Observational Studies 1 (2015) 241-290 Submitted 4/2015; Published 10/2015

Simulation-Extrapolation for Estimating Means and Causal Effects with Mismeasured Covariates

J.R. Lockwood **jrlockwood@ets.org**
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541, USA

Daniel F. McCaffrey **dmccaffrey@ets.org**
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541, USA

Abstract

Regression, weighting and related approaches to estimating a population mean from a sample with nonrandom missing data often rely on the assumption that conditional on covariates, observed samples can be treated as random. Standard methods using this assumption generally will fail to yield consistent estimators when covariates are measured with error. We review approaches to consistent estimation of a population mean of an incompletely observed variable using error-prone covariates, noting difficulties with applying these methods. We consider the application of Simulation-Extrapolation (SIMEX) as a simple and effective alternative. We provide technical conditions under which SIMEX will lead to a consistent estimator of a population mean and argue why it may function well in common settings. We use a simulation study to demonstrate its potential for removing nearly all of the bias in regression, weighting and doubly robust estimators for a population mean while maintaining precision competitive with what would be obtained without measurement error. We also discuss and evaluate options for estimating the standard error of the SIMEX mean estimator. Finally, we present an empirical example of estimating middle school effects on student achievement.

Keywords: causal inference, doubly robust estimators, estimating equations, inverse probability weighting, measurement error, non-response weighting

1. Introduction

A common problem in applied research is the estimation of a population mean from a sample with nonrandom missing data. Applications include survey nonresponse in which variables of interest are missing for some units, and observational studies of causal effects in which each unit's potential outcome under one of multiple possible treatment assignments is observed and the remaining potential outcomes are missing (Bang and Robins, 2005; Kang and Schafer, 2007; Lunceford and Davidian, 2004; Robins et al., 1994; Rubin, 1974; Scharfstein et al., 1999). A common way of constructing consistent estimators for a population mean in these settings is to assume that missing data are "ignorable" conditional on observed covariates. That is, although the observed outcome data may be a nonrandom sample from the full population, it is assumed that they are random samples from the subpopulations of units sharing the same values of the covariates. Estimators using this assumption include regression estimators, inverse probability-of-response weighted (IPW) estimators, matching or stratification estimators, and "doubly-robust" (DR) estimators that combine approaches to yield consistent estimators if either the regression or probability of response model are correct (Kang and Schafer, 2007; Rosenbaum and Rubin, 1983; Stuart, 2010).

The success of these methods requires that the covariates necessary for missing data to be ignorable contain no measurement error (ME). When both outcomes and the probability of response or treatment assignment depend on latent quantities that are observed only through error-prone surrogates, modeling with the surrogates while ignoring ME can result in bias in weighting, matching, regression and related estimators (D'Agostino and Rubin, 2000; Kuroki and Pearl, 2014; Lockwood and McCaffrey, 2015a; McCaffrey et al., 2013; Pearl, 2010; Regier et al., 2014; Steiner et al., 2011; Raykov, 2012; Yi et al., 2012).

ME in key covariates is commonplace. For example, scores from standardized achievement tests commonly are used to adjust for non-equivalent student groups in observational studies in educational research (Battauz and Bellio, 2011; Lockwood and McCaffrey, 2014), including the estimation of individual school and teacher "value-added" effects for accountability purposes (Braun, 2005; Harris, 2011; McCaffrey et al., 2004a). Test scores are error-prone measures of latent achievement (Lord, 1980). It is clear that future achievement outcomes, and perhaps other outcomes of interest, depend on these latent achievement attributes, and thus observational treatment groups need to be balanced on these attributes in order to estimate treatment effects unbiasedly. Balancing the error-prone scores is generally insufficient to balance the latent attributes when treatment assignment depends in part on the latent quantities (Yi et al., 2012). For example, in a study of a computer-based algebra tutor (Pane et al., 2014), students were administered a study pre-test designed to assess algebra knowledge, but student selection into the intervention condition was based in part on decisions by school personnel that were made prior to the administration of the pre-test. The observed pre-test scores can at best proxy for the underlying attributes that differentiate intervention from control students. Therefore adjusting the groups to be equivalent on the latent constructs measured by the pre-test, rather on than the observed pre-test scores, would be necessary to argue for unbiased estimation of the intervention effect.

ME occurs in other contexts such as survey responses, which are often used as covariates for observational studies. For example, McCaffrey et al. (2010) used student survey responses to questions about attitudes, dispositions, and alcohol, cigarette and marijuana

use at grade 7 to control for differences among students who did or did not have heavy marijuana use at grades 9 and 10 in a study of the effect of marijuana use on high school graduation. However, there is a substantial literature on the errors in student survey data including self-reported drug use (Freier et al., 1991; Martino et al., 2009).

Unfortunately, methods to correct regression, IPW, DR and other estimators of population means for covariate ME tend to be complicated. For example, McCaffrey et al. (2013) provide sufficient conditions for weights based on error-prone covariates to provide consistent estimates when outcomes and the probability of response or treatment assignment depend on latent covariates. They also develop a computational approach for calculating weights that satisfy the required conditions. It involves two steps: 1) using nonlinear ME modeling with the observed data to estimate the parameters of the true propensity score function of the latent covariates; and 2) using Monte Carlo integration to solve an integral equation for a function of the observed covariates that is unbiased for the corresponding inverse of the propensity score function evaluated at the latent covariates. The first step can be difficult. For the second step, the class of problems for which unbiased estimating functions can be computed in closed form is limited (Stefanski, 1989) and approximating the required functions also can be difficult. Approaches similar to that developed by McCaffrey et al. (2013) can be used to correct DR estimators for covariate ME but require even more assumptions, additional nonlinear ME modeling, and additional searching for unbiased estimating functions (McCaffrey and Lockwood, 2014). Corrections for recently proposed estimation methods, such as the covariate balancing propensity score method (Imai and Ratkovic, 2014) or propensity score weighting combined with exact balancing of selected covariates (Haberman, 1984; Hainmueller, 2012; Kim, 2010), have yet to be developed, but are likely to face similar or even greater challenges.

In this article we argue that Simulation-Extrapolation (SIMEX) may be a more practical and accessible method for estimating a population mean with incomplete data when covariates are measured with error. SIMEX is a general method for ME correction that provides approximately consistent estimators in models where other ME correction methods would be intractable (Carroll et al., 2006; Cook and Stefanski, 1994). SIMEX is often applied to estimating the parameters of generalized linear models with error-prone covariates (Cook and Stefanski, 1994; Fung and Krewski, 1999). In the causal inference literature, Valeri et al. (2014) use SIMEX to conduct mediation analysis in generalized linear models when a continuous mediator is measured with error. We are unaware of its application to the estimation of population means with nonresponse and covariate ME.

SIMEX has several potential advantages over other approaches for using error prone covariates to account for nonrandom missingness when estimating a population mean. First, it requires no specialized ME modeling or solutions to integral equations. It can be implemented with standard statistical routines that do not correct for ME, making it widely practical and accessible. This is true even in complex settings with multiple error-prone and error-free covariates, or with complicated estimators that combine multiple stages of weighting and regression, where alternative approaches for ME correction may be unclear. Second, it requires no distributional assumptions about latent covariates and their relationships to other covariates, which alternatives often use. Finally, as argued by Cook and Stefanski (1994), implementing SIMEX provides an automatic way of studying the sensi-

tivity of inferences to covariate ME. These features, combined with the relative complexity of alternative approaches to the problem, make SIMEX potentially attractive in practice.

In Section 2 we establish notation and assumptions about the observed and latent data. In Section 3 we review regression, IPW and DR methods for consistent estimation of a population mean with incomplete data and covariate ME. In Section 4 we review the SIMEX method, provide technical conditions under which it will lead to a consistent estimator of a population mean with incomplete data and covariate ME, suggest a method for assessing whether these conditions are likely to hold in a given setting, and discuss standard error estimation. In Section 5 we demonstrate the successful performance of SIMEX for regression, IPW and DR estimators in a simulation study. We then present an empirical example of the estimation of middle school effects on student achievement outcomes in Section 6, and conclude with a discussion and suggestions for future research in Section 7.

2. Notation and Model Assumptions

We let (Y, R, Z, X, W) be random quantities with a proper joint distribution on a probability space. We assume we are dealing with independent and identically distributed (IID) samples of size n from this joint distribution corresponding to individual units in a population, and subscript units by i when necessary. The dichotomous response indicator R is observed for all units. The outcome of interest is Y, which is observed for units when $R = 1$ but is missing when $R = 0$. We assume that the goal is to estimate the population mean $\mu_* := E[Y]$ and assume this is finite.

This framework covers several common applications. In survey sampling, Y is an outcome of interest and R indicates whether or not this outcome is observed for a sampled unit. In longitudinal settings, $R = 1$ can indicate units for which an outcome Y from a given timepoint is observed, and $R = 0$ can indicate units who have dropped out by that timepoint. In causal inference with observational data, there are two outcomes of interest: Y_0, the potential outcome of each unit had it received control ($T = 0$), and Y_1, the potential outcome of each unit had it received treatment ($T = 1$). The population average treatment effect is defined as $E[Y_1] - E[Y_0]$, so the population means of both potential outcomes must be estimated. However Y_1 is observed only for treatment units and Y_0 is observed only for control units, so there are missing data for both outcomes. When estimating $E[Y_1]$, Y refers to Y_1 and $R = T$, and when estimating $E[Y_0]$, Y refers to Y_0 and $R = (1 - T)$.

The problem in all of these cases is that $E[Y \mid R = 1]$ does not generally equal μ_* because the observed values of Y are not necessarily a random sample from the population. To get around this problem, we assume the covariates (X, Z) are sufficient for "strong ignorability" of response for inference about Y (Imbens, 2000; Rosenbaum and Rubin, 1983):

$$Y \text{ is independent of } R \text{ given } (X, Z) \text{ and } p(R = 1 \mid X, Z) > 0 \text{ for all } (X, Z). \qquad (1)$$

Following Kang and Schafer (2007), we denote the propensity score $p(R = 1 \mid X, Z)$ by $\pi(X, Z)$, and we denote the conditional expectation function $E[Y \mid X, Z]$ by $m(X, Z)$.

Although we assume that (X, Z) is required for strong ignorability, we assume that the observed covariates for each unit are (W, Z) where W is the error-prone measure of X, and X is not observed for any unit. This is in contrast to a case where X might be partially missing for some units (Rosenbaum and Rubin, 1984; D'Agostino and Rubin, 2000) because

X is never observed. Observed covariates that can be treated as error-free are denoted by Z. Each of W, X and Z can be vectors but we do not explicitly use vector notation.

We assume that $W = X + U$ for $U \sim N(0, \sigma^2)$, where σ^2 is either known, or is estimated sufficiently well from auxiliary data (e.g. replicate measurements or a validation dataset) to be treated as known. Normality of the ME is typically assumed in SIMEX, though the method is purported to be robust to deviations from normality (Carroll et al., 2006). We further assume that U is independent of (Y, R) given (X, Z), so that ME is non-differential (Carroll et al., 2006) and W satisfies "strong surrogacy" (Lockwood and McCaffrey, 2015a). This implies that W would be irrelevant to the analysis if X were observed. This assumption requires scrutiny in every context, and Lockwood and McCaffrey (2015a) discuss scenarios in which it may be more or less likely to hold.

2.1 Simulated Example

We now introduce a simulated example that we use throughout the article. The scenario has three scalar covariates X, Z_1 and Z_2. The distribution of X is a mixture of normal distributions. It is skewed, multi-modal and scaled to have mean zero and variance one. The variable Z_1 is correlated 0.3 with X and normally distributed conditional on X. Its marginal mean and variance are also zero and one. The variable Z_2 is dichotomous with mean 0.5 and independent of (X, Z_1).

We generate R using the model in the simulation example in McCaffrey et al. (2013). That example specified R as Bernoulli with

$$p(R = 1 \mid X = x, Z_1 = z_1, Z_2 = z_2) \quad = \quad G(0.5 + 1.2x + 0.5z_1 - 1.0z_2 + 0.7xz_2) \qquad (2)$$

where G is the CDF of a Cauchy random variable. This model imposes clear differences between samples with $R = 0$ and $R = 1$ on the distributions of all covariates: compared to the sample with $R = 0$, the means for the sample with $R = 1$ are almost 1 SD unit higher on X, 0.57 SD units higher on Z_1 and about 0.16 lower on Z_2.

We consider two outcome variables. The first is $Y = a + bX + cZ_1 + dZ_2 + \epsilon$, $\epsilon \sim N(0, \tau^2)$ and independent of everything, where (a, b, c, d, τ^2) are set so that $E[Y] = 0$, $\text{Var}[Y] = 1$, the covariates explain 80% of the variance in Y, the mean for $R = 1$ cases is about 0.43, and the mean for $R = 0$ cases is about -0.42. Thus $R = 1$ and $R = 0$ cases differ substantially on the distribution of Y. We refer to this as the "linear outcome" because its conditional mean function is linear in the covariates. We also consider $Y^* = (0.812 + 1.255Y)I(0.812 + 1.255Y > 0)$ where I is the indicator function. We chose 0.812 and 1.255 so that both $E[Y^*]$ and $\text{Var}[Y^*]$ are approximately 1. Among $R = 0$ cases, Y^* has mean of about 0.58 and probability about 0.39 of equaling zero, compared to a mean of about 1.42 and probability about 0.13 of equaling zero among $R = 1$ cases, and so is clearly differentiated between the two groups. We refer to Y^* as the "tobit outcome" because it follows a tobit model conditional on the covariates.

