

Generalized Additive Selection Models for the Analysis of Studies with Potentially Non-ignorable Missing Outcome Data

Daniel O. Scharfstein and Rafael A. Irizarry*

March 10, 2003

Abstract

Rotnitzky, Robins, and Scharfstein (*Journal of the American Statistical Association*; 1998) developed a methodology for conducting sensitivity analysis of studies in which longitudinal outcome data are subject to potentially non-ignorable missingness. In their approach, they specify a class of fully parametric selection models, indexed by a non- or weakly- identified selection bias function which indicates the degree to which missingness depends on potentially unobservable outcomes. Estimation of the parameters of interest proceeds by varying the selection bias function over a range considered plausible by subject-matter experts. In this paper, we focus on cross-sectional, univariate outcome data and extend their approach to a class of semiparametric selection models, using generalized additive restrictions. We propose a backfitting algorithm to estimate the parameters of the generalized additive selection model. For estimation of the mean outcome, we propose three types of estimating functions: simple inverse weighted, doubly robust, and orthogonal. We present the results of a data analysis and a simulation study.

KEY WORDS: Backfitting; Double Robustness; Inverse Weighting; Sensitivity Analysis; Smoothing

*Daniel O. Scharfstein is Associate Professor and Rafael A. Irizarry is Assistant Professor of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205.

1 Introduction

Rotnitzky, Robins, and Scharfstein (RRS; 1998) developed a methodology for conducting sensitivity analysis of studies in which longitudinal outcome data are subject to potentially non-ignorable missingness. In their approach, they specify a class of fully parametric selection models, indexed by a non- or weakly- identified selection bias function which indicates the degree to which missingness depends on potentially unobservable outcomes. Estimation of the parameters of interest proceeds by varying the selection bias function over a range considered plausible by subject-matter experts. In this paper, we focus on cross-sectional, univariate outcome data and extend their approach to a class of semiparametric selection models, using generalized additive restrictions.

1.1 Data Structure and Notation

Let $\mathbf{X} = (X_1, \dots, X_p)'$ denote a p -dimensional vector of covariates. Let Y denote the outcome of interest. Let R be an indicator that takes on the value 1 if Y is observed and 0 otherwise. The observed data for an individual are $O = (R, RY, \mathbf{X}')'$. We assume that we observe n i.i.d. copies of O ,

$$\mathbf{O} = \{O_i = (R_i, R_i Y_i, \mathbf{X}_i')' : i = 1, \dots, n\}$$

Let $n_1 = \sum_{i=1}^n R_i$ be the number of subjects with complete data and, without loss of generality, let $R_1 = \dots = R_{n_1} = 1$ and $R_{n_1+1} = \dots = R_n = 0$.

The goal is to use \mathbf{O} to draw inference about a functional of the distribution Y . For concreteness, we will focus on estimating the mean of Y , denoted by μ .

1.2 Model

In order to estimate μ , assumptions are required on the missingness mechanism. RRS considered the following model of the conditional probability of R given \mathbf{X} and Y :

$$\text{logit } P[R = 0 | \mathbf{X}, Y] = \eta(\mathbf{X}) + q(Y) \tag{1}$$

where $\eta(\mathbf{X})$ is an unknown function of \mathbf{X} and $q(Y)$ is a specified function of Y . In this model, note that $q(Y) = 0$ corresponds to the assumption of missing at random (Rubin, 1976). When $q(Y)$ is a function of Y , then the missingness mechanism is said to be non-ignorable. The function $q(Y)$ indicates the effect of the potentially unobserved outcome of the probability of missingness, after controlling for the observed covariates \mathbf{X} . RRS showed that given the function q , model (1) is a non-parametric model for the law of the observed data. As a result, the function q is not identifiable and no statistical test can reject a specific choice of q . In addition, RRS showed that, given q , μ is identified under (1). With this in mind, RRS suggested that one should perform sensitivity analysis with respect to q .

Estimation of μ under (1) requires an estimator of $\eta(\mathbf{X})$. When \mathbf{X} is high-dimensional (i.e., it cannot be well approximated by a discrete random variable with a moderate number of levels), non-parametric estimation is not feasible in moderate sized datasets due to the curse of dimensionality (Robins and Ritov, 1997). Thus, RRS recommended imposing additional restrictions on $\eta(\mathbf{X})$. In particular, they assumed $\eta(\mathbf{X})$ followed a fully parametric model, $\eta(\mathbf{X}; \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is a finite-dimensional parameter vector (e.g., $\eta(\mathbf{X}; \boldsymbol{\gamma}) = \gamma_1 + \gamma_2 X_1 + \dots + \gamma_{p+1} X_p$). By imposing these additional restrictions, the model is no longer non-parametric - it is semiparametric. In theory, one can then test for correct specification of $q(Y)$ and the restrictions imposed on $\eta(\mathbf{X})$. Thus, RRS recommend sensitivity analysis with respect to q and suggest choosing the dimension of $\eta(\mathbf{X})$ large enough so any goodness-of-fit will have very low power, but choosing the dimension small enough so that there exist estimators of μ which will have a nearly normal sampling distribution with variance small enough to be of substantive use to subject matter experts. The general idea of not using additional restrictions to estimate $q(Y)$ has been supported by a number of statisticians (see, for example, the discussions of Diggle and Kenward, 1994 and Scharfstein *et al.*, 1999)

With this in mind, in this paper, we show how to draw inference about μ under a flexible, additive model $\eta(\mathbf{X}; \boldsymbol{\gamma})$ for $\eta(\mathbf{X})$ of the following form:

$$\eta(\mathbf{X}; \boldsymbol{\gamma}) = \gamma_1(X_1) + \gamma_2(X_2) + \dots + \gamma_p(X_p) \tag{2}$$

where $\gamma_1(X_1)$ is an unknown smooth function of X_1 and $\gamma_j(X_j)$ ($j = 2, \dots, p$) are unknown smooth, mean zero functions of X_j .

1.3 Focus and Outline

Limited theoretical work exists for generalized additive models. For example, Stone (1986), Linton and Härdle (1996), and Kauermann and Opsomer (2001) present estimators and show, under regularity conditions, that they have desirable asymptotic properties. In our context, these estimators only apply to cases where $q(Y)$ in (1) is equal to 0. In this paper, we adopt a more practical point of view, similar to the seminal work of Hastie and Tibshirani (1990).

The remainder of the paper proceeds as follows. In Section 2, we show how to estimate $\boldsymbol{\gamma}(\mathbf{X}) = \{\gamma_1(X_1), \dots, \gamma_p(X_p)\}$ using linear smoothers and a backfitting algorithm. In Section 3, we present three estimators for μ derived from three types of estimating functions: simple inverse weighted, doubly robust, and orthogonal. In this section, we also discuss how to estimate the standard errors for each of these estimators. In Section 4, we present an example of how our methodology can be used to analyze data. In Section 5, we present the results of a simulation study, designed to help us understand the finite-sample operating characteristics of our three estimators and how these characteristics are impacted by the choice of smoothing parameters. In Section 6, we describe a data-analytic procedure for choosing smoothing parameters. The final section is devoted to a summary and discussion. Mathematical proofs have been incorporated into a technical report, which is available at <http://www.biostat.jhsph.edu/~dscharf>. The Splus code used in our simulations is available following the software link at <http://www.biostat.jhsph.edu/~ririzarr>.

2 Estimation of $\boldsymbol{\gamma}(\mathbf{X})$

A straight forward approach to estimating $\boldsymbol{\gamma}(\mathbf{X})$ is to use a parametric model such as cubic splines, in which case the work of RRS can be directly applied. However, this is not practical in situations where we have many covariates as we need to choose the number of knots *and* their location for each covariate. In this paper we present a more practical approach that permits use of

linear smoothers. We can, for example, use smoothing splines in which case we only need to choose a smoothing parameter for each covariate.

Under model (1,2), it can be shown (see technical report) that the functions $\boldsymbol{\gamma}(X) = \{\gamma_1(X_1), \dots, \gamma_p(X_p)\}$ must satisfy

$$\gamma_j(X_j) = \text{logit}\{E[1 - R|X_j]\} - \log \left\{ E \left[\exp \left\{ \sum_{k \neq j} \gamma_k(X_k) + q(Y) \right\} \mid R = 1, X_j \right] \right\} \quad (3)$$

for $j = 1, \dots, p$. Our algorithm will provide estimates of $\gamma_j(X_j)$ only at the X_j 's of subjects with complete data. As we will see in the next section, this is sufficient for estimation of μ .

Equality (3) suggests the use of a backfitting algorithm. In motivating the elements of our backfitting algorithm, it is useful to let $\boldsymbol{\gamma}_{-j}(\mathbf{X}) = \{\gamma_k(X_k) : k \neq j\}$, $\gamma_{j1}(X_j) = \text{logit}\{E[1 - R|X_j]\}$ and $\gamma_{j2}(X_j) = \log\{E[\exp\{\sum_{k \neq j} \gamma_k(X_k) + q(Y)\} | R = 1, X_j]\}$. With this notation, $\gamma_j(X_j) = \gamma_{j1}(X_j) - \gamma_{j2}(X_j)$.

