since it is not based on a linearity assumption.

(6) When the model for the propensity is not linear, there is some evidence that SW is better than LW, consistent with the fact that SW does not make a linearity assumption for the logit of the propensity. However gains are less dramatic than for SP over LP.

The second simulation study involved two fully observed covariates Y_1 and Y_2, and one variable Y_3 with missing values. We generated Y_1 and Y_2 as independent uniform deviates between –1 and 1, and Y_3 from a normal distribution with one of four mean structures:

(I) constant: $N\left(10, 2^2\right)$,

(II) linear : $N\left(10\left(1+Y_1+3Y_2\right), 2^2\right)$,

(III) additive : $N\left(118+\left(3Y_1-3\right)^3+\left(3Y_2-3\right)^3, 2^2\right)$,

(IV) non-additive : $N\left(10\left(1+Y_1+Y_2+4Y_1Y_2\right), 2^2\right)$.

The expected value of Y_3 for all these mean structures is 10. We simulated four response propensity structures, all of which yield an expected response rate of 0.5:

(I) constant : $\operatorname{logit}\left(pr\left(M=0 \mid Y_1, Y_2\right)\right)=0.5$,

(II) linear : $\operatorname{logit}\left(pr\left(M=0 \mid Y_1, Y_2\right)\right)=Y_1+Y_2$,

(III) additive : $\operatorname{logit}\left(pr\left(M=0 \mid Y_1, Y_2\right)\right)=Y_1^3+Y_2^3$,

(IV) non-additive : $\operatorname{logit}\left(pr\left(M=0 \mid Y_1, Y_2\right)\right)=Y_1+Y_2+3Y_1Y_2$.

For each combination of mean and response propensity structures, 500 simulated data sets of size *n*=100 were generated. Then, we compared the mean before deletion (BD) with the following eight estimators of the mean of Y_3 from the incomplete data:

(CC) The complete-case mean.

(LP) Regression prediction from a linear regression of Y_3 on Y_1 and Y_2.

(ASP) Additive Spline Prediction, namely the prediction estimator (1) based on an additive regression model of Y_3 on penalized splines for Y_1 and Y_2.

(LW) The weighted estimator (5) with weights computed as the inverse of the response propensity estimated by linear logistic regression of the missing-data indicator on Y_1 and Y_2.

(ASW) Additive Spline Weighting, namely the weighted estimator (5) with weights computed as the inverse of the response propensity estimated by an additive spline logistic regression of the missing-data indicator on Y_1 and Y_2.

(ASPW) The calibration estimator (7) with predictions computed as for ASP and weights computed as for ASW.

(SPP) Penalized spline propensity prediction based on a regression of Y_3 on the spline of Y_3^*, the linear predictor of the estimated propensity to respond from a linear logistic regression of M on Y_1, Y_2; that is, Eq. (9) with $g = 0$.

(SPPL) Penalized spline propensity prediction based on Eq. (9) with a linear parametric term for Y_2. That is, Eq. (9) with g given by Eq. (11).

We chose 15 equally spaced knots over Y_1 and Y_2 respectively and a truncated linear basis for the ASW and ASPW. We also chose 20 equally spaced knots over the estimated response propensity and a linear truncated basis for the SPP and SPPL.

Results in Table 4 and Figure 2 can be summarized as follows:

(1) The propensity spline prediction methods (SPP, SPPL) and the prediction methods based on additive splines (ASP, ASPW) do best overall, followed by the linear prediction methods LP and the weighting methods, LW and ASW; CC is much worse than the other methods since it is very biased except for simulations with a constant mean model or an MCAR response propensity.

(2) The methods based on additive splines (ASP, ASPW) do slightly better than propensity spline methods (SPP, SPPL) when the prediction model is additive, and considerably worse when the prediction model is non-additive. In that case the additive models are misspecified, and

the calibration estimator is also biased when the propensity model is not additive. In fact all the methods were biased for this rather demanding problem, but the propensity spline methods have less bias than the others.

(3) Little gain was seen from adding Y_2 to the propensity spline, since SPP and SPPL performed very similarly. Greater gains might be expected in problems with more useful covariates.

(4) Results for ASPW are very similar to those of ASP, suggesting that calibration of ASP has little effect for this particular set of problems.

(5) As in the first simulation, the weighting methods LW and ASW are less efficient than the prediction methods, and smoothing the weight by a spline does not appear to help much.