Finally, we generate the measurement errors U as independent of everything and distributed as $N(0, \sigma^2 = 0.176)$. The resulting reliability of W as a measure of X is 0.85, computed as $\text{Var}[X]/(\text{Var}[X] + \sigma^2)$ (Crocker and Algina, 1986). This can be interpreted as the correlation of two hypothetical measurements W_1 and W_2 of the same X where W_1 and W_2 have independent measurement errors. A reliability of 0.85 is consistent with

those of standardized test scores from state K-12 assessments. The joint distributions of (Y, R, Z_1, Z_2, X, W) and of (Y^*, R, Z_1, Z_2, X, W) both meet the strong ignorability and surrogacy conditions specified previously in this section. The code for the R environment (R Development Core Team, 2015) that we use to simulate data from these distributions is provided in the Appendix. The simulated data are summarized in Table 1 for reference.

Variable	Description
Y	Linear outcome with $E[Y] = 0$, $\text{Var}[Y] = 1$, and $E[Y\|R = 1] \approx 0.43$
Y^*	Tobit outcome with $E[Y^*] = 1$, $\text{Var}[Y^*] = 1$, and $E[Y^*\|R = 1] \approx 1.42$
R	Response indicator; $R = 1$ indicates outcome is observed
Z_1, Z_2	Error-free covariates
X	Unobserved covariate with $E[X] = 0$ and $\text{Var}[X] = 1$
W	Observed proxy for X with normal ME and reliability 0.85

Table 1: Summary of random variables from the simulated example.

3. Review of Approaches for Consistent Estimation

Kang & Schafer (2007) summarize regression, IPW and DR approaches for estimating μ_* in the standard (i.e., no ME) case where (X, Z) is observed for all units. In this section we review these approaches and discuss how they can be expressed as solutions to estimating equations. We use this correspondence to discuss analogs to these approaches when (W, Z) is instead observed for all units and the other assumptions of the previous section hold.

3.1 Regression

If X_i were observed and if the conditional mean function $m(X, Z) = E[Y \mid X, Z]$ were known, then estimating the population mean μ_* would be straightforward. By the law of iterated expectation, $\mu_* = E[E[Y \mid X, Z]]$ so that $(1/n)\sum_{i=1}^n m(X_i, Z_i)$ would consistently estimate the mean. In practice, $m(X, Z)$ is unknown and must be estimated using only the observed values of Y. This is possible with the strong ignorability assumption in (1), which implies that $E[Y \mid X, Z] = E[Y \mid X, Z, R = 1]$, which can be estimated because Y is observed for all $R = 1$ units. In practice it is often assumed that $m(X, Z) = \tilde{m}(X, Z, \delta_*)$ for a parametric model $\tilde{m}(X, Z, \delta)$, where δ is a K-dimensional vector of model parameters with true value δ_*. Kang & Schafer (2007) define $(1/n)\sum_{i=1}^n \tilde{m}(X_i, Z_i, \widehat{\delta})$ as the regression estimator of μ_*, where $\widehat{\delta}$ is an estimate of δ_*. This estimator is consistent for μ_* when $\tilde{m}(X, Z, \delta)$ is correctly specified, $\widehat{\delta}$ is consistent for δ_*, strong ignorability holds, and general regularity conditions hold (see Tsiatis (2007) for an example of these regularity conditions).

The two stages of estimating δ_* with $\widehat{\delta}$ and then using the $\tilde{m}(X, Z, \widehat{\delta})$ to estimate μ_* can be expressed in one stage as the solution to an estimating equation

$$(1/n)\sum_{i=1}^{n} \psi(Y_i, R_i, Z_i, X_i, \theta) = 0. \tag{3}$$

Here $\psi(Y, R, Z, X, \theta)$ is a $K+1$ dimensional function with one function corresponding to each of the parameters in $\theta = (\mu, \delta')'$ with true value $\theta_* = (\mu_*, \delta_*')'$. The key idea of estimating equations is that if $\psi(Y, R, Z, X, \theta)$ is specified such that θ_* is the unique value of θ with

the property that $E[\psi(Y, R, Z, X, \theta_*)] = 0$, and other regularity conditions hold, then the solution to Equation 3 consistently estimates θ_* (Stefanski and Boos, 2002). We assume the first component $\psi(Y, R, Z, X, \theta)_1$ corresponds to μ so that $\psi(Y, R, Z, X, \theta)_1 = \tilde{m}(X, Z, \delta) - \mu$. This satisfies the key requirement because $E[\psi(Y, R, Z, X, \theta_*)_1] = E[\tilde{m}(X, Z, \delta_*)] - \mu_* = E[E[Y|X, Z]] - \mu_*$. The remaining components of ψ depend on the function $\tilde{m}$ and the method used to estimate its parameters δ. Later we consider the case where $\tilde{m}$ is linear and δ is estimated with least squares, but nonlinear, generalized linear and other estimation methods also fit this form (see, e.g. Lunceford and Davidian 2004 and references therein).

When (Y, R, Z, W) is observed instead of (Y, R, Z, X), the estimating function must be modified from $\psi(Y, R, Z, X, \theta)$ to $\varphi(Y, R, Z, W, \theta)$ such that $E[\varphi(Y, R, Z, W, \theta_*)] = 0$. Carroll et al. (2006) provide equations for the terms associated with δ. For μ we note that if $\tilde{m}$ is replaced by $B(W, Z, \delta)$ such that $E[B(W, Z, \delta) \mid X, Z] = \tilde{m}(X, Z, \delta)$, then $\varphi(Y, R, Z, W, \theta)_1 = B(W, Z, \delta) - \mu$ meets the sufficient equality condition. This is because $B(W, Z, \delta)$ has the property that $E[B(W, Z, \delta_*) \mid X, Z] = m(X, Z)$ so that $E[B(W, Z, \delta_*)] - \mu_* = 0$. This idea naturally extends the approach of McCaffrey et al. (2013) which is a special case of the general "corrected score function" or "unbiased estimating function" approach to ME correction (Carroll et al., 2006).

In the case of a linear model, a plug-in approach would work: if $m(X, Z) = \delta_{*0} + \delta_{*1}X + \delta_{*2}Z$ then $B(W, Z, \delta) = \delta_0 + \delta_1 W + \delta_2 Z$ satisfies the requirements of an unbiased estimating function. The parameters δ_* can be estimated consistently from the observed data by $\widehat{\delta}$ obtained using standard approaches for correcting a linear model for covariate ME (Fuller, 2006), and then the consistent regression estimator for μ_* is $(1/n)\sum_{i=1}^{n} B(W_i, Z_i, \widehat{\delta})$. With normally distributed ME, a closed form for $B(W, Z, \delta)$ is also possible when $m(X, Z)$ is a polynomial using the results of Stefanski (1989). Another case where a closed form exists with normally distributed ME is the exponential function: if $m(X, Z) = \exp(\delta_{*0} + \delta_{*1}X + \delta_{*2}Z)$, then results using the normal moment generating function indicate that $B(W, Z, \delta) = \exp(\delta_0 + \delta_1 W + \delta_2 Z - \delta_1^2\sigma^2/2)$ satisfies $E[B(W, Z, \delta_*) \mid X, Z] = m(X, Z)$. Outside of these cases, finding $B(W, Z, \delta)$ typically requires Monte Carlo approximations such as those suggested by McCaffrey et al. (2013) or the general Monte Carlo methods using complex variables described by Novick and Stefanski (2002) and Carroll et al. (2006). These methods are not trivial to implement and yield approximate solutions whose validity must be checked carefully in any given setting.

3.2 Inverse Probability-of-Response Weighting

IPW estimators also rely on assumptions about the ignorability of response given observed covariates. However, these estimators avoid making explicit assumptions about the conditional mean of the outcome given those covariates. In fact, one of the purported advantages of weighting is the ability to make adjustments for imbalance in observed covariates between respondents and nonrespondents without reference to the outcomes, and possibly before they are even observed (Rubin, 2001). If the probability of response were known and X_i was observed, then the IPW estimator for μ_* would equal $\frac{\sum_{i=1}^{n} R_i \pi^{-1}(X_i, Z_i) Y_i}{\sum_{i=1}^{n} R_i \pi^{-1}(X_i, Z_i)}$, which is closely related to the Horvitz-Thompson estimator (Horvitz and Thompson, 1952). The intuition is that the estimator is a weighted mean of the observed Y_i, where the weights are chosen to reweight the (X_i, Z_i) distribution among these cases to match the population

distribution of (X, Z). This, combined with the strong ignorability assumption, allows the weighted mean of the observed Y_i to consistently estimate the population mean μ_*. Typically, the response probabilities are unknown, so a functional form for the propensity score is selected and the parameters of this model are then estimated from observed data.

Again, the two stages of estimating the parameters of the propensity score model and then using the estimated propensity scores to calculate the weighted mean can be written in one stage as the solution to an estimating equation. We let α be the parameters of the working model $\tilde{\pi}(X, Z, \alpha)$ and let $\tilde{q}(X, Z, \alpha) = 1/\tilde{\pi}(X, Z, \alpha)$. The IPW estimator solves $(1/n)\sum_{i=1}^{n} \psi(Y, R, Z, X, \theta) = 0$, where now $\theta = (\mu, \alpha')'$ with true value $\theta_* = (\mu_*, \alpha_*')'$, and where $\psi(Y, R, Z, X, \theta)_1 = R\tilde{q}(X, Z, \alpha)[Y - \mu]$. Provided $\tilde{\pi}(X, Z, \alpha_*) = \pi(X, Z)$ and $\widehat{\alpha}$ consistently estimates α_*, the IPW estimator is consistent and asymptotically normal.

When X_i is unobserved, even if we know the true propensity score function $\pi(X, Z)$ or can estimate it consistently from the observed data, we still do not know where to evaluate it. Dealing with this problem again uses the logic of unbiased estimating functions. McCaffrey et al. (2013) show that if $\tilde{q}(X, Z, \widehat{\alpha})$ is replaced by $A(W, Z, \widehat{\alpha})$ where $E[A(W, Z, \alpha) \mid X, Z] = \tilde{q}(X, Z, \alpha)$, then the solution to the resulting estimating equation is consistent if $\tilde{\pi}(X, Z, \alpha_*) = \pi(X, Z)$ and $\widehat{\alpha}$ consistently estimates α_*. The idea is that the weighted estimator using (W, Z) is constructed so that each case is given a weight that is unbiased for its true, unobserved weight determined by (X, Z). A closed form for $A(W, Z, \alpha)$ exists for logistic propensity score functions with normal ME using the normal moment generating function result noted previously. Outside of this case, numerical methods for function approximation are generally required.

3.3 Doubly Robust

Doubly robust estimators combine a model $\tilde{m}$ for the conditional mean of the outcome and a model $\tilde{\pi}$ for the probability of response to provide estimators for μ_* that are consistent if either model, but possibly not both, is correct (Bang and Robins, 2005). This can insure against model misspecification. Kang and Schafer (2007) review several ways to combine the two models to create a doubly robust estimator. Here we focus on one. When X_i is observed, the "bias corrected" DR estimator for μ_* noted by Kang and Schafer (2007) is

$$(1/n)\sum_{i=1}^{n} \tilde{m}(X_i, Z_i, \widehat{\delta}) + \frac{\sum_{i=1}^{n} R_i[Y_i - \tilde{m}(X_i, Z_i, \widehat{\delta})]\tilde{q}(X_i, Z_i, \widehat{\alpha})}{\sum_{i=1}^{n} R_i\tilde{q}(X_i, Z_i, \widehat{\alpha})}, \tag{4}$$

where, as previously defined, $\tilde{m}(X, Z, \delta)$ is a working model for the conditional mean and $\tilde{q}(X, Z, \alpha)$ is a weighting function equal to the reciprocal of the working model for the propensity score function [1] This estimator is the solution to an estimating equation $(1/n)\sum_i \psi(Y_i, R_i, Z_i, X_i, \theta) = 0$, where now $\theta = (\mu, \delta', \alpha')'$ with true value $\theta_* = (\mu_*, \delta_*', \alpha_*')'$, and where $\psi(Y, R, Z, X, \theta)_1 = R\tilde{q}(X, Z, \alpha)(Y - \tilde{m}(X, Z, \delta)) + (\tilde{m}(X, Z, \delta) - \mu)$. This estimator is consistent and asymptotically normal provided either $\tilde{q}(X, Z, \alpha_*) = \pi(X, Z)^{-1}$ or $\tilde{m}(X, Z, \delta_*) = m(X, Z)$. The intuition for this estimator is that when $\tilde{m}$ is correctly specified, the term on the right in Equation 4 converges to zero and the term on the left

1. Kang and Schafer (2007) present (4) as a variation of $(1/n)\sum_{i=1}^{n} \tilde{m}(X_i, Z_i, \widehat{\delta}) + (1/n)\sum_{i=1}^{n} R_i[Y_i - \tilde{m}(X_i, Z_i, \widehat{\delta})]\tilde{q}(X_i, Z_i, \widehat{\alpha})$. As with the IPW estimator, normalizing the weights tends to improve the precision of the estimator and is commonly used in practice.

converges to μ_* regardless of whether $\tilde{\pi}$ is correctly specified, whereas when $\tilde{\pi}$ is correct, the term on the left converges to a possibly incorrect value $\tilde{\mu}$ while the term on the right corrects this bias by converging to $\mu_* - \tilde{\mu}$.