The first step of our algorithm is to obtain an unbiased estimate $\hat{\gamma}_{j1}(X_{j,i})$ of $\gamma_{j1}(X_{j,i})$. We write

$$\hat{\gamma}_{j1}(X_{j,i}) \approx \gamma_{j1}(X_{j,i}) + \epsilon_{j1,i} \quad (4)$$

where $E[\epsilon_{j1,i} | X_{j,i}] = 0$ for all $i = 1, \dots, n$. For example, one could use the adjusted dependent variable iterative smoothing procedure of Hastie and Tibshirani (1990). That is, for each $j = 1, \dots, p$, one can fit a generalized additive model with a logistic link that smooths with respect to X_j with degrees of freedom $df_{\gamma_{j1}}$. At this stage we are not interested in estimates with small variance but rather a functional of the data for which (4) holds. We therefore use smoothing splines with degrees of freedom $df_{\gamma_{j1}}$ large enough so that it is reasonable to believe that (4) holds. In our data analysis and simulations, we used $df_{\gamma_{j1}} = 24$. Notice that the estimate $\hat{\gamma}_{j1}(X_j)$ will be a very rough function of X_j .

Considering $\boldsymbol{\gamma}_{-j}$ as known, we define $\omega_{j2}(\mathbf{X}, Y; \boldsymbol{\gamma}_{-j}) = \exp\{\sum_{k \neq j} \gamma_k(X_k) + q(Y)\}$. By a Taylor

series expansion, we note that

$$\log\{\omega_{j2}(\mathbf{X}, Y; \boldsymbol{\gamma}_{-j})\} \approx \gamma_{j2}(X_j) + \frac{\omega_{j2}(\mathbf{X}, Y; \boldsymbol{\gamma}_{-j}) - E[\omega_{j2}(\mathbf{X}, Y; \boldsymbol{\gamma}_{-j})|X_j, R = 1]}{E[\omega_{j2}(\mathbf{X}, Y; \boldsymbol{\gamma}_{-j})|X_j, R = 1]}$$

Thus,

$$\log\{\omega_{j2}(\mathbf{X}_i, Y_i; \boldsymbol{\gamma}_{-j})\} \approx \gamma_{j2}(X_{j,i}) + \epsilon_{j2,i} \quad (5)$$

where $E[\epsilon_{j2,i}|X_{j,i}] = 0$, for all $i = 1, \dots, n_1$. Together, (4) and (5) imply that

$$\hat{\gamma}_{j1}\{(X_{j,i}) - \log(\omega_{j2}(\mathbf{X}_i, Y_i; \boldsymbol{\gamma}_{-j}))\} \approx \gamma_j(X_{j,i}) + \epsilon_{j,i} \quad (6)$$

where $E[\epsilon_{j,i}|X_{j,i}] = 0$, for all $i = 1, \dots, n_1$. Since $\boldsymbol{\gamma}_{-j}$ is unknown, we use estimates from the previous iteration of the backfitting algorithm to estimate $\omega_{j2}(\mathbf{X}_i, Y_i; \boldsymbol{\gamma}_{-j})$.

Notice that the “adjusted dependent data” $\hat{\gamma}_{j1}(X_{j,i}) - \log\{\omega_{j2}(\mathbf{X}_i, Y_i; \boldsymbol{\gamma}_{-j})\}$ has the form: smooth function of X_j plus noise. Thus, a natural estimate $\hat{\gamma}_j(X_{j,i})$ is obtained by applying a linear smoother that yields an appropriate amount of smoothness. For $j > 1$, we normalize the current estimate so that it has mean zero. Then we iterate the algorithm until there is not much change in the estimates.

Let $\hat{\boldsymbol{\gamma}}_{-j}^{(l)}(\mathbf{X}) = \{\hat{\gamma}_1^{(l)}(X_1), \dots, \hat{\gamma}_{j-1}^{(l)}(X_{j-1}), \hat{\gamma}_{j+1}^{(l-1)}(X_{j+1}), \dots, \hat{\gamma}_p^{(l-1)}(X_p)\}'$, where $\hat{\gamma}_j^{(l)}(X_j)$ is the estimate of $\gamma_j(X_j)$ on the l th iteration of our backfitting algorithm. Let $\hat{\boldsymbol{\gamma}}(\mathbf{X}) = \{\hat{\gamma}_1(X_1), \dots, \hat{\gamma}_p(X_p)\}'$ denote our final estimator of $\boldsymbol{\gamma}(\mathbf{X})$. Let $\mathbf{1}_k$ be a vector of ones of length k . We now present the second step of our algorithm in detail:

0. Set $\hat{\gamma}_2^{(0)} = \dots = \hat{\gamma}_p^{(0)} = 0$. Set $l = 1$.

1. Sequentially in $j = 1, \dots, p$, smooth the “adjusted dependent data”, $\{\hat{\gamma}_{j1}(X_{j,i}) - \log\{\omega_{j2}(\mathbf{X}_i, Y_i; \hat{\boldsymbol{\gamma}}_{-j})\} : i = 1, \dots, n_1\}$, using a cubic smoothing spline with specified degrees of freedom (df_{γ_j}), to yield an estimate $\hat{\gamma}_j^{(l)}(X_{j,i})$ for subjects with $R_i = 1$. For $j > 1$, center the estimate by subtracting an estimate of its mean (set $\hat{\gamma}_j^{(l)}(X_{j,i}) = \hat{\gamma}_j^{(l)}(X_{j,i}) - \hat{m}_j^{(l)}$) and absorb the estimated mean into $\hat{\gamma}_1^{(l)}(X_{1,i})$ (set $\hat{\gamma}_1^{(l)}(X_{1,i}) = \hat{\gamma}_1^{(l)}(X_{1,i}) + \hat{m}_j^{(l)}$), where $\hat{m}_j^{(l)} =$

$\frac{1}{n} \sum_{i=1}^n R_i \hat{\gamma}_j^{(l)}(X_{j,i}) (1 + \exp\{\mathbf{1}'_{p-1} \hat{\boldsymbol{\gamma}}_{-j}^{(l)}(\mathbf{X}_i) + \hat{\gamma}_j^{(l)}(X_{j,i}) + q(Y_i)\})$ (This estimate of the mean is derived from the fact that $E[\gamma_j(X_j)] = E[R\gamma_j(X_j)(1 + \exp\{\mathbf{1}'_p \boldsymbol{\gamma}(\mathbf{X}) + q(Y)\})]$).

2. If $\frac{\sum_{j=1}^p \|\hat{\gamma}_j^{(l)} - \hat{\gamma}_j^{(l-1)}\|^2}{\sum_{j=1}^p \|\hat{\gamma}_j^{(l-1)}\|^2} < \epsilon$ (where ϵ is some small number), then stop and let $\hat{\gamma}_j(X_{j,i}) = \hat{\gamma}_j^{(l)}(X_{j,i})$ for $i = 1, \dots, n_1$. Otherwise set $l = l + 1$ and repeat steps 1 and 2.

In our above algorithm, we used smoothing splines as an example of a linear smoother to estimate one-dimensional conditional expectations. Notice that one could have used any linear smoother, such kernel, loess, wavelets, etc.. However, for simplicity in this paper, we only use smoothing splines. Since in Splus, the smoothness parameter is controlled by the degrees of freedom option, we use df as the smoothness parameter. Since there exists a one-to-one correspondence between df and the smoothness parameter usually associated with smoothing splines (see Hastie and Tibshirani, 1990), there is no ambiguity.

By allowing $df_{\gamma_{1j}}$ and df_{γ_j} to increase with the sample size, we conjecture that, under regularity conditions, $\hat{\boldsymbol{\gamma}}(\mathbf{X})$ can be shown to converge in probability to $\boldsymbol{\gamma}(\mathbf{X})$ (uniformly in \mathbf{X}) at rate $n^{1/4+\epsilon}$ for some $\epsilon > 0$. It is this rate that is critical for our estimators of μ in the next section to have \sqrt{n} rates of convergence (Newey, 1990). Following the ideas in non-parametric statistics, one should be able to specify an appropriate set of assumptions to make the the degrees of freedom big enough so that the distance between the true $\boldsymbol{\gamma}(\cdot)$ and the expected value of our estimator $\hat{\boldsymbol{\gamma}}(\cdot)$ gets arbitrarily “small,” while the number of data points grows fast enough so that the variance of our estimator decreases to zero. Usually by imposing assumptions on the smoothness of the true functions, we can achieve any rate n^q for $0 < q < 1/2$ (e.g., Stone, 1986; Linton and Härdle, 1996; Kauermann and Opsomer, 2002). Different smoothness assumptions yield different q . Thus, we conjecture that there exist a set of smoothness assumptions as well as a data adaptive procedure for choosing the degrees of freedoms so that we can achieve rates with $q > 1/4$.