This simulation study does not display the potential for SPPL to yield gains in precision over SPP when the covariates other than the propensity are predictive of the outcome. We thus simulated two additional cases where the logit of response was linear in $Y_1 + Y_2$, but the mean was more strongly correlated with $Y_1 - Y_2$. In the first case the mean of Y_3 is a nonlinear additive function of Y_1 and Y_2, namely

$$Y_3|Y_1, Y_2 \sim N((3*Y_1 - 3)^3\text{-}(32/27)*(3*Y_2 - 3)^3,\ 2^2).$$

In the second case the mean of Y_3 is a nonlinear non-additive function of Y_1 and Y_2, namely

$$Y_3|Y_1, Y_2 \sim N(10 * (1 + Y_1 - 3*Y_2 + 3*Y_1*Y_2),\ 2^2):$$

The results of these simulations are shown in Table 5. Note that in these cases both SPP and SPPL display minimal bias, but SPPL shows the expected gain in precision, reflected in lower RRMSE's. As might be expected, the methods that fit additive splines, ASP and ASPW, have the lowest RRMSE's in the additive case, but are inferior to SPP and SPPL in the non-additive case.

The results of any simulation study are limited by the choice of populations simulated, and should be interpreted with caution. Our conclusion from these simulations is that the predictions from spline models can yield relatively robust estimates of the population mean.

With several covariates, additive splines work well when the effects of the variables are additive, but the propensity spline method provides an attractive alternative way of addressing the "curse of dimensionality".

7. Extensions to Monotone and General Patterns

The propensity spline model (9) of the previous section can be extended to a monotone pattern by applying it sequentially to each block of missing variables. Missing values of covariates are replaced by their predictions in this sequence of regressions. Multiple imputation versions of this approach, where draws from the predictive distribution are imputed rather than means, and extensions to general patterns of missing data based on the sequential imputation methods of Raghunathan et al. (2001), will be examined in future research.

8. Conclusions

Despite the large literature devoted to nonignorable missing data adjustments, we believe that the key to successful treatment of missing data is to measure covariates that are predictive of the missing values, and to model carefully the relationships between the missing variables and these covariates. Likelihood-based methods based on multivariate models for the data are useful tools for making efficient use of the available data, but standard models such as the multivariate normal imply linear additive relationships between the variables that may be too simplistic in certain settings. We propose easily-fitted spline models that yield regression predictions that are more robust to nonlinearity in the relationship between the missing variables and the covariates, under the MAR assumption. A key idea is to single out the propensity score for this robust form of modeling.

A limitation of the work on propensity spline methods described here is that it is focuses on point estimation. Inferences for the propensity spline prediction model require valid estimates of standard errors, and ideally Student t corrections for small samples. Possible approaches to estimating standard errors include:

(a) computing the estimate on a set of bootstrap samples, and calculating a bootstrap standard error from the sample variance over the bootstrap samples, or from percentiles of the bootstrap distribution;

(b) ignoring sampling error in estimating the propensity $\hat{p}(Y_1,...,Y_{p-1})$, and using asymptotic standard errors for the model (9) based on standard linear mixed model formulae.

(c) using the propensity spline prediction model to multiply-impute draws from the predictive distribution of the missing values, and then using multiple imputation methods for estimating the variance (e.g. Rubin, 1987; Little and Rubin, 2002, chapter 10).

Simulations comparing these approaches are currently underway. Future work will also consider extensions to general patterns of missing data and non-normal outcomes, based on extensions of the sequential imputation method of Raghunathan et al. (2001).

Acknowledgments

This research was supported by National Science Foundation Grant DMS 9408837. We greatly appreciate the helpful comments of Bin Nan, Trivellore Raghunathan, an associate editor and a referee.

References

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

Cheng, P.E. (1994). Nonparametric Estimation of Mean Functionals with Data Missing at Random. *Journal of the American Statistical Association*, 89, 81-87.

David, M., Little, R.J.A., Samuhel, M.E. and Triest, R.K. (1983). Imputation models based on the propensity to respond. *Proceedings of the Business and Economics Section, American Statistical Association 1983*, 168-173

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

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models, *Annals of Economic and Social Measurement* 5, 475-492.

Horowitz, J.L. and Manski, C.F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data, *Journal of the American Statistical Association*, 95, 77-88 (with discussion).

Little, R.J. and Rubin, D.B. (1999). Comment on "Adjusting for Non-Ignorable Drop-out Using Semiparametric Models" by D. O. Scharfstein, A. Rotnitsky and J.M. Robins. *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., and Wang, Y-X (1996). Pattern-mixture models for multivariate incomplete data with covariates, *Biometrics* **52**, 98-111.

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

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 Rotnitsky (2001). Comment on "Inference for semiparametric models: some questions and an answer" by P. Bickel and J. Kwon. *Statistica Sinica*, 11, 920-936.

Robins, J.M., Rotnitsky, 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.

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.

Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys, *Journal of the American Statistical Association* 72, 538-543.

Ruppert, D. and Carroll R.J. (2000). Spatially adaptive penalties for spline fitting, *Australia and New Zealand Journal of Statistics* 42, 205-223.