When X is measured with error by W, satisfying surrogacy, then again this approach can be extended using unbiased estimating functions. If functions $A(W, Z, \alpha)$, $B(W, Z, \delta)$, and $C(W, Z, \alpha, \delta)$ satisfy $E[A(W, Z, \alpha) \mid X, Z] = \tilde{q}(X, Z, \alpha)$, $E[B(W, Z, \delta) \mid X, Z] = \tilde{m}(X, Z, \delta)$ and $E[C(W, Z, \alpha, \delta) \mid X, Z] = \tilde{m}(X, Z, \delta)\tilde{q}(X, Z, \alpha)$, then

$$E[R(A(W, Z, \alpha)Y - C(W, Z, \alpha, \delta))] + E[B(W, Z, \delta)] = E[Y]$$

if either $\tilde{q}(X, Z, \alpha_*) = \pi(X, Z)^{-1}$ or $\tilde{m}(X, Z, \delta_*) = m(X, Z)$. The estimating function $\psi(Y, R, Z, X, \theta)_1$ for μ_* is replaced by $\varphi(Y, R, Z, W, \theta)_1 = R(A(W, Z, \alpha)Y - C(W, Z, \alpha, \delta)) + (B(W, Z, \delta) - \mu)$ which has mean zero for $\mu = \mu_*$ if either $\tilde{q}(X, Z, \alpha_*) = \pi(X, Z)^{-1}$ or $\tilde{m}(X, Z, \delta_*) = m(X, Z)$. Using this result to obtain an estimator requires estimating the parameters of two potentially nonlinear models of (X, Z) from the observed data and then solving three integral equations. If X given Z, and W given (X, Z), are both normally distributed, then closed-form expressions exist for all three functions. However, solutions do not exist for some values of the parameters, and the practical performance of these estimators has not been studied. Similar ideas can be used to extend the DR estimator of Bang and Robins (2005) that includes the reciprocal of the estimated propensity score as a covariate in $\tilde{m}(X, Z)$ to the case of ME. Extending the DR approach noted by Kang and Schafer (2007) that uses weighted regression with inverse propensity score weights is possible, but may be impractical because of the potentially large number of functions of (X, Z) that must be estimated and for which corresponding unbiased estimating functions must be found.

3.4 Summary

The general method of unbiased estimating functions can be used to correct certain regression, IPW and DR estimators for covariate ME. However the methods generally require specialized ME modeling and the solution or approximation of unbiased estimating functions that must be tailored to the particular estimator being used. These steps can be difficult. For example, for the simulated data scenario described in Section 2 which includes a complicated marginal distribution for X, a Cauchy link for the propensity score function that includes an interaction term involving X, and a nonlinear conditional mean function (for the tobit outcome), closed form solutions for either estimating the model parameters or determining unbiased estimating functions are unlikely to be available for any of the common mean estimators. In the remaining sections we consider SIMEX as an alternative approach that is more practical and accessible.

4. SIMEX for Estimation of a Population Mean

In this section we 1) review the SIMEX method; 2) provide technical conditions under which it will lead to a consistent estimator of a population mean with incomplete data and covariate ME; 3) provide an example in which these conditions can be verified; 4) suggest a method for assessing whether these conditions are likely to hold in other settings; and 5) discuss standard error estimation.

4.1 Review of SIMEX

To describe SIMEX, we introduce notation similar to that used by Devanarayan and Stefanski (2002) and Lockwood and McCaffrey (2015b). We consider hypothetical data in which X is measured with ME with variance $(1+\lambda)\sigma^2$ for $\lambda \geq -1$, where σ^2 is the variance of the ME in the observed data. The resulting vector of data $(Y, R, Z, X, W_{(1+\lambda)})$ has a joint distribution that we denote by $P_{(1+\lambda)}$. At $\lambda = 0$, $(1+\lambda)\sigma^2 = \sigma^2$, and so $(Y, R, Z, X, W_{(1)})$ has the same distribution $P_{(1)}$ as the observed data. When $\lambda > 0$ the hypothetical data have larger ME than the observed data. At the other extreme, $W_{(0)}$, corresponding to $\lambda = -1$, denotes hypothetical data with no ME so that $W_{(0)} = X$, and $P_{(0)}$ equals the distribution of (Y, R, Z, X). We assume that $(Y_i, R_i, Z_i, X_i, W_{(1+\lambda)i})$ for $i = 1, \ldots, n$ are independent and identically distributed with distribution $P_{(1+\lambda)}$ and we use $\{Y_i, R_i, Z_i, W_{(1)i}\}$ to denote the observed data $(Y_1, R_1, Z_1, W_{(1)1}), (Y_2, R_2, Z_2, W_{(1)2}), \ldots, (Y_n, R_n, Z_n, W_{(1)n})$ across units $i = 1, \ldots, n$, where it is understood that Y_i is observed only when $R_i = 1$. When there is no risk for confusion, we drop the subscripting by i to simplify notation.

As discussed in the previous section, if there were no ME, the estimator of $\widehat{\theta}_n$ of θ_* is the solution to $(1/n)\sum_i \psi(Y_i, R_i, Z_i, X_i, \theta) = 0$. We assume for the remainder that $\psi(Y, R, Z, X, \theta)$ is correctly specified and that other standard regularity conditions hold so that $\widehat{\theta}_n \xrightarrow{p} \theta_*$. We refer to the "naïve" estimator of θ_* as the solution to the equation $(1/n)\sum_i \psi(Y_i, R_i, Z_i, W_{(1)i}, \theta) = 0$ which ignores ME by plugging $W_{(1)i}$ into the equation in place of X_i. We use $\widehat{\theta}_{(1),n}$ to denote this estimator. For large samples, $\widehat{\theta}_{(1),n}$ converges to the solution to $E[\psi(Y, R, Z, W_{(1)}, \theta)] = 0$, which we call $\theta_{(1)}$. Following this notational convention, we denote the solution to $E[\psi(Y, R, Z, W_{(1+\lambda)}, \theta)] = 0$ by $\theta_{(1+\lambda)}$ for $\lambda > -1$. We denote $\lim_{\lambda \to -1} \theta_{(1+\lambda)}$ by $\theta_{(0)}$. Under technical assumptions that we discuss later, $\theta_{(0)} = \theta_*$. If this holds, then the logic of SIMEX is easy to state: if we had samples from $P_{(1+\lambda)}$ for various values of $\lambda > 0$, which can be generated by adding extra ME to the observed data, we could estimate $\theta_{(1+\lambda)}$ and develop a model for how $\theta_{(1+\lambda)}$ varies as a function of λ. Evaluating this model at $\lambda = -1$ would then give us an estimate of θ_*.

The SIMEX algorithm proceeds as follows. First a sequence of values $\lambda_1, \ldots, \lambda_L$ is established, ranging from 0 to an upper bound commonly taken as 2 (Cook and Stefanski, 1994), with $L = 20$ typically providing a sufficiently fine grid. Then for each $\ell = 1, \ldots, L$ synthetic covariates $W^*_{(1+\lambda_\ell)i,b} = W_{(1)i} + U^*_{(\lambda_\ell)i,b}$ are generated, where the simulated $U^*_{(\lambda_\ell)i,b} \sim$ IID $N(0, \lambda_\ell \sigma^2)$. The simulation step is repeated for $b = 1, \ldots, B$, where B is as large as desired to reduce Monte Carlo error to an acceptably low level. The synthetic covariates thus have larger ME variance of $(1+\lambda_\ell)\sigma^2$ compared to σ^2 for the $W_{(1)i}$ themselves. In fact, due to the assumptions of surrogacy and normal ME, $\{Y_i, R_i, Z_i, X_i, W^*_{(1+\lambda_\ell)i,b}\}$ are independent and identically distributed according to $P_{(1+\lambda_\ell)}$ for each b. This is the key property of the simulated data from SIMEX: the synthetic data have the same distribution as hypothetical data with ME variance $(1+\lambda_\ell)\sigma^2$ rather than σ^2. Therefore the naïve estimator $\widehat{\theta}_{(1+\lambda_\ell),b,n}$, equal to the solution to $(1/n)\sum_i \psi(Y_i, R_i, Z_i, W^*_{(1+\lambda_\ell)i,b}) = 0$, is a consistent estimator of $\theta_{(1+\lambda_\ell)}$. This is true for each $b = 1, \ldots, B$, and in practice unnecessary noise in this estimator is reduced by averaging it over the $b = 1, \ldots, B$ sets of simulated measurement errors. We denote this average by

$$\widehat{\theta}_{(1+\lambda_\ell),\overline{B},n} = (1/B)\sum_{b=1}^{B} \widehat{\theta}_{(1+\lambda_\ell),b,n}.$$

This step is performed for $\lambda_1, \ldots, \lambda_L$, which leads to the sequence of synthetic estimators $\widehat{\theta}_{(1+\lambda_1),\overline{B},n} \cdots, \widehat{\theta}_{(1+\lambda_L),\overline{B},n}$.

The final step of SIMEX fits a parametric function $\mathcal{G}(\lambda, \gamma)$ to $\widehat{\theta}_{(1+\lambda_1),\overline{B},n}, \ldots, \widehat{\theta}_{(1+\lambda_L),\overline{B},n}$ to estimate $\widehat{\gamma}_n$ and then extrapolates the prediction from that model back to $\lambda = -1$ to estimate the target by $\mathcal{G}(-1, \widehat{\gamma}_n)$. This extrapolated value serves as the final estimator. The extrapolation step is the "Achilles Heel" of SIMEX because the true functional relationship between λ and $\theta_{(1+\lambda)}$ is typically unknown and therefore the extrapolation, estimated from the observed data, will be only approximate (Cook and Stefanski, 1994).

Figure 1 demonstrates a hypothetical SIMEX projection. The naïve estimator $\widehat{\theta}_{(1),n}$ using observed data is located at $\lambda = 0$. The averages $\widehat{\theta}_{(1+\lambda_\ell),\overline{B},n}$ of the naïve estimators across B Monte Carlo replications using synthetic data with additional ME are located at each selected $\lambda_\ell > 0$. A curve is fitted to these data points and then extrapolated back to $\lambda = -1$ to obtain the SIMEX estimator. The goal is to use the extrapolation to reduce or remove bias in the naïve estimator as an estimator of θ_*.

The SIMEX algorithm repeatedly uses naïve estimators. It avoids the complexities of finding unbiased estimating equations of Section 3. The standard software and models that would be used if there were no ME are all that are required. The trade-off is that the extrapolation function must be modeled and may not be exact.

4.2 Conditions for Consistent Estimation

The general theory of estimating equations can be used to establish conditions under which SIMEX will yield a consistent estimator of a population mean. General arguments supporting the consistency and asymptotic normality of the SIMEX estimator for cases where estimators can be expressed as solutions to estimating equations, such as the estimators of the mean described in Section 3, are provided by Cook and Stefanski (1994), Stefanski and Cook (1995) and Carroll et al. (1996). In this section we establish technical conditions under which SIMEX will lead to a consistent estimator of a population mean with incomplete data and covariate ME, along the way providing some general results for SIMEX that build on previous work by Cook and Stefanski (1994) and Carroll et al. (1996).

Carroll et al. (1996) show that under general regularity conditions, like those required for the solution of Equation 3 to be consistent, $\lim_{n\to\infty} \widehat{\theta}_{(1+\lambda),\overline{B},n} = \theta_{(1+\lambda)}$. This is true for each value of λ. We let $\widehat{\mu}_{(1+\lambda),\overline{B},n}$ denote the first element of $\widehat{\theta}_{(1+\lambda),\overline{B},n}$, the estimate of μ, and $\mu_{(1+\lambda)}$ equal its limit. For extrapolation step, we choose a parametric model for $\mu_{(1+\lambda)}$, which we call $\mathcal{G}(\lambda, \gamma)$. For example, the function may be quadratic so that $\mathcal{G}(\lambda, \gamma) = \gamma_0 + \gamma_1\lambda + \gamma_2\lambda^2$. We can extrapolate all the parameters, but our interest is in μ_*, so we focus only on the extrapolation of $\widehat{\mu}_{(1+\lambda),\overline{B},n}$. We estimate the parameters of $\mathcal{G}$ by fitting the model to the $\ell = 1, \ldots, L$ simulation-based estimates, $\widehat{\mu}_{(1+\lambda_\ell),\overline{B},n}$, for the chosen λ_ℓ values. The resulting parameter estimates can be written as: $\widehat{\gamma}_n = \sum_{\ell=1}^{L} \omega_\ell \widehat{\mu}_{(1+\lambda_\ell),\overline{B},n}$, where ω_ℓ are vectors that depend only on the values of λ_ℓ and do not depend on the observed data or the sample size. The SIMEX estimator of μ_* is $\mathcal{G}(-1, \widehat{\gamma}_n)$. As $n \to \infty$, $\widehat{\gamma}_n \xrightarrow{p} \widetilde{\gamma} = \sum_{\ell=1}^{L} \omega_\ell \mu_{(1+\lambda_\ell)}$ and, by Slutsky's theorem (Serfling, 1980), $\mathcal{G}(-1, \widehat{\gamma}_n) \xrightarrow{p} \mathcal{G}(-1, \widetilde{\gamma})$. We have not assumed that $\mathcal{G}$ is correctly specified, so consequently we cannot assume that $\mathcal{G}(-1, \widetilde{\gamma})$ equals $\mu_{(0)}$, the limit of $\mu_{(1+\lambda)}$ as λ approaches -1. However, if $\mathcal{G}$ is correctly