Asymptotic theory typically provides little guidance as to the choice of degrees of freedom in moderate sized datasets. As a result, data analysts either perform sensitivity analysis with respect to the degrees of freedom or use a model selection procedure. In Section 6, we propose one such procedure.

Finally, it is important to note that setting $df_{\gamma_j} = 1$ for all $j = 1, \dots, p$ is equivalent to assuming that $\gamma_1(X_1) + \dots + \gamma_p(X_p)$ is a linear combination of X_1, \dots, X_p (i.e., the fully parametric approach of RRS). Thus, our algorithm can be used to assess how inference changes as one increases the flexibility of the selection model.

3 Estimation of μ

To estimate μ , we consider the class of augmented inverse probability weighted estimating functions for μ . Throughout, model (1) is assumed to hold. To ease notation, define $\pi(\mathbf{X}, Y; \boldsymbol{\gamma}) = 1/[1 + \exp\{\mathbf{1}'_p \boldsymbol{\gamma}(\mathbf{X}) + q(Y)\}]$. Furthermore, define

$$U(O; \mu; \boldsymbol{\gamma}; \phi) = \frac{R}{\pi(\mathbf{X}, Y; \boldsymbol{\gamma})}(Y - \mu) + \left\{1 - \frac{R}{\pi(\mathbf{X}, Y; \boldsymbol{\gamma})}\right\} \phi(\mathbf{X}; \mu; \boldsymbol{\gamma}) \quad (7)$$

where $\phi(\mathbf{X}; \mu; \boldsymbol{\gamma})$ is some specified function of \mathbf{X} and μ and $\boldsymbol{\gamma}$. This class of augmented inverse weighted estimating functions has the following properties (see technical report for proofs):

Property 1:

Regardless of the choice of $\phi(\mathbf{X}; \mu; \boldsymbol{\gamma})$, $U(O; \mu; \boldsymbol{\gamma}; \phi)$ has mean zero, provided model (2) is correctly specified.

Property 2:

If

$$\phi(\mathbf{X}; \mu; \boldsymbol{\gamma}) = E[Y \exp\{q(Y)\} | R = 1, \mathbf{X}] / E[\exp\{q(Y)\} | R = 1, \mathbf{X}] - \mu, \quad (8)$$

then $U(O; \mu; \boldsymbol{\gamma}; \phi)$ has mean zero even if model (2) is incorrectly specified.

Property 3:

If $\phi(\mathbf{X}; \mu; \boldsymbol{\gamma})$ is chosen so that

$$E[\{Y - \mu - \phi(\mathbf{X}; \mu; \boldsymbol{\gamma})\} \exp\{\mathbf{1}'_p \boldsymbol{\gamma}(\mathbf{X}) + q(Y)\} | R = 1, X_j] = 0 \text{ for all } j = 1, \dots, p \quad (9)$$

then, $U(O; \mu; \gamma; \phi)$ is orthogonal to the nuisance tangent space (Newey, 1992) for the nuisance parameters $\gamma_1, \dots, \gamma_p$.

These facts suggest that we estimate μ as the solution to

$$S(\mathbf{O}; \mu; \hat{\gamma}; \hat{\phi}) = \frac{1}{n} \sum_{i=1}^n U(O_i; \mu; \hat{\gamma}; \hat{\phi}) = 0 \quad (10)$$

where $\hat{\phi}$ is a consistent estimator of targeted function ϕ . We now consider some special cases.

3.1 Simple Inverse Weighted Estimator

The simplest estimator is to choose $\phi = 0$, in which case $\hat{\phi} = 0$ as well. In this situation, the estimator of μ can be solved in closed form as

$$\hat{\mu}_{IW} = \hat{E} \left[\frac{RY}{\pi(\mathbf{X}, Y; \hat{\gamma})} \right] / \hat{E} \left[\frac{R}{\pi(\mathbf{X}, Y; \hat{\gamma})} \right]. \quad (11)$$

In this estimator, the observed values of the outcomes are inverse weighted by the estimated probability of being observed. The denominator converges in probability to one. It can be shown that the resulting estimator is consistent and asymptotically normal, provided model (2) is correctly specified (see Property 1 above). It is difficult to derive the influence function for $\hat{\mu}_{IW}$, but we can estimate its standard error by using non-parametric bootstrap.

3.2 Doubly Robust Estimator

The second estimator is motivated from Properties (1) and (2) above. To estimate the targeted $\phi(\mathbf{X}; \mu; \gamma)$ in Property 2, we need to estimate $E[Y \exp\{q(Y)\} | R = 1, \mathbf{X}] / E[\exp\{q(Y)\} | R = 1, \mathbf{X}]$. Due to the curse of dimensionality, this cannot be done non-parametrically. To proceed, we place enough lower-dimensional restrictions (i.e, parametric or semi-parametric) on the conditional distribution of Y given \mathbf{X} and $R = 1$ so that the above ratio of expectations can be well estimated in finite samples. Let $f(Y | \mathbf{X}, R = 1; \rho)$ be a lower-dimensional model for the conditional distribution

of Y given \mathbf{X} and $R = 1$, indexed by parameter ρ . Let $\hat{\rho}$ be a consistent estimator of ρ , which behaves well in finite samples. Then, we can estimate the targeted $\phi(\mathbf{X}; \mu; \gamma)$ by

$$\hat{\phi}(\mathbf{X}; \mu; \gamma) = \frac{\int y \exp\{q(y)\} f(y|R = 1, \mathbf{X}; \hat{\rho}) dy}{\int \exp\{q(y)\} f(y|R = 1, \mathbf{X}; \hat{\rho}) dy} - \mu$$

Example: In our simulations and data analysis, we assumed that

$$\log(Y) = \rho_1(X_1) + \dots + \rho_p(X_p) + \nu$$

where $\rho_1(X_1)$ is an unknown smooth function of X_1 and $\rho_j(X_j)$, for $j = 2, \dots, p$, are unknown, smooth, mean zero (given $R = 1$) functions of X_j , and, given $R = 1$ and \mathbf{X} , ν is normally distributed with mean 0 and variance ρ_{p+1}^2 . Here $\rho = (\rho_1, \dots, \rho_p, \rho_{p+1})$, where the ρ_1, \dots, ρ_p 's can be estimated by $\hat{\rho}_1, \dots, \hat{\rho}_p$ using an additive model backfitting algorithm and smoothing splines with specified degrees of freedom (df_{ρ_j}). The variance ρ_{p+1}^2 can be estimated by $\hat{\rho}_{p+1}^2$, the residual sums of squares divided by the “degrees of freedom for error” (see Hastie and Tibshirani, 1990). As with the degrees of freedom in estimating the $\gamma(\mathbf{X})$, df_{ρ_j} must increase with the sample size (at an appropriate rate) in order for the resulting estimators to be consistent. Again, the asymptotic theory is not very useful in moderate sample sizes, so we recommend sensitivity analysis or model selection. Note that $df_{\rho_j} = 1$ for all $j = 1, \dots, p$ is equivalent to assuming that $E[\log(Y)|R = 1, \mathbf{X}]$ is linear in the X_j 's (i.e., a fully parametric model for the conditional distribution of Y given $R = 1$ and \mathbf{X}).

Plugging $\hat{\phi}(\mathbf{X}; \mu, \gamma)$ into Equation (10), we obtain a closed form estimator for μ :

$$\hat{\mu}_{DR} = \frac{1}{n} \sum_{i=1}^n \left[\frac{R_i Y_i}{\pi(\mathbf{X}_i, Y_i; \hat{\gamma})} + \left\{ 1 - \frac{R_i}{\pi(\mathbf{X}_i, Y_i; \hat{\gamma})} \right\} \frac{\int y \exp\{q(y)\} f(y|R = 1, \mathbf{X}_i; \hat{\rho}) dy}{\int \exp\{q(y)\} f(y|R = 1, \mathbf{X}_i; \hat{\rho}) dy} \right]$$

This estimator is called doubly robust since it will be consistent and asymptotically normal (CAN) if model (2) *or* the lower dimensional model for the conditional distribution of Y given \mathbf{X} and $R = 1$ is correctly specified (see Robins, 1999; Scharfstein, Rotnitzky, and Robins, 1999; Robins, Rotnitzky, and Scharfstein, 2000; Robins and Rotnitzky, 2001). If both models are correct, then

the estimator will be semi-parametric efficient (i.e., it will be a semiparametric estimator in the sense that it will be CAN under model (2) and attain the efficiency bound when the the model for the conditional distribution of Y given \mathbf{X} and $R = 1$ is correctly specified). The influence function for $\hat{\mu}_{DR}$ when both models are correct is $U(O; \mu; \gamma; \phi)$. In this case, the asymptotic variance of $\hat{\mu}_{DR}$ is $E[U(O; \mu; \gamma; \phi)^2]$ and it can be consistently estimated by $\frac{1}{n} \sum_{i=1}^n U(O_i; \hat{\mu}_{DR}; \hat{\gamma}; \hat{\phi})^2$. When either model (2) or the model for the conditional distribution of Y given \mathbf{X} and $R = 1$ fails to hold, the influence function becomes difficult to compute, in which case we recommend non-parametric bootstrap.