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.

Scharfstein, D. and Irizarry, R. Generalized additive selection models for the analysis of nonignorable missing data. *Biometrics*, in press.

Scharfstein, D., Rotnitsky, A. and Robins, J. (1999). Adjusting for nonignorable dropout using semiparametric models, *J. Am. Statist. Assoc.* **94**, 1096-1146 (with discussion).

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models, *Journal of the American Statistical Association* 97, 1042-1054.

Table 3. Simulation Study Comparing Estimators with a Single Covariate.

Propensity Model→		Constant (MCAR)(I)		Linear (II)		Cubic (III)		SINE (IV)	
↓ Mean Model		STDBIAS	RRMSE	STDBIAS	RRMSE	STDBIAS	RRMSE	STDBIAS	RRMSE
Constant (I)	BD	-3	0	0	0	1	0	-4	0
	CC	-3	25	-5	40	1	39	-4	35
	LP	-3	25	-8	85	1	53	-6	45
	SP	-3	25	-9	94	0	55	-6	48
	LW	-3	25	-6	115	0	52	-6	52
	SW	-3	25	-5	107	0	54	-6	54
	SPW	-3	25	-8	116	-1	58	-5	55
Linear (II)	BD	6	0	1	0	4	0	2	0
	CC	8	36	504	493	286	293	269	295
	LP	5	5	3	15	3	9	3	10
	SP	5	5	3	18	2	10	3	10
	LW	6	5	11	92	23	19	-26	43
	SW	6	5	22	79	32	27	9	22
	SPW	5	5	3	23	2	11	3	11
Cubic (III)	BD	-4	0	6	0	6	0	2	0
	CC	-1	28	383	466	247	333	228	239
	LP	-6	6	-149	147	-68	53	-38	13
	SP	-4	2	0	11	-1	3	0	3
	LW	-3	6	17	42	16	5	-13	5
	SW	-4	3	25	42	27	12	12	11
	SPW	-4	2	2	11	0	3	1	3
SINE (IV)	BD	6	0	-2	0	-1	0	-5	0
	CC	5	24	430	421	168	160	408	395
	LP	6	12	50	75	-79	61	182	144
	SP	6	1	-43	65	-20	10	-1	3
	LW	6	12	-2	33	-87	52	156	114
	SW	6	11	5	31	-51	32	72	39
	SPW	6	1	-42	65	-19	10	-2	3

Table 4. Simulation Study Comparing Estimators with Two Covariates.

Propensity Model → / ↓ Mean Model		Constant (MCAR) (I) STDBIAS	Constant (MCAR) (I) RRMSE	Linear (II) STDBIAS	Linear (II) RRMSE	Additive (III) STDBIAS	Additive (III) RRMSE	Non-Additive (IV) STDBIAS	Non-Additive (IV) RRMSE
Constant (I)	BD	5	0	-4	0	-4	0	1	0
	CC	7	35	-3	42	-6	48	0	52
	LP	7	36	-5	57	-6	57	-5	56
	ASP	7	36	-3	60	-6	58	-5	58
	LW	7	36	-4	61	-6	56	-6	66
	ASW	7	36	-4	62	-5	57	-7	71
	SPP	7	36	-4	59	-5	57	-5	59
	SPPL	7	36	-4	58	-5	57	-5	58
	ASPW	7	37	-3	64	-4	59	-5	66
Linear (II)	BD	1	0	5	0	3	0	-8	0
	CC	-1	25	232	240	147	137	168	188
	LP	1	1	5	1	3	1	-8	1
	ASP	1	1	5	1	3	1	-8	1
	LW	2	1	6	25	3	5	-113	82
	ASW	2	1	11	25	9	6	-111	90
	SPP	1	1	4	4	1	2	-22	5
	SPPL	1	1	5	1	3	1	-8	1
	ASPW	1	1	5	1	3	1	-8	1
Additive (III)	BD	-5	0	2	0	8	0	-3	0
	CC	-2	25	272	264	180	166	191	233
	LP	-3	5	38	21	36	16	24	25
	ASP	-5	0	3	0	9	0	-2	1
	LW	-6	5	7	48	19	30	-91	134
	ASW	-5	3	15	40	25	26	-103	141
	SPP	-5	4	10	11	24	9	51	40
	SPPL	-5	3	14	11	23	8	69	39
	ASPW	-5	0	3	0	9	0	-2	1
Non-Additive (IV)	BD	2	0	-2	0	-7	0	7	0
	CC	6	35	116	154	71	90	303	393
	LP	5	28	-82	113	-35	55	221	215
	ASP	5	30	-80	131	-34	67	222	243
	LW	5	27	-6	33	-5	30	250	218
	ASW	5	28	-2	33	-4	34	253	224
	SPP	5	18	-11	22	-6	19	154	106
	SPPL	5	18	-10	23	-5	21	155	111
	ASPW	6	29	-18	90	-13	59	236	308

Table 5. Supplemental Simulations where Covariates Other than the Propensity Score are Predictive of Outcome.