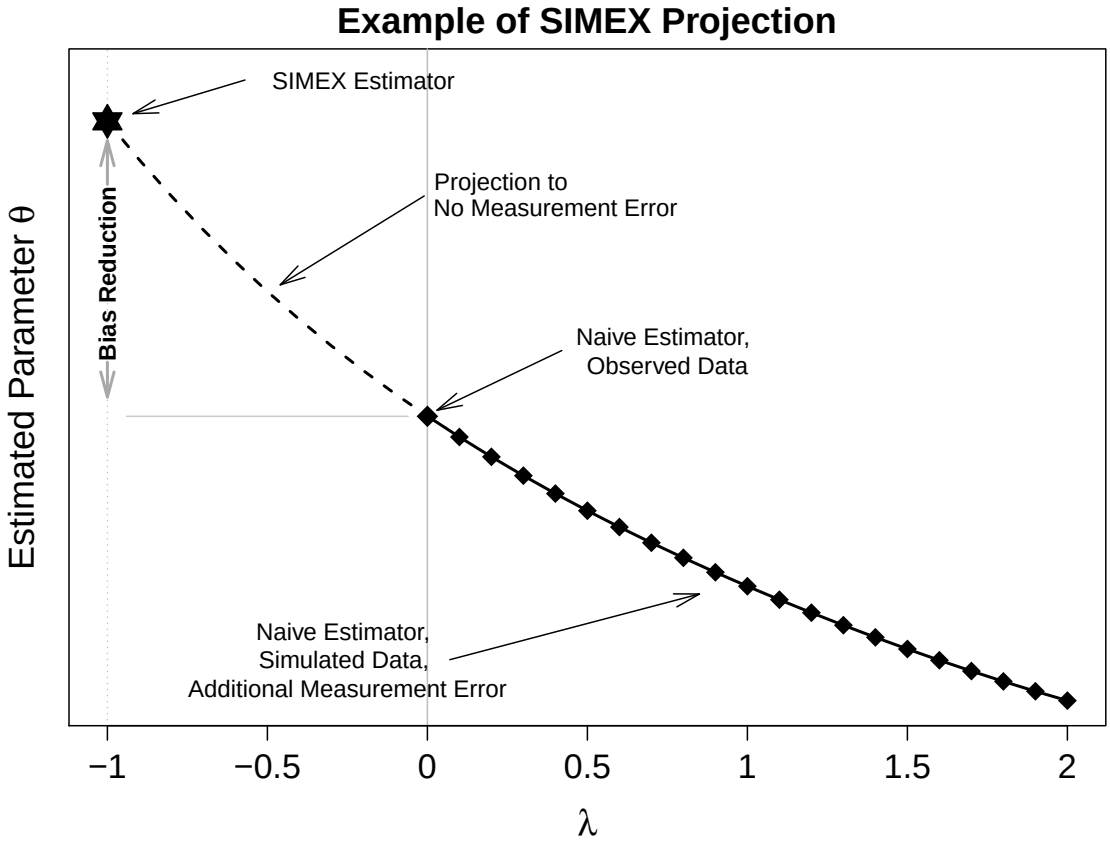

Figure 1: Example of SIMEX projection. The naïve estimator of θ_* using observed data is located at $\lambda = 0$. Averages of naïve estimators across B Monte Carlo replications using synthetic data are located at each selected $\lambda > 0$. The SIMEX estimator is obtained by fitting a curve to the naïve estimators and projecting that function, located at $\lambda = -1$.

specified so that $\mu_{(1+\lambda)} = \mathcal{G}(\lambda, \gamma_*)$ for some γ_*, then $\sum_{\ell=1}^{L} \omega_\ell \mu_{(\mathbf{1}+\lambda_\ell)} = \gamma_*$, provided L is greater than the dimension of γ. Consequently, if we specify the extrapolation function correctly, then $\mathcal{G}(-1, \widehat{\gamma}_n) \xrightarrow{p} \mathcal{G}(-1, \gamma_*)$ and $\mathcal{G}(-1, \gamma_*) = \mu_{(0)}$. Now all we need for consistency is $\mu_{(0)} = \mu_*$; that is, $\mu_{(1+\lambda)}$ must converge to μ_* as λ approaches -1. Recall that $\mu_{(1+\lambda)}$ is the first element of $\theta_{(1+\lambda)}$. Thus, if $\lim_{\lambda \to -1} \theta_{(1+\lambda)} = \theta_*$; that is, if the solution to $E[\psi(Y, R, Z, W_{(1+\lambda)}, \theta)] = 0$ converges to the solution of $E[\psi(Y, R, Z, X, \theta)] = 0$, and we correctly specify the extrapolation function, then $\mathcal{G}(-1, \widetilde{\gamma})$ equals μ_* and the SIMEX estimator is consistent. We provide sufficient conditions for the solution to $E[\psi(Y, R, Z, W_{(1+\lambda)}, \theta)] = 0$ to converge to the solution of $E[\psi(Y, R, Z, X, \theta)] = 0$ in the Appendix. These conditions apply to SIMEX estimation in any parametric model where parameter estimates can be expressed as solutions to estimating equations, not only the mean estimation problem.

4.3 Example: SIMEX for the Regression Estimator of the Mean

As an example to provide intuition about how the solution to $E[\psi(Y, R, Z, W_{(1+\lambda)}, \theta)] = 0$ can converge to the solution of $E[\psi(Y, R, Z, X, \theta)] = 0$, we show it holds for the case of the regression estimator for a mean with a linear model with only one covariate X measured by $W = X + U$. We assume $Y = \delta_{*0} + \delta_{*1}X + \epsilon$, where ϵ has mean zero and is independent of X and R. We assume the linear model coefficients are estimated by least squares and the mean for the population is predicted using those estimates. Then ψ has three components:

$$\psi(Y, R, X, \theta) = \begin{pmatrix} \mu - \delta_0 - \delta_1 X \\ (Y - \delta_0 - \delta_1 X)R \\ X(Y - \delta_0 - \delta_1 X)R \end{pmatrix}$$

and $\theta = (\mu, \delta_0, \delta_1)$. This gives that $E[\psi(Y, R, W_{(1+\lambda)}, \theta)]$ equals

$$\begin{pmatrix} \mu - \delta_0 - \delta_1 E[W_{(1+\lambda)}] \\ (E[Y \mid R = 1] - \delta_0 - \delta_1 E[W_{(1+\lambda)} \mid R = 1])p(R = 1) \\ (E[W_{(1+\lambda)}Y \mid R = 1] - \delta_0 E[W_{(1+\lambda)} \mid R = 1] - \delta_1 E[W^2_{(1+\lambda)} \mid R = 1])p(R = 1) \end{pmatrix}. \quad (5)$$

Details given in the Appendix show that the solution to $E[\psi(Y, R, W_{(1+\lambda)}, \theta)] = 0$ has

$$\mu_{(1+\lambda)} = \delta_{*0} + \delta_{*1}(1 - \frac{\nu^2}{\nu^2 + (1+\lambda)\sigma^2})E[X \mid R = 1] + \frac{\delta_{*1}\nu^2}{\nu^2 + (1+\lambda)\sigma^2}E[X],$$

where $\nu^2 = E[X^2 \mid R = 1] - E[X \mid R = 1]^2$. It is clear that, as λ approaches -1, $\mu_{(1+\lambda)}$ will converge to $\delta_{*0} + \delta_{*1}E[X]$, which would be the value from the solution to $E[\psi(Y, R, X, \theta)] = 0$ and which is the population mean under the assumed model.

4.4 Monte Carlo Verification of Conditions

The linear regression example is atypical in that $E[\psi(Y, R, W_{(1+\lambda)}, \theta)]$ and its unique zero can be computed in closed form. In general these operations will be intractable. Thus it will not be feasible to directly verify, as we did above, that the required limiting behavior as $\lambda \to -1$ holds. Similarly, the sufficient conditions in the Appendix often cannot be verified. However, if $P_{(1)}$ and thus $P_{(1+\lambda)}$ are known, it is possible to check by simulation that the required conditions hold for any estimation method (i.e. any given set of estimating equations). Suppose that for select values of $\lambda \geq -1$, including -1, we can draw samples from $P_{(1+\lambda)}$. By the large sample results in the previous section, for very large n, the solution $\widehat{\theta}_{(1+\lambda)}$ to $n^{-1}\sum_{i=1}^{n}\psi(Y_i, R_i, Z_i, W_{(1+\lambda)i}) = 0$ is essentially equal to the solution $\theta_{(1+\lambda)}$ to $E[\psi(Y, R, Z, W_{(1+\lambda)})] = 0$. Thus, we can apply our estimation method to the very large simulated data set for each value of λ to obtain $\theta_{(1+\lambda)}$ and trace these as a function of λ to verify convergence as $\lambda \to -1$, and to explore the shape of the extrapolation function. Because the data are simulated, data for $-1 \leq \lambda < 0$ can be generated.

We used this procedure to verify that the required convergence holds for various mean estimators in the simulated scenario described in Section 2. We generated a sample with $n = 10^7$ observations. We generated $W_{(1+\lambda)}$ for 300 equally-spaced λ values between -1 and 2 inclusive. Figure 2 traces the solution to $E[\psi(Y^*, R, Z, W_{(1+\lambda)}, \theta)] = 0$ as a function of λ

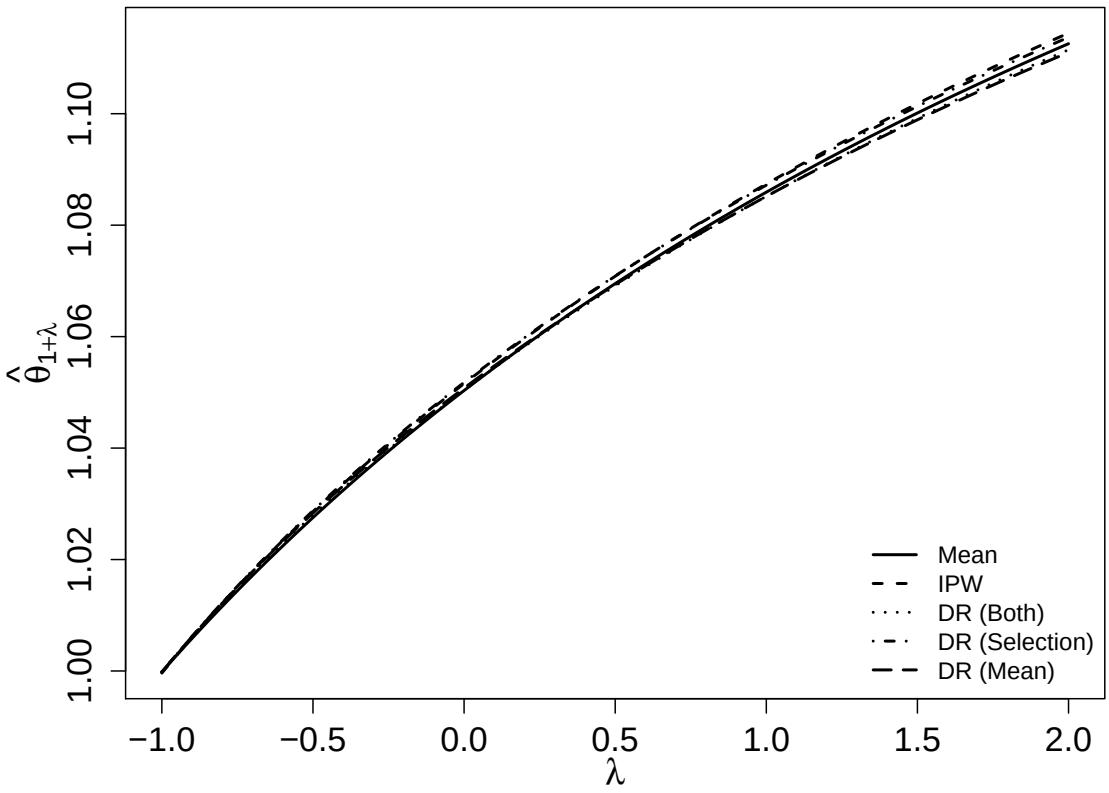

Figure 2: Extrapolation functions for five estimators for $E[Y^*]$: 1) the regression estimator; 2) the IPW estimator; 3) the bias-corrected DR estimator with correct propensity score function and correct regression function; 4) the bias-corrected DR estimator with correct propensity score function and incorrect regression function; and 5) the bias-corrected DR estimator with incorrect propensity score function and correct regression function.

for five estimators for $E[Y^*]$, the mean of the tobit outcome: 1) the regression estimator; 2) the IPW estimator; 3) the bias-corrected DR estimator with correct propensity score function and correct regression function; 4) the bias-corrected DR estimator with correct propensity score function and incorrect regression function, where the specified regression function included only W; and 5) the bias-corrected DR estimator with incorrect propensity score function and correct regression function, where the specified propensity score function included only W, but used the correct Cauchy link function[2]. The figure clearly shows that, although the extrapolation functions for the different mean estimators have slightly

2. For the estimators involving the mean function for Y^*, parameters of the tobit model were estimated using the tobit() function in the AER package (Kleiber and Zeileis, 2008) for the R environment, and $m(X, Z)$ was computed using methods appropriate for the tobit model.

different curvatures, all converge to the actual population mean of Y^* as $\lambda \to -1$. Similar results hold for the same set of five estimators applied to the linear outcome Y (Figure 8 in the Appendix) as well as for Y^* and Y where the Cauchy link for the propensity score model in Equation 2 was replaced by a logistic link (Figures 9 and 10 in the Appendix).