3.3 Orthogonal Estimating Function Estimator

Our final estimator is motivated by Properties 1 and 3 above and is one for which we can derive its influence function when model (2) is correctly specified. When ϕ is chosen so that $U(O; \mu; \gamma; \phi)$ is orthogonal to the nuisance tangent space for γ , Newey (1992) and Robins, Mark, and Newey (1994) showed that the estimator for μ based on solving (10) will have influence function $U(O; \mu; \gamma; \phi)$, provided that the estimators for γ are estimated at rate $n^{1/4+\epsilon}$ (for some $\epsilon > 0$) and $\hat{\phi}$ is a consistent estimator of ϕ . In Section 2, we conjectured that, under regularity conditions, our estimator of γ converges at the above rate.

Property 3 above tells us that ϕ must satisfy the constraint (9). There are an infinite number of solutions to (9), which are difficult to characterize in general. With this in mind, we consider the special case where we impose the additional assumption that

$$\phi(\mathbf{X}; \mu; \gamma) = \sum_{j=1}^p \phi_j(X_j; \mu; \gamma)$$

where for given μ, γ , $\phi_1(X_1; \mu; \gamma)$ is a function of X_1 and $\phi_j(X_j; \mu; \gamma)$ are mean zero functions of X_j , for $j = 2, \dots, p$. With this constraint, we have that

$$\phi_j(X_j; \mu; \gamma) = \frac{E[\{Y - \mu - \sum_{k \neq j} \phi_k(X_k; \mu; \gamma)\} \exp\{\mathbf{1}'_p \gamma(\mathbf{X}) + q(Y)\} | R = 1, X_j]}{E[\exp\{\mathbf{1}'_p \gamma(\mathbf{X}) + q(Y)\} | R = 1, X_j]} \quad (12)$$

for all $j = 1, \dots, p$. This suggests a weighted backfitting algorithm with weights $w = \exp\{\gamma_1(X_1) + \dots + \gamma_p(X_p) + q(Y)\}$ (Hastie and Tibshirani 1990, Chapter 6). Let $\hat{\phi}_j^{(l)}(X_j)$ denote the estimate of $\phi_j(X_j)$ on the l th iteration of the backfitting algorithm (for ease of notation, we suppress the dependence of ϕ_j on the fixed parameters μ, γ):

0. Set $\hat{\phi}_2^{(0)} = \dots = \hat{\phi}_p^{(0)} = 0$. Set $l = 1$.

1. Sequentially in $j = 1, \dots, p$, let

$$\hat{\phi}_j^{(j)}(X_j) = \hat{E} \left[\left\{ Y - \mu - \sum_{k=1}^{j-1} \hat{\phi}_k^{(l)}(X_k) - \sum_{k=j+1}^p \hat{\phi}_k^{(l-1)}(X_k) \right\} \mid R = 1, X_j \right]$$

where for some function $g(\mathbf{X}, Y)$, $\hat{E}[g(\mathbf{X}, Y) \mid R = 1, X_j]$ is a weighted smooth of the “data”, $\{g(\mathbf{X}_i, Y_i) : i = 1, \dots, n_1\}$ obtained using a weighted smoothing spline procedure with specified degrees of freedom (df_{ϕ_j}), to yield estimates $\hat{\phi}_j^{(l)}(X_{j,i})$ for all subjects. For $j > 1$, center the estimate by subtracting off an estimate of its mean (set $\hat{\phi}_j^{(l)}(X_{j,i}) = \hat{\phi}_j^{(l)}(X_{j,i}) - \hat{m}_j^{(l)}$) and absorb the estimated mean into $\hat{\phi}_1^{(l)}(X_{1,i})$ (set $\hat{\phi}_1^{(l)}(X_{1,i}) = \hat{\phi}_1^{(l)}(X_{1,i}) + \hat{m}_j^{(l)}$), where $\hat{m}_j^{(l)} = \frac{1}{n} \sum_{i=1}^n R_i \phi_j^{(l)}(X_{j,i}) (1 + \exp\{\mathbf{1}'_p \boldsymbol{\gamma}(\mathbf{X}_i) + q(Y_i)\})$ (This estimate of the mean is derived from the fact that $E[\phi_j(X_j)] = E[R \phi_j(X_j) (1 + \exp\{\mathbf{1}'_p \boldsymbol{\gamma}(\mathbf{X}) + q(Y)\})]$).

2. If $\frac{\sum_{j=1}^p \|\hat{\phi}_j^{(l)} - \hat{\phi}_j^{(l-1)}\|^2}{\sum_{j=1}^p \|\hat{\phi}_j^{(l-1)}\|^2} < \epsilon$ (where ϵ is some small number), then stop and let $\hat{\phi}_j(X_{j,i}) = \hat{\phi}_j^{(l)}(X_{j,i})$ for $i = 1, \dots, n_1$. Otherwise set $l = l + 1$ and repeat steps 1 and 2.

Since there is no reason to believe that the solution to (12) should be very smooth, we suggest taking df_{ϕ_j} large. In our data analysis and simulations, we set $df_{\phi_j} = 12$. As with the previous two estimation procedures involving degrees of freedom, we also conjecture that as the degrees of freedom df_{ϕ_j} increase with the sample size (at an appropriate rate), the estimator $\hat{\phi}(\mathbf{X}; \mu; \boldsymbol{\gamma}) = \sum_{j=1}^p \hat{\phi}_j(\mathbf{X}; \mu; \boldsymbol{\gamma})$ can be shown to be a consistent estimator of the unique solution to (12).

We now solve (10), with $\hat{\phi}(\mathbf{X}; \mu; \boldsymbol{\gamma})$, using an iterative Newton-Raphson algorithm. We denote the solution $\hat{\mu}_{ORTH}$. The influence function for $\hat{\mu}_{ORTH}$ is $U(O; \mu; \boldsymbol{\gamma}; \phi)$. The asymptotic variance of $\hat{\mu}_{ORTH}$ is $E[U(O; \mu; \boldsymbol{\gamma}; \phi)^2]$ and it can be consistently estimated by $\frac{1}{n} \sum_{i=1}^n U(O_i; \hat{\mu}_{DR}; \hat{\boldsymbol{\gamma}}; \hat{\phi})^2$.

4 Analysis of ACTG 175

To illustrate our approach, we reanalyze data from the AIDS Clinical Trial Group (ACTG) 175 (Hammer *et al.*, 1996), which was also the motivating example behind the sensitivity analysis approach developed by Scharfstein *et al.* (1999). To review, ACTG 175 was a randomized, double blind trial designed to evaluate nucleotide monotherapy vs. combination therapy in HIV infected individuals with CD4 counts between 200 and 500/mm³. 2,467 subjects were randomized to one of four treatment arms: 619 to AZT (600 mg/day) alone, 613 to AZT (600 mg/day) + ddI (400 mg/day), 615 to AZT (600 mg/day) + ddC (2.25 mg/day), and 620 to ddI (400 mg/day) alone (Hammer *et al.*, 1996). CD4 counts were scheduled to be collected at baseline, week 8, and then every 12 weeks thereafter. Additional baseline characteristics were also collected, including age and IV drug use. In the interest of space, we will only consider the AZT+ddI treatment arm.

One goal of the investigators was to estimate the treatment-specific means of CD4 cell count at week 56 had all subjects remained on their assigned treatment through that week. The analytic objective is complicated by the presence of missing values at week 56, which arise due to loss to follow-up, skipped clinic visits, and discontinuation of assigned therapy. Our analysis will not distinguish between the multiple causes of missingness. Here, we define a completer as a subject who stays on treatment and is measured at week 56; all others are called drop-outs. The percent of drop-out in the AZT+ddI arm was 33.6%. Among completers, the mean CD4 at week 56 is 384.96, with an associated standard error of 8.53. The above estimate of the targeted mean is only valid if the completer and drop-outs are similar on measured and unmeasured risk factors (i.e., random drop-out or missing completely at random). This latter assumption is unlikely to hold, as it is well known from other studies that completers tend to be very different than drop-outs.

In our analysis, we let Y be the CD4 count at week 56 under full compliance. This outcome is observed when the subject is a completer and thus we let R be the completion indicator. In addition, we use three predictors ($p = 3$) in the selection model (2) with $X_1 =$ IV Drug User at Baseline, $X_2 =$ Age at Baseline, and $X_3 =$ CD4 at Baseline. Throughout the analysis, we used smoothing splines with $df_{\gamma_2} = df_{\gamma_3} = df_{\gamma}$, $df_{\rho_2} = df_{\rho_3} = df_{\rho}$, $df_{\phi_2} = df_{\phi_3} = 12$ to estimate γ_j , ρ_j and ϕ_j , for

$j = 2, 3$. Also, recall from Section 2 that we set $df_{\gamma_{j1}} = 24$. Since X_1 is a binary factor we did not need smoothing splines to estimate its effect, instead we estimated it in the usual manner.