	Additive Mean		Non-Additive Mean	
METHOD	STDBIAS	RRMSE	STDBIAS	RRMSE
BD	6	0	-3	0
CC	-25	31	-115	99
LP	-3	13	-64	42
ASP	5	1	-64	51
LW	-2	27	-7	43
ASW	1	18	-9	42
SPP	-1	31	-25	21
SPPL	2	11	-23	12
ASPW	5	1	-16	36

Figure 1. Histograms of RRMMSE for Methods other than CC Analysis, for Simulation 1 with a Single Covariate.

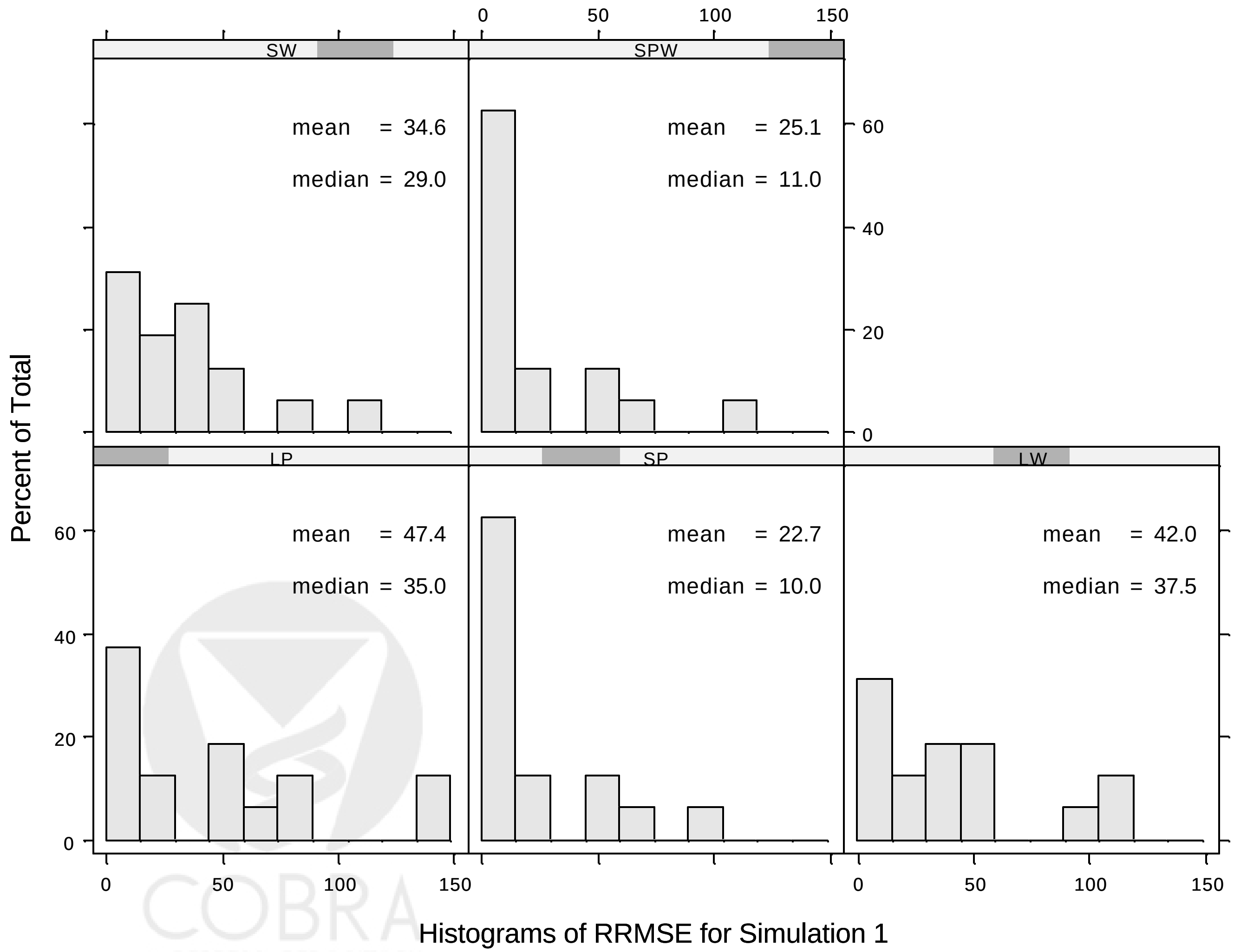

Figure 2. Histograms of RRMMSE for Methods other than CC Analysis, for Simulation 2 with Two Covariates.