The large simulated dataset can also be used to assess the adequacy of different SIMEX extrapolation functions for any given estimator. For example, Figure 3 plots the true extrapolation function for the bias-corrected DR estimator with incorrect propensity score function and correct regression function. It also includes quadratic and quartic approximations fit to the values for $\lambda > 0$. Both fitted approximations fit the true extrapolation essentially perfectly for $\lambda > 0$ but the quadratic fit diverges by a small amount as λ approaches -1 and does not fully correct for the bias. This conservative behavior of the quadratic extrapolation function is a known general result for SIMEX (Carroll et al., 2006). Alternatively the quartic extrapolation diverges only trivially from the correct limiting value. The results are similar for the other estimators and settings, except that the divergences are smaller for the regression estimator than the others.

The obvious limitation to this approach in practice is that $P_{(1)}$ is unknown; if it were known then $E[Y]$ could be computed directly. However, the method could be used in applications with an approximation to $P_{(1)}$ informed by the observed data. For example, $\tilde{m}(X, Z)$ and/or $\tilde{\pi}(X, Z)$ could be specified, perhaps using models and parameters estimated with the data. Values for Z could be drawn from the empirical distribution, and values for X could be generated according to some hypothesized distribution conditional on Z. Next, Y and R could be generated using the working estimates of $\tilde{m}(X, Z)$ and $\tilde{\pi}(X, Z)$, and finally, $W_{(1+\lambda)}$ could be generated using the ME distribution evaluated at different values of λ. Sensitivity analyses could be done to explore alternative values for the models, parameters, and the distribution of X. These analyses could provide confidence that the required limiting behavior of the estimating equation solutions holds in models similar to those generating the observed data, and may inform the choice of the extrapolation function.

4.5 Standard Error Estimation for the SIMEX Estimator

The SIMEX estimator is a complicated function of the observed and simulated data. Thus some work is required to estimate standard errors. Carroll et al. (2006) suggest three general methods for SIMEX standard error estimation. The first is a resampling method such as the jackknife or bootstrap, in which the entire SIMEX procedure is replicated with resampled datasets. The second uses SIMEX projection itself to estimate the standard error. Suppose there is a typical way of estimating the sampling covariance matrix $\widehat{V}(\widehat{\theta}_{(1+\lambda_\ell),b,n})$ of $\widehat{\theta}_{(1+\lambda_\ell),b,n}$; for example, using a sandwich variance estimator. In addition, let $s^2_{(1+\lambda_\ell),B,n}$ equal the sample covariance matrix of $\{\widehat{\theta}_{(1+\lambda_\ell),b,n}\}_{b=1}^{B}$. Then Carroll et al. (2006) argue that projecting

$$(1/B)\sum_{b=1}^{B}\widehat{V}(\widehat{\theta}_{(1+\lambda_\ell),b,n}) - s^2_{(1+\lambda_\ell),B,n} \tag{6}$$

to $\lambda = -1$ approximates the sampling variance for the final SIMEX estimator. The third method is to express the SIMEX estimator as the solution to a system of estimating equations that include both the model parameters and the parameters of the extrapolation

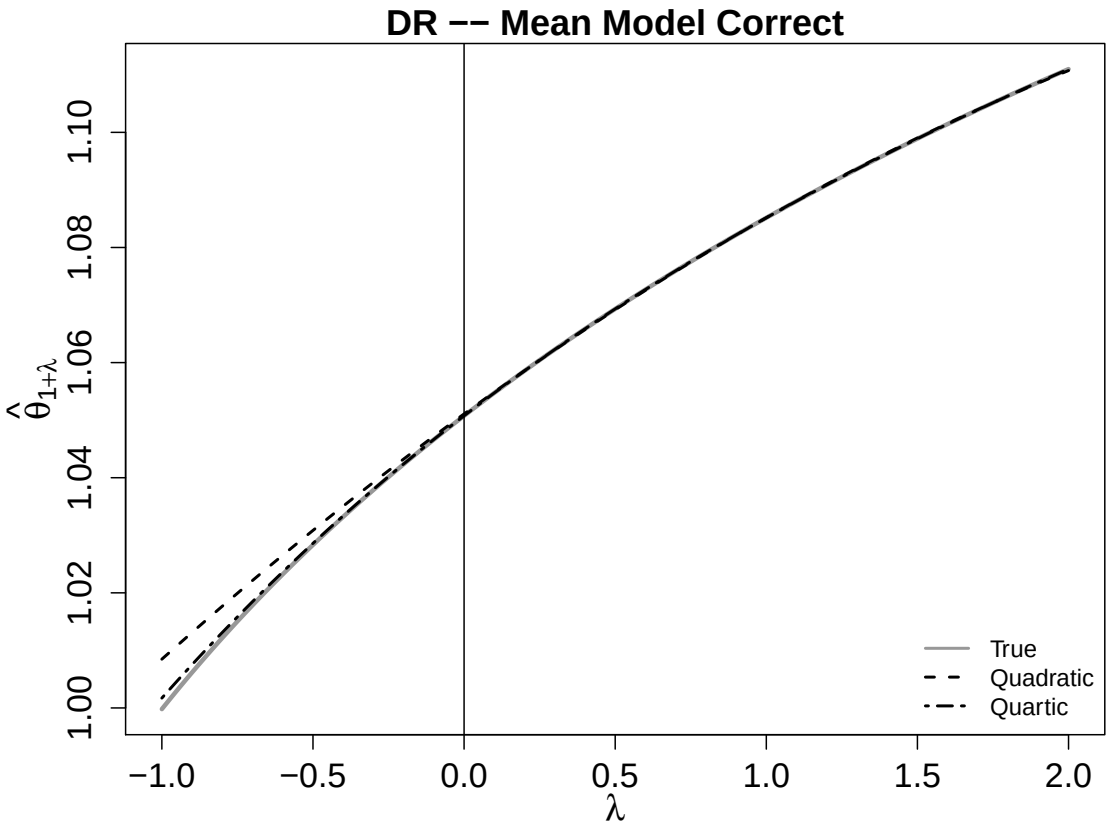

Figure 3: Extrapolation function, and quadratic and quartic approximations, for the bias-corrected DR estimator with incorrect propensity score and correct regression functions.

function, and to use asymptotic standard errors available from general estimating equation theory. Details of this approach are provided in Appendix B.4 of Carroll et al. (2006).

Each approach has advantages and disadvantages. The resampling method is straightforward to implement, albeit computationally burdensome, because the SIMEX procedure itself can be computationally intensive and each bootstrap or jackknife replication requires replicating the full SIMEX procedure. The projection method requires no resampling, but $\widehat{V}$ must be available, an extrapolation function must be chosen, and the final standard error approximation is valid only for large n and small ME. The full estimating equation approach is asymptotically correct, but it generally will be tedious to construct and check all of the required terms. In Section 5 we consider the performance of the bootstrap and projection methods in a simulation study, and in Section 6 we use the bootstrap in the case study of middle school effects on student achievement.

5. Simulation Study of Estimator Performance

The development to this point has not considered how well SIMEX mean estimators perform in finite samples. In this section we evaluate the performance of SIMEX estimators, and estimated standard errors, again in the context of the simulated scenario from Section 2.

5.1 Performance of SIMEX Estimators

We conducted four simulations using four different data sample sizes of $n = 500, 1000, 5000$ and 25000. For each sample size we generated $M = 500$ independent datasets, and for each such dataset we implemented 20 estimators for $E[Y]$, the mean of the linear outcome. These were determined by the five classes of estimators used in the previous section (regression, IPW, and three different versions of bias-corrected DR), and each was implemented in four different ways: 1) We applied the estimator to (Y, R, Z_1, Z_2, X). We call this the "ideal" approach. It would not be feasible in practice and we include it to calibrate the performance of the feasible estimators that use $W_{(1)}$. 2) We applied the estimator to $(Y, R, Z_1, Z_2, W_{(1)})$ without any correction for ME. We call this the "naïve" approach. 3) We applied the estimator to $(Y, R, Z_1, Z_2, W_{(1)})$ using SIMEX with the quadratic extrapolation function, which we label "SIMEX(2)"; and 4) we applied the estimator to $(Y, R, Z_1, Z_2, W_{(1)})$ using SIMEX with the quartic extrapolation function, which we label "SIMEX(4)". Each of the SIMEX estimators was implemented using $B = 500$ generations of the synthetic data and a λ grid consisting of 20 equally-spaced points from 0 to 2.

We used the M independent replications of the simulation to calculate the bias, standard error (SE) and root mean squared error (RMSE) of each of these 20 estimators for $E[Y] = 0$ and for each sample size n. These results are summarized in Figure 4. For each estimation class, the figure plots the bias, standard error, and RMSE. In each panel of the figure, the values for the ideal, naïve and SIMEX estimators are plotted by n. The scale of the vertical axis is 100 times the scale of the outcome variable, so that, for example, a value of "6" on the vertical axis corresponds to 0.06 outcome units. Because the outcomes were constructed to have population standard deviation 1, the values on the vertical axis can be thought of as percentages of a standard deviation unit of the outcome. The combination of $M = 500$, $B = 500$ and a λ grid consisting of 20 points led to Monte Carlo error in our estimates that was sufficiently small to not affect our substantive conclusions.

The ideal estimator is effectively unbiased for all estimation classes and has the smallest RMSE for every estimator class and sample size. Alternatively, the naïve estimator is always biased and has the largest RMSE for every estimator class and sample size. Both types of SIMEX estimators perform well under all conditions. SIMEX(2) removes most of the bias for all estimator classes. It is also relatively precise; its RMSE dominates the naïve estimator and is not much worse than the ideal estimator for all classes and sample sizes. SIMEX(4) removes virtually all of the bias, but the extra flexibility of the extrapolation function adds a little variance and thus it does not improve on SIMEX(2) with respect to RMSE until the sample size is very large.

We repeated the simulation for the tobit outcome Y^* and obtained similarly successful performance of SIMEX. The results are summarized in Figure 11 in the Appendix and are extremely similar to those presented in Figure 4. SIMEX is again competitive with the ideal in terms of RMSE for all estimator classes and sample sizes.

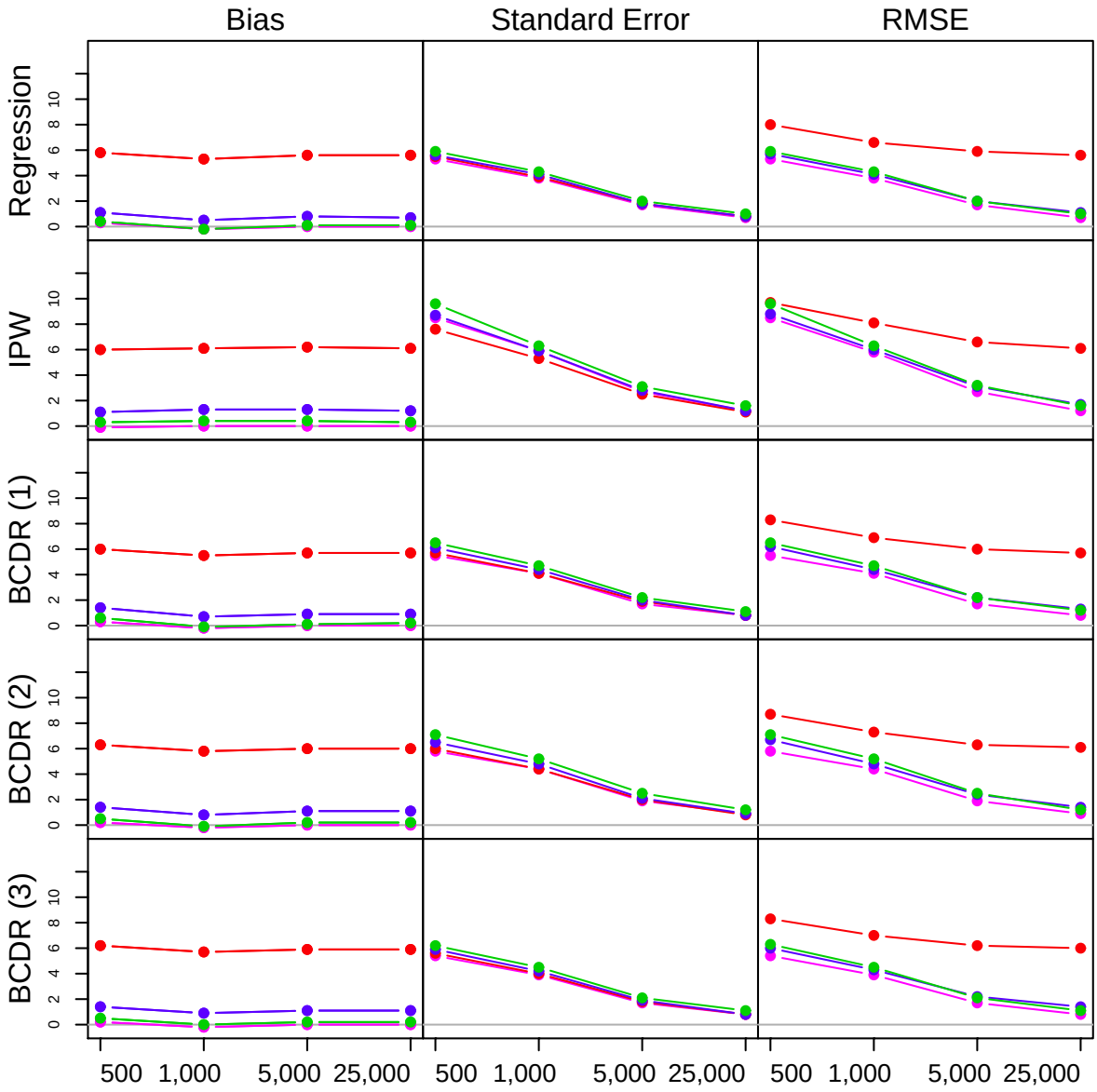

Figure 4: Performance of SIMEX estimators for mean $E[Y]$ of the linear outcome. Estimators include regression, IPW, bias corrected DR (BCDR) with correct mean and propensity score models (1), correct propensity score and incorrect mean models (2), correct mean and incorrect propensity score models (3). Within each panel the magenta line is for the "ideal" estimator, the red line is for the "naïve" estimator, the blue line is for "SIMEX(2)" with the quadratic extrapolation, and the green line is for "SIMEX(4)" with the quartic extrapolation. The scale of the vertical axis corresponds to percentages of a standard deviation unit of Y.