In conducting a sensitivity analysis of these data, it is useful to restrict attention to a simple class of selection bias functions $q(Y)$, which includes missing at random. With this in mind, we parameterized the selection bias function $q(Y)$ in (1) using an interpretable, practically non-identifiable, one-dimensional parameter (α) and obtained estimates of μ as we varied α over a plausible range. It is important to emphasize the importance of choosing a parameterization that is clear and meaningful to subject matter experts so that they can encode their beliefs. Inference will be sensitive to the choice of parameterization, but it is important to recognize that the aim of sensitivity analysis is not to find the “truth,” but to report an analysis which reasonably reflects an expert’s beliefs about selection bias. Experts may find it useful to consider multiple parameterizations and evaluate the stability of the final inferences. In considering alternative parameterizations, it is important to recognize that the selection bias parameters will have different interpretations. As a result, it may be tricky for experts to specify “comparable” ranges for the alternative set of parameters (see Section 7.2.3 of Scharfstein *et al.*, 1999) for a related discussion). As a start, it may prove useful to normalize (e.g., by a location-scale change) the alternative selection bias functions so that they cover the same range over the range of observed Y ’s. Depending on the relative shape of these functions, experts may be more willing to consider a similar range for the selection bias parameters.

For illustrative purposes, we let $q(Y) = \alpha \log(Y)$, where α is interpreted as the log odds ratio of drop-out for subjects who differ by one unit in $\log(Y)$. Here, $\alpha = 0$ is equivalent to missing at random (i.e., drop-out only depends on measured risk factors); $\alpha \neq 0$ is equivalent to non-ignorable missingness (i.e., drop-out depends on potentially unmeasured factors). $\alpha < 0$ (> 0) indicates that, above and beyond measured factors sicker (healthier) subjects are more likely to drop out. For example, $\alpha = -0.5(0.5)$ indicates that a k -fold ($k > 1$) decrease in CD4 at week 56 leads to a $k^{0.5}$ -fold increase (decrease) in the odds of dropping out.

In Table 1, we present the three different estimators of μ (simple inverse weighted - SIW;

doubly robust - DR; orthogonal - ORTH) as a function of the smoothing parameters df_γ and df_ρ for five levels of the selection bias parameter $\alpha = -1.0, -0.5, 0.0, 0.5, 1.0$. In the table, we present bootstrapped standard errors (based on 250 resamples) as well as analytic standard errors for the ORTH estimator. On a SUN Enterprise 6500, the slowest time (over varying choices of degrees of freedom and α) to compute the SIW, DR, and ORTH (including standard errors) estimators was 6, 7, and 9 seconds, respectively. Note that the analytic standard errors are comparable to the bootstrapped standard errors, except at large α 's where they tend to be slightly larger. This may be due to issues of model misspecification or finite samples. In the table, the bold values denote the estimator-specific “optimal” choice of degrees of freedom based on a model selection criteria discussed in Section 6.

In terms of the estimates themselves, we first notice that the uncertainty in the choice of α has a much greater influence on the magnitude of the resulting estimate as compared to the choice of estimator and associated degrees of freedom. Within levels of α , we see that the DR estimates are relatively insensitive to df_γ , but much more sensitive to the df_ρ . For $\alpha < 0$, the SIW estimate is lower and has greater variability than the DR (with $df_\rho > 1$) and ORTH estimates. For $\alpha \geq 0$, the estimates are much more comparable. For all α , it is interesting to note that the DR (with $df_\rho > 1$) and ORTH estimates are quite comparable in both magnitude and variability. Given the robust nature of the DR estimator and the fact that one does not pay a price in terms of standard error, we prefer using this estimator.

In Figures 1a-1d, we present estimates of $\gamma_2(X_2)$, $\gamma_3(X_3)$, $\rho_2(X_2)$, and $\rho_3(X_3)$ for various choices of df_γ and df_ρ . Here α is assumed equal to -0.5 . In Figures 1a and 1b, we see that the linear selection model ($df_\gamma = 1$) suggests that the probability of non-compliance tends to decrease sharply with baseline age and increase slightly with baseline CD4. With higher degrees of freedom, we see a slight non-linearity. Specifically, with $df_\gamma = 3$, the probability of non-compliance appears to attenuate for the very young and very old and tends to be lowest in the 2nd quartile of baseline CD4 and highest in the 3rd and 4th quartiles. With $df_\gamma = 6$, we start to see exaggerated non-linearities, suggestive of overfitting. However, there does not appear to be strong evidence against a linear

model. This is in agreement with the fact that the estimates of μ obtained with $df_\gamma = 1$ and $df_\gamma = 3$ are quite similar.

In Figures 1c and 1d, we see that the linear model ($df_\rho = 1$) for the conditional mean of $\log(Y)$ among compliers suggests that the mean increases slightly with age and increases sharply with baseline CD4. With $df_\rho = 3, 6$, the effect of age is similar to that of the linear model and the effect of baseline CD4 becomes attenuated at high levels of baseline CD4. In this case there does seem to be an important difference between the $df_\rho = 1$ and $df_\rho > 1$ which results in a difference in the estimates obtained for μ . The model selection procedure presented in Section 6 chooses $df_\rho > 1$.

5 Simulation Study

For each of 500 simulated datasets, we generated data for 613 subjects. For each subject, we generated two continuous baseline covariates X_1 and X_2 , where

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \text{Normal} \left(\begin{bmatrix} \mu_{X_1} \\ \mu_{X_2} \end{bmatrix}, \begin{bmatrix} \sigma_{X_1}^2 & \tau \sigma_{X_1} \sigma_{X_2} \\ \tau \sigma_{X_1} \sigma_{X_2} & \sigma_{X_2}^2 \end{bmatrix} \right) \quad (13)$$

$\mu_{X_1} = 0$, $\mu_{X_2} = 0$, $\sigma_{X_1} = 10$, $\sigma_{X_2} = 100$, and $\tau = -0.25$. We assume model (1,2) for the conditional distribution of R given \mathbf{X} and Y with $q(Y) = \alpha \log(Y)$. The true values of $\gamma_1(X_1)$ and $\gamma_2(X_2)$ are depicted by the curve of circles in Figures 2a and 2b, respectively. These curves are clearly non-linear and are well approximated by a smoothing spline with degrees of freedom of at least 3. For illustrative purposes, we chose $\alpha = 1.0$. We also assume that the conditional distribution of $\log(Y)$ given \mathbf{X} and $R = 1$, $f(y|\mathbf{X}, R = 1)$, follows a normal distribution with mean $\rho_1(X_1) + \rho_2(X_2)$ and variance ρ_3^2 , where $\rho_3 = 0.05$ and $\rho_1(X_1)$ and $\rho_2(X_2)$ are depicted by the curve of circles in Figures 2c and 2b, respectively. These curves are also non-linear and are well approximated by a smoothing spline with degrees of freedom of at least 12. Given these models, we then generated Y

from the conditional distribution of Y given \mathbf{X} , which has cumulative distribution function

$$F(y|\mathbf{X}) = \frac{\int \frac{1}{\pi(\mathbf{X},w;\boldsymbol{\gamma})} I(w \leq y) f(w|\mathbf{X}, R = 1) dw}{1 + c(\mathbf{X}) \exp\{\gamma_1(X_1) + \gamma_2(X_2)\}}$$

where $c(X) = \int \exp\{\alpha \log(y)\} f(y|\mathbf{X}, R = 1) dy$. Generation proceeded using the inverse cumulative distribution function technique. Finally, we drew R from its conditional distribution given \mathbf{X} and Y . In this simulation, the true μ was 226.3 and the true proportion of missingness was 30%. Throughout the simulation, we assumed $df_{\gamma_1} = df_{\gamma_2} = df_{\gamma}$, $df_{\rho_1} = df_{\rho_2} = df_{\rho}$, $df_{\phi_1} = df_{\phi_2} = 12$ and $df_{\gamma_{j1}} = 24$.

In Figures 2a and 2b, we present three sets of estimates of $\gamma_1(X_1)$ and $\gamma_2(X_2)$ for one particular realization of the simulation, respectively. The estimates correspond to various choices of df_{γ} . In both figures, the estimates for $df_{\gamma} = 12$ overfit the true $\gamma_1(X_1)$ and $\gamma_2(X_2)$ functions. For $\gamma_1(X_1)$, the choice of $df_{\gamma} = 1$ overestimates the truth at low values of X_1 , while $df_{\gamma} = 3$ does a relatively better job overall of estimating the truth. A similar pattern is observed for $\gamma_2(X_2)$. In Figures 2c and 2d, we present three sets of estimates of $\rho_1(X_1)$ and $\rho_2(X_2)$ for one particular realization of the simulation, respectively. The dots represent the resulting observed values of Y . The estimated curves correspond to various choices of df_{ρ} . In both figures, $df_{\rho} = 12$ does the best job of estimating the truth.

In Figures 3a - 3c, we present the bias, Monte Carlo standard error (MCSE), and mean squared error (MSE) of *i*) simple inverse weighted estimator (SIW), *ii*) doubly robust estimator with correct specification of the distributional form for the law of Y given \mathbf{X} and $R = 1$ (DR), *iii*) doubly robust estimator with incorrect specification of the distributional form of Y given \mathbf{X} and $R = 1$ (DRRW), and *iv*) orthogonal estimator (ORTH), as a function of df_{γ} . While Figure 3a shows the large bias for the observed data mean estimator, we have excluded this estimator from Figures 3b and 3c because the mean squared error will be at least 36 (bias squared), much larger than the competing estimators. For the DR estimator, we used $df_{\rho} = 12$. For the DRRW estimator, we assumed that the conditional distribution of Y given \mathbf{X} and $R = 1$ follows of gamma distribution

with $E[Y|\mathbf{X}, R = 1]$ linear in X_2 . All estimators used $\alpha = 1$. In Figure 3a, we see that the DR and ORTH estimators are unbiased for all df_γ . For DR, this is expected because the law of Y given \mathbf{X} and $R = 1$ is correctly specified. For ORTH, this is surprising. As expected, we see that for $df_\gamma \leq 2$, SIW and DRRW are positively biased and for $df_\gamma > 2$, they are unbiased. In Figures 3b and 3c, we see that DR and ORTHO have equivalent MCSE (MSE) and that they are lower than SIW and DRRW, for all df_γ . For DR, this is expected since it is efficient. Again, the results for ORTH are surprising. The MCSE (MSE) for SIW and DRRW are U-shaped functions of df_γ , with SIW uniformly dominating DRRW.

The reason why the SIW estimators using degrees of freedom that are too small are biased is because incorrect estimation of $\gamma(\mathbf{X})$ results in larger weights for observations that “don’t deserve it”. We looked closely at one simulated dataset with $df_\gamma = 1$ and $df_\gamma = 3$. We noticed that larger weights were obtained for $df_\gamma = 1$. This observation leads us to conjecture that controlling the degree of smoothing in generalized additive models may help in reducing the sensitivity of augmented inverse probability weighted estimating functions to skewness in the weights (Scharfstein *at al.*, 1999, Section 7.2.2).

We investigated the bias, MCSE, MSE of *i*) doubly robust estimator with correct specification of the law of R given \mathbf{X} and Y and the distributional form of the law of Y given \mathbf{X} and $R = 1$ (DR), *ii*) doubly robust estimator with incorrect specification of the distribution of R given \mathbf{X} and Y and correct specification of the distributional form of the law of Y given \mathbf{X} and $R = 1$ (DRWR). All estimators used $\alpha = 1$. For DRWR, we set $df_\gamma = 1$; for all other estimators we set $df_\gamma = 3$. We found that DR had slightly negative bias for $df_\rho < 12$, while DRWR had relatively severe positive bias for $df_\rho < 12$. For $df_\rho \geq 12$, the two estimators were unbiased (as expected). We found that the MCSE for DRWR was much higher than that of DR for $df_\rho < 6$, but the difference disappears for $df_\rho \geq 6$. We found that the MSE for DR increases monotonically with df_ρ . The DRWR estimator has U-shaped MSE: for $df_\rho < 6$, the MSE is quite poor, but improves substantially for $df_\rho \geq 6$. DRWR and DR have comparable MSE for $df_\rho \geq 12$.

To assess the performance of our analytic variance formula for the ORTH estimator, we com-

pared the average of the estimated standard error (ASE) to the MCSE as a function of df_γ . For df_γ ranging from 1 to 12, the ASE was between 2.4% and 3.0% smaller than the MCSE.

To evaluate the behavior of the estimates based on general additive models when in fact the data is generated with a linear model we performed another simulation. As before, for each of 500 simulated datasets, we generated data for 613 subjects. For each subject, we generated two continuous baseline covariates using (13) but this time with $\gamma_1(X_1)$, $\gamma_2(X_2)$, $\rho_1(X_1)$, and $\rho_2(X_2)$ were chosen as linear functions of X_1 and X_2 . We evaluated the the percentage growth of MSE for the DR estimator of using different degrees of freedom (df_γ ranging from 2 to 12 and df_ρ ranging from 2 to 8) as opposed to fitting a linear model ($df_\gamma = df_\rho = 1$). The percentage growth (not shown) was relatively insensitive to the choice of df_γ . The maximum percentage growth was 0.6% which occurred at $df_\rho = 8$. This suggests that in situations where one is not sure the correct model is linear one should use general additive models.

In summary, our simulations support the use of the doubly robust and orthogonal estimators. With these alternatives, the simple inverse weighted estimator should not be used. The choice of degrees of freedom can be critically important. Making general comments about choosing the degrees of freedom to use in practical situations is difficult. In the next section, we consider one criterion for selecting the smoothing parameters. There we conduct a small simulation study to see how well the criterion selects the correct degrees of freedom.

6 Choosing Smoothing Parameters

For our three estimators above, the degrees of freedom df_{γ_j} and df_{ρ_j} must be specified. While researchers may have some sense of the underlying “smoothness” of γ_j and ρ_j , it is useful to have an aid in choosing their associated degrees of freedom. Ideally, one would like to choose these parameters automatically by minimizing some penalized objective function (Lin and Zhang, 1999; Shively, Kohn and Wood, 1999; Wood, 2000). While one piece of an objective function could be a quadratic form based on our estimating functions for μ , it is not clear how, in our problem, to add in a penalty term to control the degrees of smoothing. Instead, we use the model selection

technique of Pan (2001). Let \mathbf{df} be the vectors of degrees of freedom required for the particular estimator of μ . For a particular estimating function, he suggests minimizing, with respect to \mathbf{df} , the expected predicted bias ($EPB(\mathbf{df})$), where

$$EPB(\mathbf{df}) = E_{\mathbf{O}}[E_{\mathbf{O}^*}[|S\{\mathbf{O}^*; \hat{\mu}(\mathbf{O}), \hat{\gamma}(\mathbf{O}; \mathbf{df}), \hat{\phi}(\mathbf{O}; \mathbf{df})\}|]]$$

where \mathbf{O}^* is an i.i.d. sample of O independent of \mathbf{O} , and $\hat{\gamma}(\mathbf{O}; \mathbf{df})$, $\hat{\phi}(\mathbf{O}; \mathbf{df})$, and $\hat{\mu}(\mathbf{O})$ are estimators of γ , ϕ , and μ using data \mathbf{O} and specified degrees of freedom \mathbf{df} (Note the change of notation for the estimators - this serves to emphasize the data and degrees of freedom from which they are derived). The logic is that if \mathbf{df} is correctly specified, then EPB should be close to zero (at least for large sample sizes). Since only data \mathbf{O} are available, Pan (2001) suggests estimating $EPB(\mathbf{df})$ by a bootstrap smoothed cross validation estimator. That is, we estimate $EPB(\mathbf{df})$ by

$$\widehat{EPB}(\mathbf{df}) = \frac{1}{B} \sum_{b=1}^B |S\{\mathbf{O}^{*b-}; \hat{\mu}(\mathbf{O}^{*b}), \hat{\gamma}(\mathbf{O}^{*b}; \mathbf{df}), \hat{\phi}(\mathbf{O}^{*b}; \mathbf{df})\}|$$

where B refers to the number of bootstrapped samples, \mathbf{O}^{*b} is the b th bootstrapped sample drawn from \mathbf{O} with replacement, and \mathbf{O}^{*b-} and the observations in \mathbf{O} but not in \mathbf{O}^{*b} .

We applied this criterion (with $B = 250$) to the three estimators used in the data analysis in Section 4. In Table 1, for each α , the bolded estimates indicates the degrees of freedom which minimized the EPB criterion. With the exception of $\alpha = -1.0$, the optimal df_{γ} was always 1. For the DR estimates, the optimal df_{ρ} varied according α . The optimal choice of df_{ρ} was always greater than 1.

We also applied the EPB criterion (with $B = 250$) to the simple inverse weighted (SIW) and doubly robust (DR) estimators in our simulation study. For each of 50 Monte Carlo simulations, we computed the EPB for the SIW and DR estimators as a function of various choices of df_{γ} and df_{ρ} . In Table 2, we present the median EPB as a function of the specified degrees of freedom. In parentheses, we also include the number of times (out of 50) that the specified degrees of freedom were considered optimal. While there is great variability, we see that this criterion does a good job,

on average, of identifying the correct degrees of freedom (i.e., $df_\gamma \geq 3$ and $df_\rho \geq 12$).

7 Summary and Discussion

In this paper, we showed how to draw inference about the mean of an outcome, which is believed to be informatively missing on some study subjects. Our approach is much more flexible than other approaches, which rely on much more stringent modeling assumptions.