In addition to these 20 estimators assessed in Figures 4 and 11, we also evaluated the performance of the DR estimator of Kang and Schafer (2007) that uses weighted regression modeling with inverse propensity score weights. We did this using the same incorrect regression and incorrect propensity score function used for the bias corrected DR estimator. The performance of the ideal, naïve and SIMEX estimators for this DR class were virtually identical to those of the bias corrected form for both the linear and tobit outcomes.

Finally, we also considered a replication of the simulation reported in Figures 4 and 11 where the reliability of $W_{(1)}$ was 0.70 rather than 0.85. These are reported in Figures 12 and 13 in the Appendix. In both cases, the bias in the naïve estimators was much larger and both SIMEX(2) and SIMEX(4) still dominated the naïve estimators in terms of RMSE for all estimator classes and sample sizes. However unlike the case with reliability 0.85, SIMEX(4) had lower RMSE than SIMEX(2) for all estimator classes when the sample size was 1,000 or more. This is because the larger ME variance leads to more extreme nonlinearity in the true extrapolation function, so the bias reduction afforded by SIMEX(4) is relatively more important. The conclusion that a more flexible extrapolation function may be beneficial when the ME variance is larger probably generalizes to many other applications of SIMEX.

5.2 Performance of Standard Error Estimators

In this section we consider the performance of the bootstrap and projection methods for feasible standard error estimation. Figure 5 presents results for the linear outcome with $n = 500$ and reliability 0.85. The estimator classes are organized into the same five blocks as in Figure 4, and the ideal, naïve and two SIMEX estimators are labeled on the left in the same color scheme as used in Figure 4. For each of the 20 estimators, Figure 5 provides the following. The black filled circle is the standard deviation of the true sampling distribution of each estimator, estimated using the $M = 500$ Monte Carlo simulations described in the previous section. This value is used as the benchmark for the feasible estimators because it is correct up to Monte Carlo error from the 500 simulated datasets. To evaluate the performance of the bootstrap standard error estimator, we selected a random sample of 10 of the 500 simulated datasets, and for each of the samples, we then computed the bootstrap standard deviation of each of the mean estimators using 100 bootstrap samples. These values are plotted with the gray open circles, and their means across the 10 datasets are plotted with the gray triangle. To evaluate the performance of the projection standard error estimator, we used estimating equation theory to derive $\widehat{V}$ for the regression, IPW and bias-corrected DR estimators. Details for these computations, which are more broadly applicable to estimating standard errors of these population mean estimators in the case of no measurement error, are provided in the Appendix. We used these values to estimate the standard error of the ideal and naïve estimators for each of $M^* = 200$ random datasets, and also used these values to compute the projection standard error estimators for the SIMEX estimators. These values are plotted with the open dark yellow circles in Figure 5, and their means across the 200 datasets are given by the orange squares.

The bootstrap standard errors appear to function quite well. They appear to be essentially unbiased for the true standard errors for all estimator classes. The standard error estimators using $\widehat{V}$ appear to be reasonable for the ideal and naïve estimators, but slightly biased downward when used in the projection calculation for the SIMEX estimators, par-

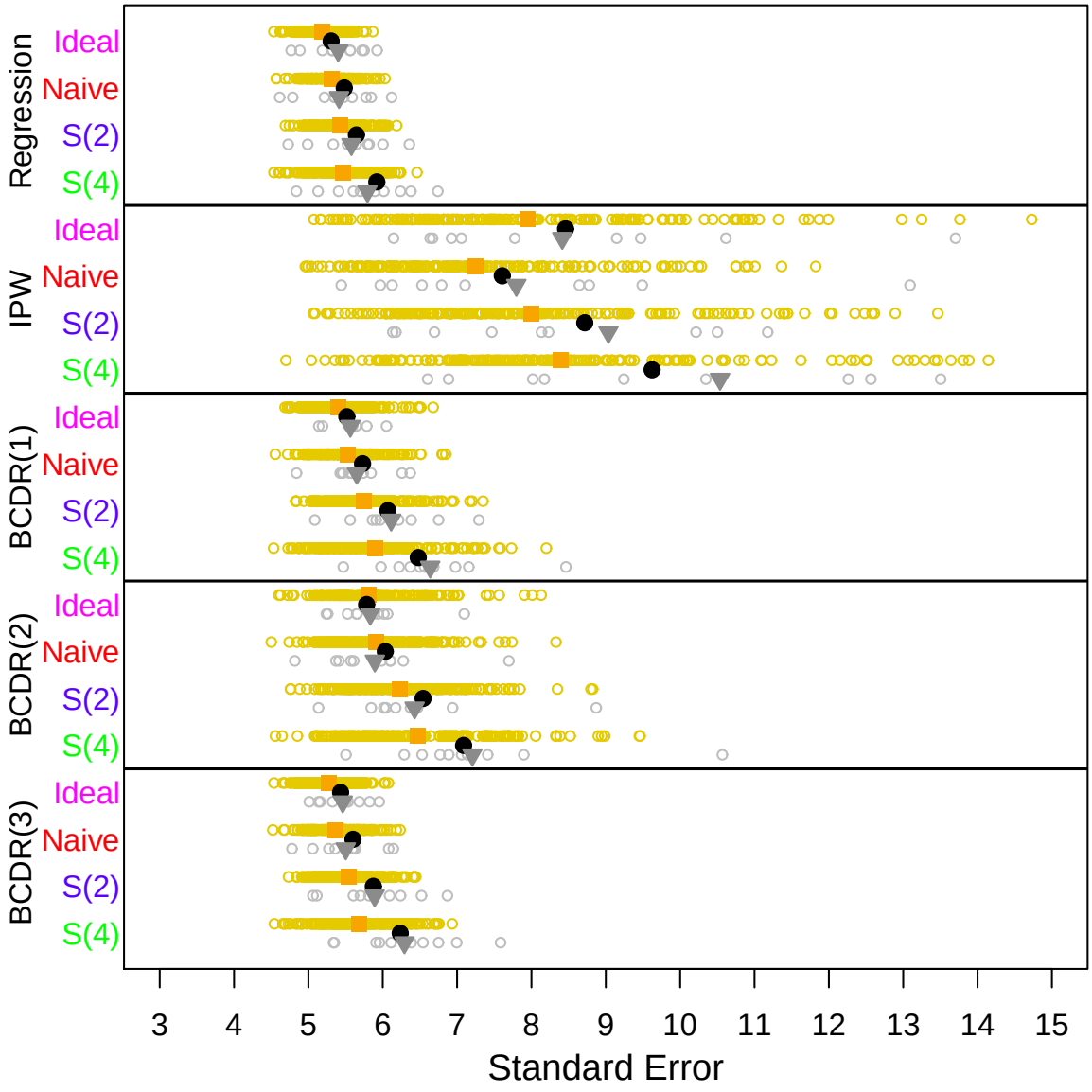

Figure 5: Summary of estimated standard errors for different estimators of the mean of the linear outcome, with $n = 500$ and reliability of $W_{(1)}$ equal to 0.85. Full description in the text.

ticularly for quartic SIMEX. This may be due to the fact that the projection calculation is generally valid only when the sample size is large and the ME variance is small. Plots similar to Figure 5 but for other scenarios including $n = 1000$, the tobit outcome, and reliability of 0.70 are given in Figures 14 to 20 in the Appendix. Due to computational limitations, we considered only $n = 500$ and 1000. These alternative scenarios replicate the

basic conclusions from the analyses presented in Figure 5. The bootstrap standard errors function well across scenarios, while the projection standard errors appear to be negatively biased, particularly for the quartic SIMEX estimator. These results suggest that the bootstrap is probably the safest approach to standard error estimation in practice, thus we use it in our empirical example in Section 6.

6. Example: Estimating Middle School Effects

This section summarizes an example in which we use SIMEX to estimate population average treatment effects on student math achievement for middle schools in a large suburban school district. The data include achievement test scores and demographic characteristics for 6,552 grade 6 students nested in 24 middle schools. The covariates Z_i for each student i include indicator variables for whether the student is black, Hispanic, male, eligible for special education, eligible for gifted instruction, and eligible for free or reduced price lunches (FRL). The observed grade 5 math and language test scores $(W_{i,\text{math}}, W_{i,\text{lang}})$ are assumed to be error-prone measures of latent grade 5 math and language achievement $(X_{i,\text{math}}, X_{i,\text{lang}})$. The observed test scores are on a standardized scale where the district mean is zero and the district standard deviation is 1, so that score points are in student-level standard deviation units. The number of grade 6 students in each middle school ranges from 142 to 371 (mean 273) and there is substantial variation across schools in the distributions of $(Z_i, W_{i,\text{math}}, W_{i,\text{lang}})$. For example, the school-level averages of $W_{i,\text{math}}$ range from about -0.42 to about 0.70, so that schools range in average prior math achievement by more than 1 student-level standard deviation unit of observed scores. Similarly, the school percentages of FRL-eligible students range from 6% to 85%.

The outcome of interest is the grade 6 math achievement test score, also standardized to have mean zero and standard deviation 1 in the district. Consistent with the potential outcomes framework for causal effects, we assume that each student i has a potential grade 6 math achievement outcome $Y_{i,s}$ for each school $s = 1, \ldots, 24$, indicating what the grade 6 test score would have been for student i had he or she attended school s [3].

We define the causal effect of school s for student i as $Y_{i,s} - (1/24)\sum_{t=1}^{24} Y_{i,t}$; i.e the difference between the potential outcome under school s and the average of the potential outcomes across all schools. We assume that for each s, $Y_{i,s}$ are IID samples from the population of students and we denote $E[Y_{i,s}]$ by μ_s, the mean grade 6 test score that would be observed if all students in the population attended school s. Using the definition of the individual causal effects, the population average effect of school s is $\Delta_s = \mu_s - \overline{\mu}$ where $\overline{\mu}$ is the equally-weighted mean of μ_s across the 24 schools. Estimating $\{\Delta_s\}$ requires estimating the 24 population means μ_s in the presence of nonrandom missing data because $Y_{i,s}$ are observed only for the corresponding subpopulations of students who attended each school. That is, because each student attended only one school, $Y_{i,s}$ is observed only for the school s that student i attended, and $Y_{i,t}$ is missing for the 23 schools where $t \neq s$.

3. Implicit in our use of potential outcomes is the SUTVA or no-interference assumption that students' outcomes do not depend on other students assigned to their school (Rubin, 1980). The plausibility of this assumption can be questioned. But regardless of whether or not it holds, failure to properly control for selection effects due to ME in the covariates would contribute to bias in the estimated school effects.

Grade 6 is the first year of middle school in this district. The grade 5 test scores $(W_{i,\text{math}}, W_{i,\text{lang}})$ were measured while the students were still in elementary school. Because middle school selection decisions are unlikely to be based on the observed test scores, it is reasonable to assume that observed differences across middle schools in grade 5 test scores reflect differences in latent achievement $(X_{i,\text{math}}, X_{i,\text{lang}})$. It is also reasonable to assume that the relationship between the grade 6 test scores and $(W_{i,\text{math}}, W_{i,\text{lang}}, Z_i)$ is driven by the relationship between the grade 6 test scores and $(X_{i,\text{math}}, X_{i,\text{lang}}, Z_i)$. Thus, given the available covariates, it is probably more reasonable to assume strong ignorability holds conditional on $(X_{i,\text{math}}, X_{i,\text{lang}}, Z_i)$ than to assume that it holds conditional on the observed covariates $(W_{i,\text{math}}, W_{i,\text{lang}}, Z_i)$, making the methods developed in this article appropriate.

A challenge with applying SIMEX in this context is that test score ME is heteroskedastic with a complicated error structure. The observed test scores $(W_{i,\text{math}}, W_{i,\text{lang}})$ are estimates of latent abilities $(X_{i,\text{math}}, X_{i,\text{lang}})$ obtained from item response theory (IRT) models (Lord, 1980; van der Linden and Hambleton, 1997). Such estimates have the property that the ME variance of $W_{i,\text{math}}$, for example, varies as a function of $X_{i,\text{math}}$ according to a known function $g_{\text{math}}(X_{i,\text{math}})$ called the squared "conditional standard error of measurement (CSEM)" function (Lord, 1980). The CSEM function for a given test is determined by the number of items and their psychometric characteristics. These functions are available in our data. Using the methods of Lockwood and McCaffrey (2014), we estimate that the average reliabilities of the grade 5 math and language test scores in this population are 0.87 and 0.84, respectively. These are consistent with the values considered in our simulation studies.