We presented three types of estimators. The simple inverse weighted and orthogonal estimators require correct specification of a semiparametric generalized additive selection model. The doubly robust estimator requires correct specification of either the generalized additive selection model or a model for the conditional distribution of the outcome given covariates and selection. In general, we recommend using the doubly robust estimator. If computing time is a big issue, then the orthogonal estimator has been shown to be a nice alternative.

Our estimators require specification of various smoothing parameters. The choice of these parameters are critically important. At a minimum, we suggest performing sensitivity analysis with respect to these parameters. From a more formal perspective, we have provided one data analytic technique for choosing these parameters.

This paper has raised a number of challenging, open research questions including: What are the precise regularity conditions required to establish that our estimator of γ converges at rate $n^{1/4+\epsilon}$; How can our results be extended to longitudinal and time-to-event studies with time-dependent covariates?; Are there better criterion for choosing smoothing parameters? How can one construct an automatic procedure for selecting smoothing parameters.

In terms of longitudinal and time-to-event studies, we believe that constructing doubly robust estimators under the class of non-ignorable models considered by Rotnitzky, Robins, and Scharfstein (1998), Scharfstein, Rotnitzky, and Robins (1999), and Scharfstein and Robins (2002) will be highly unlikely. For the cross-sectional outcome setting considered in this paper, we were actually lucky to obtain doubly robust estimators. In emphasize this point, Rotnitzky and Robins (2001) characterized the space of link functions for the $P[R = 0|\mathbf{X}, Y]$ for which doubly robust estimation

is possible. As stated by these authors, the characterization is “very complex and non-intuitive.” While the logit link is in the space of link functions, the probit and log minus log links are not. For the longitudinal and time-to-event settings, it will be relatively straightforward to construct simple inverse weighted estimators, but construction of orthogonal estimating functions remains an a seemingly tractable but unexplored technical exercise.

In addition to working on the construction of a penalized objective function which allows for automatic selection of smoothing parameters, future work will be focused on the development of a fully Bayesian approach. In such an approach, the smoothing parameter selector is obtained automatically (Berry, Carroll, and Ruppert, 2002). The further advantage of a Bayesian approach is that it is helpful in decision making as it allows formal incorporation of prior beliefs about selection bias so that the ultimate inference about μ can be summarized through a posterior distribution instead of through a sensitivity analysis (see Scharfstein, Daniels, and Robins, 2003).

Bibliography

- Berry, S., Carroll, R., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association* **97**, 160–169.
- Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (disc: P73-93). *Applied Statistics* **43**, 49–73.
- Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., and Merigan, T. C. (1996). A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine* **335**, 1081–1090.
- Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall.
- Kauermann, G. and J.D., O. (2002). Local likelihood estimation in generalized additive models. *Scandinavian Journal of Statistics* .

- Lin, X. and Xiang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B* **61**.
- Linton, O. B. and Härdle, W. (1996). Estimation of additive regression models with known links. *Biometrika* **83**, 529–540.
- Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5**, 99–135.
- Pan, W. (2001). Model selection in estimating equations. *Biometrics* **57**, 529–534.
- Robins, J. (1999). Robust estimation in sequentially ignorable missing data and causal inference models. In: *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, 6–10.
- Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48**, 479–495.
- Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine* **16**, 285–319.
- Robins, J. M. and Rotnitzky, A. (2001). On double robustness. *Statistica Sinica* **11**(4), 920–936.
- Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: E. M. Halloran and D. Berry (eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, 1–94. New York: Springer-Verlag.
- Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–590.

- Scharfstein, D., Daniels, M., and Robins, J. (2003). Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes. *Biostatistics* To Appear.
- Scharfstein, D. O. and Robins, J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika* **89**(4).
- 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**, 1096–146.
- Shively, S., Kohn, R., and Wood, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior (with discussion). *Journal of the American Statistical Association* **94**, 777–806.
- Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics* **14**, 590–606.
- Wood, S. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B* **62**, 413–428.

Table 1: Three estimates^a of μ (simple inverse weighted - SIW; doubly robust - DR; orthogonal - ORTH) as a function of the smoothing parameters df_γ and df_ρ for five levels of the selection bias parameter $\alpha = -1.0, -0.5, 0.0, 0.5, 1.0$. Bootstrapped^b and analytic^c standard errors in parentheses. Bold quantities denote estimator-specific optimal choice of degrees of freedom based on EPB criterion.

α	Estimator	df_ρ	df_γ			
			1	3	6	9
-1.0	SIW	-	347.5 ^a (11.8 ^b)	346.5 (10.9)	346.1 (11.4)	346.1 (11.5)
		1	362.7 ^a (10.2 ^b)	362.0 (9.9)	362.0 (10.2)	362.2 (10.2)
	DR	3	358.8 (9.2)	358.4 (9.1)	358.6 (9.3)	359.0 (9.3)
		6	358.7 (9.0)	358.4 (9.0)	358.5 (9.2)	359.0 (9.1)
		9	359.0 (8.9)	358.7 (8.8)	358.9 (9.0)	359.3 (9.0)
		-	363.6^a (8.4^b, 8.9^c)	363.9 (8.4, 8.9)	364.4 (8.3, 9.0)	364.6 (8.3, 9.0)
-0.5	SIW	-	367.9 (9.0)	366.9 (8.7)	366.9 (8.8)	366.8 (8.9)
		1	378.3 (9.3)	376.9 (9.2)	376.9 (9.4)	376.7 (9.3)
	DR	3	372.4 (8.1)	372.2 (8.2)	372.4 (8.3)	372.5 (8.2)
		6	371.7 (8.2)	371.6 (8.3)	371.9 (8.3)	372.0 (8.2)
		9	371.5 (8.2)	371.5 (8.3)	371.8 (8.3)	371.9 (8.3)
		-	373.0 (8.3, 8.4)	373.1 (8.3, 8.4)	373.5 (8.2, 8.5)	373.6 (8.2, 8.5)
0.0	SIW	-	380.8 (8.6)	379.8 (8.5)	380.1 (8.6)	380.0 (8.7)
		1	389.9 (9.6)	388.0 (9.6)	387.8 (9.7)	387.2 (9.7)
	DR	3	382.3 (8.2)	382.4 (8.3)	382.6 (8.4)	382.4 (8.3)
		6	381.2 (8.4)	381.5 (8.4)	381.8 (8.4)	381.7 (8.4)
		9	380.7 (8.4)	381.1 (8.4)	381.5 (8.4)	381.5 (8.4)
		-	382.1 (8.4, 8.4)	382.2 (8.5, 8.5)	382.4 (8.4, 8.6)	382.3 (8.3, 8.6)
0.5	SIW	-	391.4 (8.8)	390.5 (8.8)	391.1 (9.0)	390.9 (9.2)
		1	400.1 (10.3)	397.9 (10.3)	397.4 (10.5)	396.5 (10.5)
	DR	3	390.9 (8.7)	391.6 (8.8)	391.8 (8.9)	391.3 (8.9)
		6	389.4 (8.9)	390.4 (8.9)	390.8 (8.9)	390.5 (8.9)
		9	388.7 (9.0)	389.8 (8.9)	390.3 (8.9)	390.2 (8.9)
		-	391.0 (8.7, 8.8)	391.1 (8.8, 8.9)	391.1 (8.7, 9.1)	390.9 (8.6, 9.1)
1.0	SIW	-	402.4 (9.6)	401.6 (9.8)	402.5 (10.1)	402.3 (10.3)
		1	409.4 (11.5)	407.4 (11.6)	406.5 (11.8)	405.1 (12.0)
	DR	3	398.7 (9.8)	400.5 (9.9)	400.7 (10.0)	399.9 (10.1)
		6	396.7 (10.0)	399.1 (9.9)	399.7 (9.9)	399.2 (10.0)
		9	395.8 (10.2)	398.2 (10.0)	399.0 (10.0)	398.8 (10.1)
		-	400.2 (9.3, 9.7)	400.3 (9.3, 10.0)	400.2 (9.2, 10.1)	399.9 (9.1, 10.2)

Table 2: Median EPB as a function of the specified degrees of freedom for SIW and DR estimators. In parentheses, the number of times (out of 50) that the specified degrees of freedom were considered optimal.

Estimator	df_ρ	df_γ					
		1	2	3	4	6	8
SIW	-	8.4 (3)	6.6 (8)	6.6 (26)	6.7 (7)	7.1 (6)	8.2 (0)
DR	1	3.8 (0)	3.6 (1)	3.6 (3)	3.6 (0)	3.7 (0)	3.8 (1)
	2	3.6 (0)	3.5 (0)	3.5 (2)	3.6 (0)	3.7 (0)	3.7 (0)
	6	3.6 (1)	3.5 (1)	3.5 (0)	3.5 (1)	3.6 (0)	3.7 (0)
	12	3.6 (2)	3.5 (6)	3.5 (1)	3.5 (1)	3.6 (0)	3.7 (1)
	18	3.6 (0)	3.5 (5)	3.5 (4)	3.5 (1)	3.6 (1)	3.6 (0)
	24	3.9 (1)	3.6 (4)	3.6 (4)	3.6 (2)	3.6 (1)	3.8 (1)
	30	4.0 (1)	3.7 (2)	3.6 (2)	3.6 (0)	3.7 (0)	3.9 (0)

Figure 1: Estimates of $\gamma_2(X_2)$, $\gamma_3(X_3)$, $\rho_2(X_2)$, and $\rho_3(X_3)$ for various choices of df_γ and df_ρ . Selection bias function in (1) is assumed equal to $-0.5 \log(Y)$.

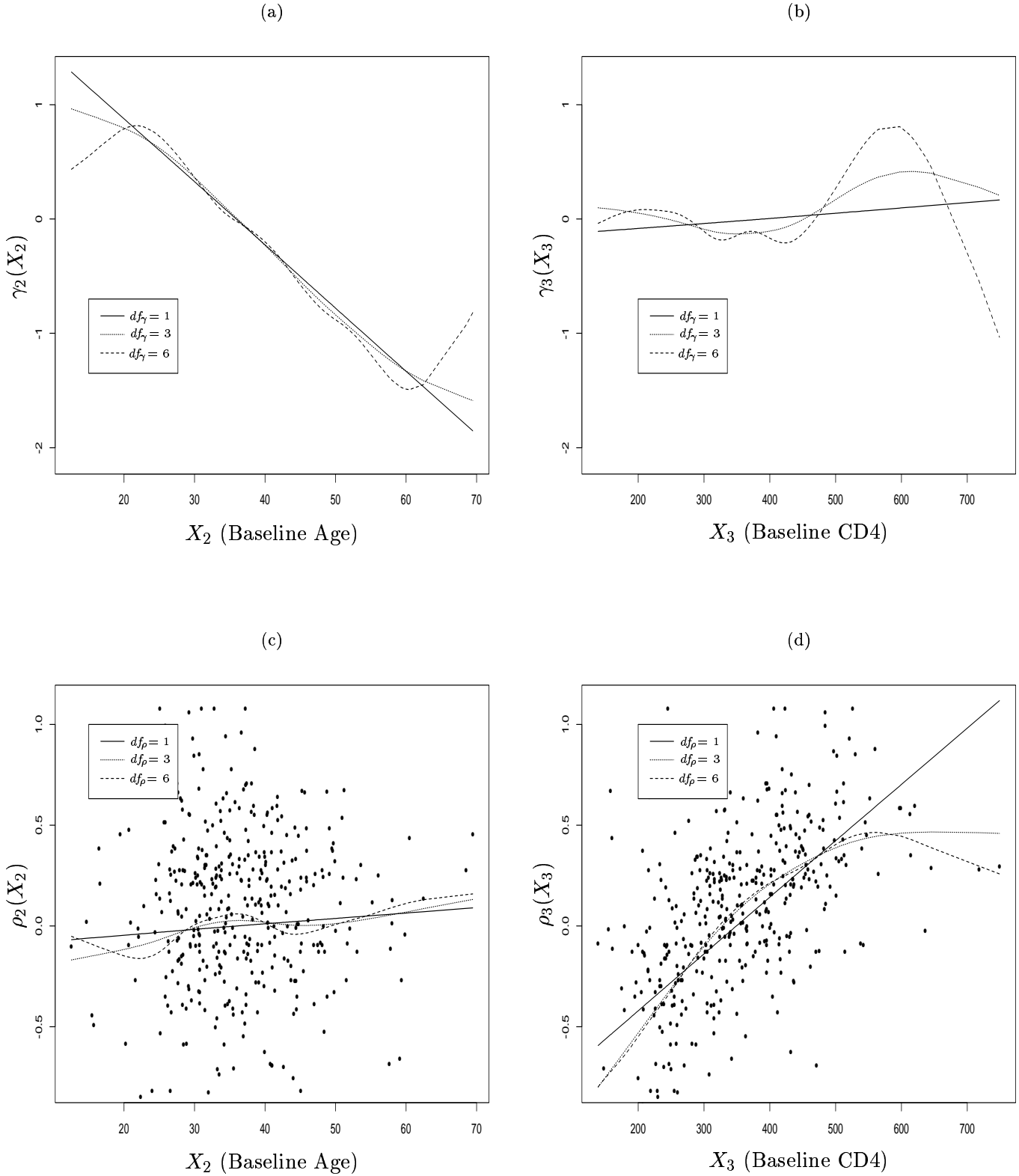


Figure 2: True values (circles) and estimates of $\gamma_1(X_1)$, $\gamma_2(X_2)$, $\rho_1(X_1)$, and $\rho_2(X_2)$ for various choices of df_γ and df_ρ .

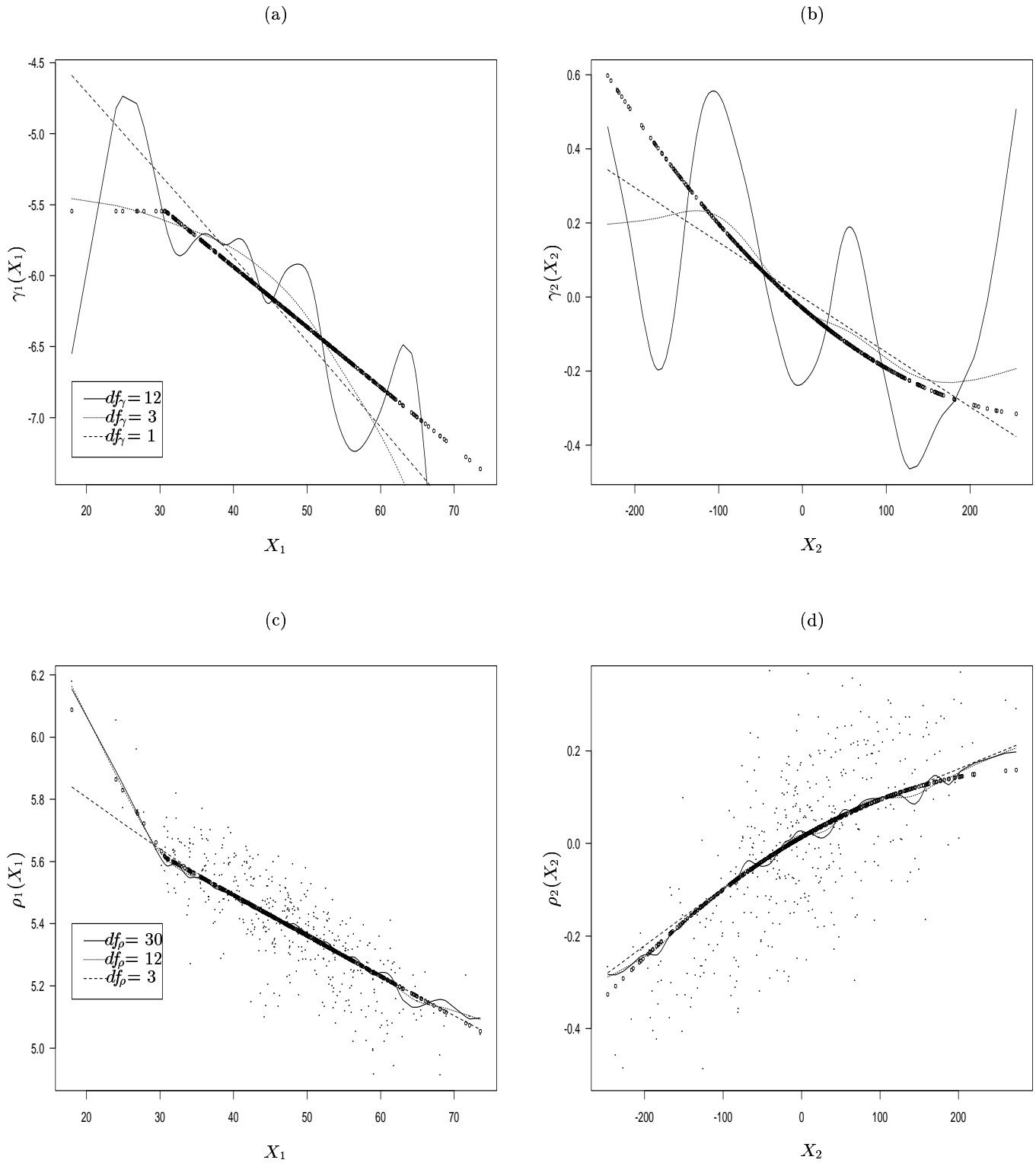


Figure 3: Bias, monte carlo standard error, and mean squared error of *i*) simple inverse weighted estimator (SIW), *ii*) doubly robust estimator with correct specification of the distributional form for the law of Y given X and $R = 1$ (DR), *iii*) doubly robust estimator with incorrect specification of the distributional form of Y given X and $R = 1$ (DRRW), and *iv*) orthogonal estimator (ORTH), as a function of df_γ .