The difficulty with using the CSEM function in SIMEX is that the ME variance for any individual student's score is unknown because it depends on the latent achievement. It is thus unclear what ME variance to use for each test score during the simulation phase of SIMEX. Lockwood and McCaffrey (2015b) evaluate different plug-in estimators for the unknown variances $g_{\text{math}}(X_{i,\text{math}})$ and $g_{\text{lang}}(X_{i,\text{lang}})$ for their effectiveness in SIMEX applications. Their analyses suggest that treating estimates of $E[g_{\text{math}}(X_{i,\text{math}})|W_{i,\text{math}}]$ and $E[g_{\text{lang}}(X_{i,\text{lang}})|W_{i,\text{lang}}]$ as the known variance for each test score results in SIMEX estimators with reasonable properties. They discuss how to estimate these quantities from the observed data, and we used those methods here to compute ME variances $\sigma^2_{i,\text{math}}$ and $\sigma^2_{i,\text{lang}}$ to use for each student's grade 5 test scores during SIMEX.

To estimate the effects, we used quadratic SIMEX with the bias-corrected DR estimator to estimate each of the 24 means μ_s and then used these to estimate the effects $\{\Delta_s\}$. For each school we specify the mean as linear in $(X_{i,\text{math}}, X_{i,\text{lang}}, Z_i)$, which is reasonable given the generally high linear correlations between test scores from the same students across years. We specify the propensity score model using a linear predictor in $(X_{i,\text{math}}, X_{i,\text{lang}}, Z_i)$ and the Cauchy link function. The SIMEX estimates for $\{\mu_s\}$ were obtained using $B = 500$ generations of the synthetic data and a λ grid consisting of 20 equally-spaced points from 0 to 2 for each school. Standard errors were computed using 100 bootstrap samples of the students and applying quadratic SIMEX to estimate $\{\mu_s\}$ for each sample. The bootstrap distributions of the estimated effects $\{\Delta_s\}$ were unimodal and approximately symmetric, so that 95% confidence intervals using ± 1.96 bootstrap standard errors are suitable.

Figure 6 presents the estimated $\{\Delta_s\}$, where schools are sorted by their estimated effects. It also gives 95% bootstrap confidence intervals. The effects for most schools are statistically significantly different from zero and range from approximately -0.3 to 0.3 standard

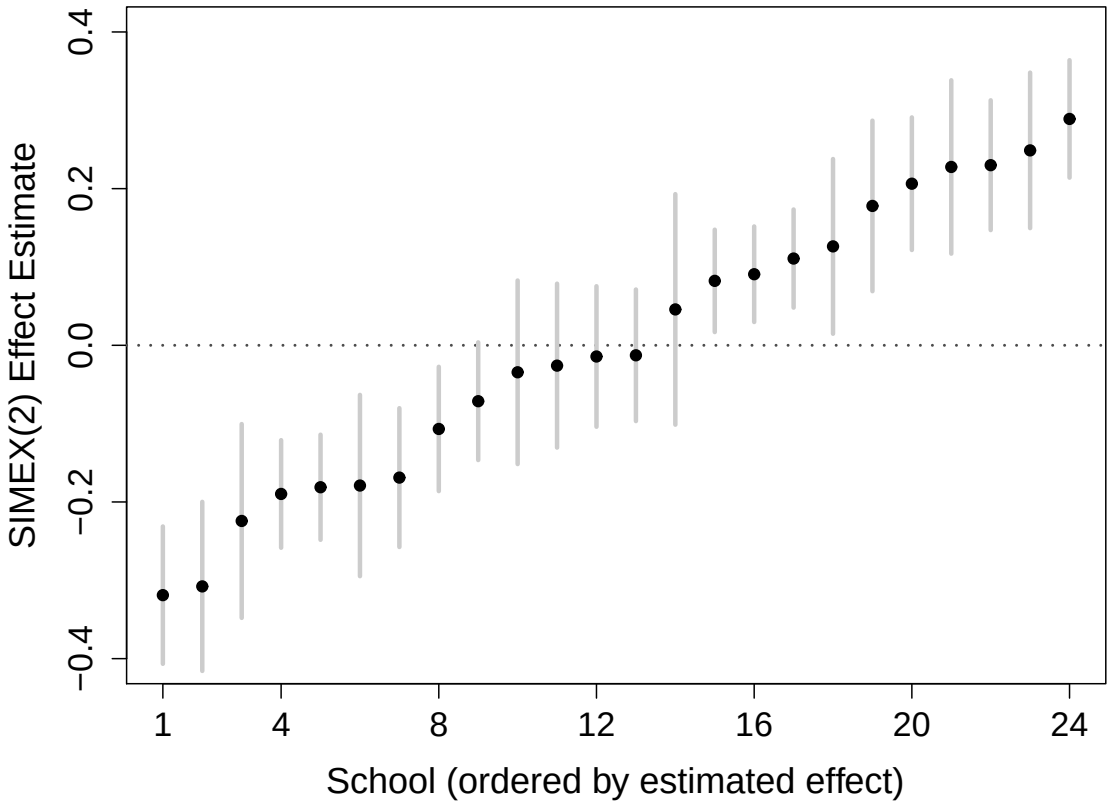

Figure 6: School effects estimated by quadratic SIMEX, with 95% bootstrap confidence intervals given by gray bars.

deviation units of the grade 6 math score. A method of moments estimate of the variance among schools effects (Morris, 1983) indicates that schools account for about 3% of the total variance in the grade 6 math scores. This is consistent with other research on the magnitude of between-school variation in achivement growth (e.g., Kane and Staiger, 2002).

Because the truth is unknown, we have to use indirect evidence to assess whether SIMEX improved the estimates of $\{\Delta_s\}$ relative to the naïve estimates. One of the consequences of failing to correct for covariate ME in estimates of group effects is that the resulting bias in the estimates tends to be positively correlated with group means of the latent covariates (Fuller, 2006) - that is, estimates for schools with high means of $(X_{\text{math}}, X_{\text{lang}})$ are too high and those for schools with low means are too low. Therefore, if SIMEX is effective at reducing bias in the estimated school effects, it should be the case that SIMEX tends to decrease the naïve estimates for schools with high average prior achievement and increase the naïve estimates for schools with low average prior achievement. To examine this, we follow Lockwood and McCaffrey (2014) by using our longitudinal data to compute the school averages of the grade 5 test scores from the prior year cohort of grade 6 students.

This is an entirely different group of students than those used to compute the school effects. Thus any relationship of the differences between the naïve and SIMEX estimates based on the estimation cohort of students, and the school averages of the incoming test scores for the prior year cohort, indicates systematic differences, rather than any spurious correlation resulting from computing estimates for a group of students and then relating the estimates to characteristics of those same students.

Figure 7 plots the differences between quadratic SIMEX estimated school effects and

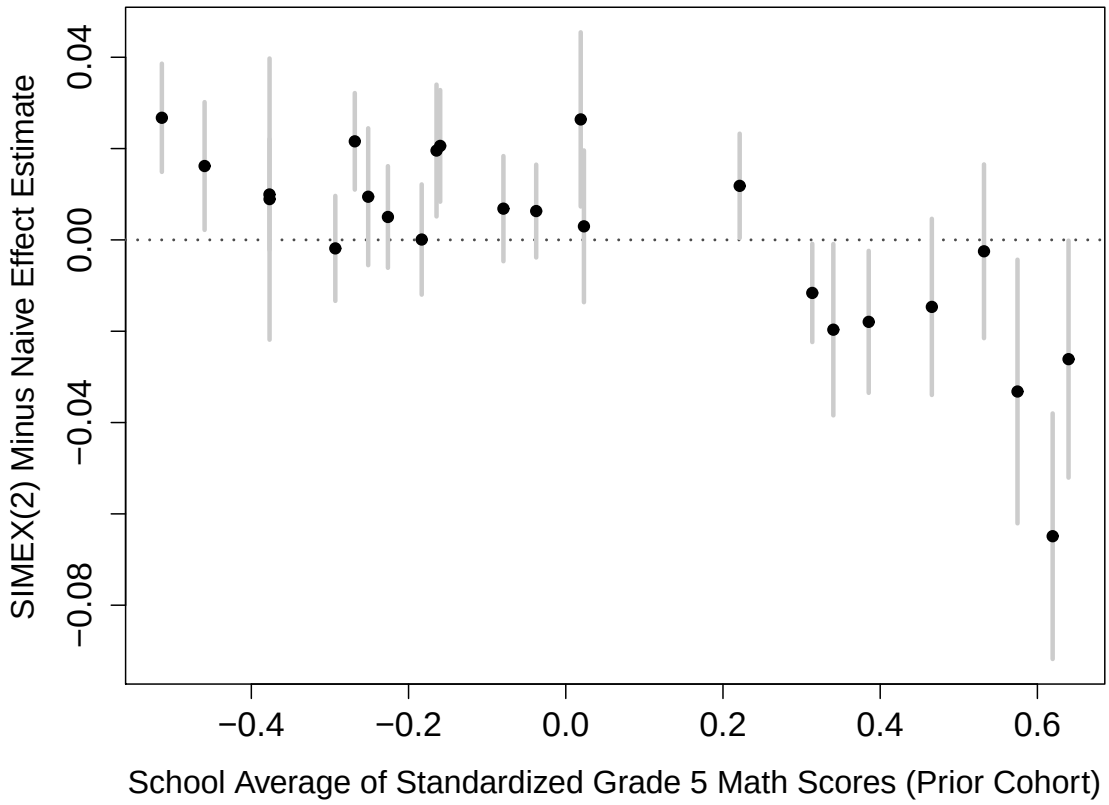

Figure 7: Differences between quadratic SIMEX estimated school effects and naïve estimated school effects as a function of school averages of standardized grade 5 math scores from the prior cohort of students. 95% bootstrap confidence intervals for the differences given by gray bars.

naïve estimated school effects as a function of school averages of grade 5 math scores from the prior cohort of students. An advantage of using the bootstrap over alternative standard error computations is that it automatically provides uncertainty estimates for complicated functions of the data and estimated parameters; in this case we present 95% bootstrap confidence intervals for the differences between the SIMEX and naïve estimated school effects. The figure demonstrates the predicted pattern of SIMEX tending to increase the estimates

for schools with low average prior achievement and decrease the estimates for schools with high average prior achievement. Many of these changes are statistically significant. The adjusted R^2 of a quadratic regression of the differences on the average prior math scores is 0.66 with a 95% bootstrap confidence interval of (0.51, 0.81), clearly indicating a relationship between general achievement attributes of schools' students and the corresponding changes in the estimated effects due to ME correction. This suggests that the SIMEX estimates are indeed less biased than the naïve estimates.

Although the adjustments due to SIMEX are sensible, they are of relatively modest magnitude in this example. The standard deviation among the SIMEX estimated school effects is 0.182 standard deviation units of student achievement, and the standard deviation of the changes is 0.021, so that the magnitude of the changes is about 12% as large as the magnitude of the effects. The modest change in this context makes sense given that the prior tests are fairly reliable and there are two of them, helping to reduce bias in the naïve estimators caused by ME (Lockwood and McCaffrey, 2014). However, given that a primary concern about school or teacher accountability measures is that they will contain systematic errors correlated with student background characteristics such as prior achievement (Braun, 2005), there may be value in removing this bias even if its magnitude is only modest.

7. Discussion

Estimation of a population mean from a nonrandom sample with error-prone covariates can be achieved through a straightforward application of the SIMEX method. However, it is distinct from common applications of SIMEX, which typically estimate the parameters of parametric models. Estimating a population mean from a nonrandom sample requires estimating parameters as well as evaluating functions of those parameters at various values of the data. An advantage of SIMEX is that no special adaptations of the method are necessary for this novel application in which alternative approaches to dealing with error-prone covariates are generally challenging.

Our theoretical development addressed the case where the ME is homoskedastic with variance σ^2. However, heteroskedastic measurement error commonly arises in applications, including psychometrics (Battauz and Bellio, 2011), economics (Sullivan, 2001) and biostatistics (see Delaigle & Meister, 2007 and references therein). SIMEX extends straightforwardly to the case where the ME variance is heteroskedastic with known variances σ_i^2 for each unit by using σ_i^2 to generate the synthetic data for each unit. However, it is common in applications for σ_i^2 to be unknown, such as the example with IRT-based achievement scores in Section 6. Appropriate SIMEX methods for the case of unknown σ_i^2 when replicate measures W_{ij} are available for each X_i are provided by Carroll and Wang (2008), Devanarayan and Stefanski (2002) and Wang et al. (2009), and could be applied in the current setting. As noted in Section 6, applying SIMEX in the heterskedastic case where the ME variance is a known function of X_i is addressed by Lockwood and McCaffrey (2015b).

The need to know or make assumptions about the ME distribution is a fundamental limitation of correcting regression, weighting or DR estimators of a population mean for covariate measurement error. Outside of limited special cases, both the methods reviewed in Section 3 and SIMEX require such information. How this information is obtained will depend on the setting. In some cases validation data containing X itself are available, and in

other cases there may be replicate error-prone measures from the same unit that can be used to assess the ME distribution (Carroll et al., 2006). As noted in Section 6, in educational and psychometric applications involving trait estimates from IRT or factor analytic models, information about ME distribution is commonly available from the analysis methods used to create the observed scores. The validity of ME-corrected estimators such as SIMEX hinges on the ME assumptions being reasonable. With SIMEX and other methods, using a value of σ^2 that is too small will tend to undercorrect the naïve estimator for ME, and using a value that is too large will tend to overcorrect. A useful feature of SIMEX is that the projection process provides insight into how much, and in what direction, ME matters for the target estimand. If σ^2 is unknown, the SIMEX procedure can be applied with different plausible values in a sensitivity analysis. As in Section 6, the bootstrap could be used in such settings to create a confidence interval for the difference between the uncorrected and corrected estimates.

Another difficulty with any method in this context is the correct specification of either $\pi(X, Z)$ or $m(X, Z)$ when there is ME. This is because ME obscures relationships (Carroll et al., 2006). In particular, common methods for building the propensity score model, such as adding terms until the $R = 0$ and $R = 1$ covariate distributions are balanced, do not work because balance of $W_{(1)}$ does not guarantee balance of X. Due to these challenges to model building and the increased likelihood of model misspecification, DR may be comparatively more attractive in the presence of ME than in the standard case. DR is also appealing over IPW when there is ME because ME can reduce the efficiency of the estimator relative to estimation with error-free covariates. The regression component of the DR estimator can improve efficiency and potentially offset the extra uncertainty created by ME. SIMEX is particularly valuable for DR estimation because implementing DR using the methods of McCaffrey and Lockwood (2014), which were detailed in Section 3, is challenging in practice, whereas SIMEX is straightforward.

A key limitation specific to SIMEX is the projection to $\lambda = -1$. Several analyst decisions are required to conduct this projection, including the choice of the number B of synthetic data replications, the grid of λ values, and the functional form of the projection function. Each of these choices can affect the final estimator in terms of bias, variance, or both. The choice of the projection function is most difficult. For example, as demonstrated in Figure 3, different extrapolation functions may fit the generated data essentially equivalently but still lead to different projections. If the estimates are relatively insensitive to the values of λ, projection may risk introducing variance without sufficient bias reduction to compensate. Work needs to be done on using data from the sample of estimates across values of λ to decide whether or not to project, and what projection function to use. Alternatively, combinations of the naïve and SIMEX estimators might yield more accurate estimates, and data from the sample of estimates across values of λ may be used to determine how to weight the two components. Whether or not such methods could improve on SIMEX remains an area for future research, but such methods may be particularly valuable in the estimation of means with nonrandom samples because such estimates can be noisy and because SIMEX may be the most applicable method for handling ME in such problems.

Future work might also consider alternative methods for weight estimation such as boosted models (McCaffrey et al., 2004b), covariate balancing propensity scores (Imai and Ratkovic, 2014), the super learner (Pirracchio et al., 2015), or minimum discriminant infor-

mation adjustment (Haberman, 1984; Hainmueller, 2012). These approaches tend to yield weights which better reduce differences among groups or yield more precise estimates of the mean than traditional weights based on logistic regression. SIMEX is likely to be the most practically feasible solution for many of these approaches in applications with error-prone covariates, and determining if their benefits extend to such applications is an area for future research. Similarly, matching is also sometimes used for estimating means in the context of causal modeling. Lockwood and McCaffrey (2015a) show that matching in the presence of error-prone variables can be quite challenging. SIMEX would be straightforward to apply to matching, and the close relationship between matching and weighting suggests that the resulting estimators may perform well, but additional study is needed.

Acknowledgments

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D140032 to ETS. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We thank Bob Mislevy, Lili Yao, Rebecca Zwick, an Associate Editor and an anonymous reviewer for helpful comments on an earlier draft.

Appendix

Data Generation Code for Simulation Examples

```
datagen <- function(n=1000, reliability=0.85, seed){
  set.seed(seed)

  ## Generate X using mixture of normals and set E[X]=0 and Var[X]=1
  mixprops <- c(.275, .475, .0666, .0667, 0.0667, 0.05)
  stopifnot(sum(mixprops)==1)
  mus   <- c(0, -2, 2.25, 3.25, 4.25, -6)
  vs    <- c(1, 1, .25, .25, .25, .25)
  xmean <- sum(mixprops*mus)
  xvar  <- sum(mixprops*vs) + sum(mixprops*(mus-xmean)^2)

  mix   <- sample(1:length(mixprops), size=n, replace=TRUE, prob=mixprops)
  d     <- data.frame(x = rnorm(n, mean=mus[mix], sd=sqrt(vs[mix])))
  d$x   <- (d$x - xmean)/sqrt(xvar)

  ## Generate Z1 and Z2
  d$z1  <- .3*d$x + sqrt(1-.3^2)*rnorm(n)
  d$z2  <- as.integer(runif(n) < 0.5)

  ## Generate error-prone measure W
  sigsq <- (1-reliability)/reliability
  d$w   <- d$x + rnorm(n, sd=sqrt(sigsq))

  ## Generate R
  d$p   <- pt(0.5 + 1.2*d$x + 0.5*d$z1 - 1.0*d$z2 + 0.7*d$x*d$z2, df = 1)
  d$R   <- as.integer(runif(n) < d$p)

  ## Generate Y (linear outcome) subject to E[Y]=0, Var[Y]=1, and covariates
  ## having R^2=0.80.
  ## "vary" is true population variance of explained part of Y
  ## "vare" is selected to give R^2 = 0.80
  d$y   <- -0.20 + 1.0*d$x + 0.6*d$z1 + 0.4*d$z2
  vary  <- 1.0^2 + 0.6^2 + 2*(1.0)*(0.6)*0.3 + (0.4^2)*(1/4)
  vare  <- 0.25*vary
  d$y   <- d$y + rnorm(n, mean=0, sd=sqrt(vare))
  d$y   <- (1/sqrt(vary + vare)) * d$y

  ## Generate Y* (tobit outcome conditional on covariates) by censoring a linear
  ## transformation of Y where the parameters of the linear transformation are
  ## chosen so that E[Y*] and Var[Y*] are both approximately 1
  d$ystar  <- (0.812 + 1.255*d$y) * as.numeric(0.812 + 1.255*d$y > 0)
  d$ystar0 <- as.integer(d$ystar==0)
```

```
  ## print summaries
  cat("\nMarginal\n")
  print(t(sapply(d, function(x){ c(mean = mean(x), sd = sd(x))})))

  cat("\nR=0 vs R=1 means\n")
  print(t(sapply(d, function(x){ tapply(x, d$R, mean) })))

  return(d)
}
```

Technical Conditions for Section 4

Let $Q = (Y, R, Z)$ to simplify notation. Here we establish sufficient conditions for the solution of $E[\psi(Q, W_{(1+\lambda)}, \theta)] = 0$ to converge to the solution of $E[\psi(Q, X, \theta)] = 0$. Assume:

1. θ takes values in a compact set $\Theta \subset \mathcal{R}^{K+1}$
2. $\psi(Q, x, \theta)$ is continuous in x for every $\theta \in \Theta$
3. $E[\psi(Q, W_{(1+\lambda)}, \theta)]$ and $E[\psi(Q, X, \theta)]$ exist for every $\theta \in \Theta$, and $E[\psi(Q, W_{1+\lambda}, \theta)]$ is continuous in θ
4. For each λ, $\theta_{(1+\lambda)}$ is the unique element of Θ satisfying $E[\psi(Q, W_{(1+\lambda)}, \theta)] = 0$, and θ_* is the unique element of Θ satisfying $E[\psi(Q, X, \theta)] = 0$
5. $\lim_{\lambda \to -1} E[\psi(Q, W_{1+\lambda}, \theta)]$ exists for every $\theta \in \Theta$
6. The following interchange of limit and integration holds for each $\theta \in \Theta$:

$$\lim_{\lambda \to -1} E[\psi(Q, W_{(1+\lambda)}, \theta)] = E[\lim_{\lambda \to -1} \psi(Q, W_{(1+\lambda)}, \theta)]$$

   and this convergence is uniform in θ.

Under these conditions $\lim_{\lambda \to -1} \theta_{(1+\lambda)}$ exists and is equal to θ_*.

Proof: As $\lambda \to -1$, $(Q, W_{(1+\lambda)}) \xrightarrow{a.s.} (Q, X)$. Assumption 2 and the continuous mapping theorem give that $\psi(Q, W_{(1+\lambda)}, \theta) \xrightarrow{a.s.} \psi(Q, X, \theta)$ for every $\theta \in \Theta$. By Assumptions 3, 5 and 6, the functions $h_{(1+\lambda)}(\theta) = E[\psi(Q, W_{(1+\lambda)}, \theta)]$ are continuous in θ and converge uniformly to $h_{(0)}(\theta) = E[\psi(Q, X, \theta)]$ as $\lambda \to -1$. Assumption 1 guarantees that $\{\theta_{(1+\lambda)}\}$ has at least one limit point $\tilde{\theta} \in \Theta$ and there is a subsequence $\{\theta_{(1+\lambda^*)}\}$ that converges to $\tilde{\theta}$. Assumption 6 gives that for any $\epsilon > 0$, for λ^* sufficiently close to -1, $|h_{(1+\lambda^*)}(\theta) - h_{(0)}(\theta)| < \epsilon$ for all θ. In particular then $|h_{(1+\lambda^*)}(\theta_{(1+\lambda^*)}) - h_{(0)}(\theta_{(1+\lambda^*)})| = |h_{(0)}(\theta_{(1+\lambda^*)})| < \epsilon$ so that $\lim_{\lambda^* \to -1} h_{(0)}(\theta_{(1+\lambda^*)}) = 0$. The uniform convergence also gives that $h_{(0)}(\theta)$ is continuous and so $h_{(0)}(\tilde{\theta}) = 0$. By Assumption 4, θ_* is the unique 0 of $h_{(0)}$ and so $\tilde{\theta} = \theta_*$ must be the unique limit of $\{\theta_{(1+\lambda)}\}$.

Assumption 1 is common in developing asymptotics for parametric models (e.g. Wald, 1949). The remaining assumptions are rather benign other than Assumption 6. The interchange may be reasonable given that standard estimating equation asymptotics require ψ to have a second moment at least at the true value of the parameter (Stefanski and Boos, 2002). The uniform convergence would generally be very strict but might be less so given Assumption 1. In any case, pointwise convergence of functions is not sufficient to guarantee convergence of zeros.

Details for Linear Regression Estimator in Section 4

Here we provide the details on the solution to $E[\psi(Y, R, W_{1+\lambda}, \theta)] = 0$ for the linear regression example of Section 4, where $E[\psi(Y, R, W_{1+\lambda}, \theta)]$ is given in Equation 5. Solving $(E[Y \mid R = 1] - \delta_0 - \delta_1 E[W_{1+\lambda} \mid R = 1])p(R = 1) = 0$ yields $\delta_{0,1+\lambda} = E[Y \mid R = 1] - \delta_{1,1+\lambda} E[W_{1+\lambda} \mid R = 1]$. Also, $E[Y \mid R = 1] = \delta_{*0} + \delta_{*1} E[X \mid R = 1]$ and $E[W_{1+\lambda} \mid R = 1] = E[X \mid R = 1]$, since U and U_λ are independent of R. This gives $\delta_{0,1+\lambda} = \delta_{*0} + \delta_{*1} E[X \mid R = 1] - \delta_{1,1+\lambda} E[X \mid R = 1]$.

We can now use $(E[W_{1+\lambda} Y \mid R = 1] - \delta_0 E[W_{1+\lambda} \mid R = 1] - \delta_1 E[W_{1+\lambda}^2 \mid R = 1])p(R = 1) = 0$ to solve for $\delta_{1,1+\lambda}$. Plugging in $\delta_{0,1+\lambda}$ for δ_0 gives

$$\begin{aligned}
0 &= E[W_{1+\lambda} Y \mid R = 1] - \delta_{0,1+\lambda} E[W_{1+\lambda} \mid R = 1] - \delta_{1,1+\lambda} E[W_{1+\lambda}^2 \mid R = 1] \\
&= E[W_{1+\lambda} Y \mid R = 1] - (E[Y \mid R = 1] - \delta_{1,1+\lambda} E[W \mid R = 1]) E[W_{1+\lambda} \mid R = 1] - \\
&\qquad \delta_{1,1+\lambda} E[W_{1+\lambda}^2 \mid R = 1] \\
&= E[W_{1+\lambda} Y \mid R = 1] - E[Y \mid R = 1] E[W_{1+\lambda} \mid R = 1] - \\
&\qquad \delta_{1,1+\lambda} (E[W_{1+\lambda}^2 \mid R = 1] - E[W_{1+\lambda} \mid R = 1]^2) \\
&= \delta_{*0} E[W_{1+\lambda} \mid R = 1] + \delta_{*1} E[W_{1+\lambda} X \mid R = 1] - \delta_{*0} E[W_{1+\lambda} \mid R = 1] - \\
&\qquad \delta_{*1} E[X \mid R = 1] E[W_{1+\lambda} \mid R = 1] - \delta_{1,1+\lambda} (E[W_{1+\lambda}^2 \mid R = 1] - E[W_{1+\lambda} \mid R = 1]^2).
\end{aligned}$$

This yields

$$\delta_{1,1+\lambda} = \frac{\delta_{*1}(E[W_{1+\lambda} X \mid R = 1] - E[X \mid R = 1] E[W_{1+\lambda} \mid R = 1])}{E[W_{1+\lambda}^2 \mid R = 1] - E[W_{1+\lambda} \mid R = 1]^2}.$$

We also have $E[W_{1+\lambda} X \mid R = 1] = E[X^2 \mid R = 1]$, so the numerator equals $E[X^2 \mid R = 1] - E[X \mid R = 1]^2$. Furthermore, $E[W_{1+\lambda}^2 \mid R = 1] = E[X^2 \mid R = 1] + E[(U + U_\lambda)^2 \mid R = 1] = E[X^2 \mid R = 1] + (1 + \lambda)\sigma^2$. This gives,

$$\delta_{1,1+\lambda} = \frac{\delta_{*1}(E[X^2 \mid R = 1] - E[X \mid R = 1]^2)}{E[X^2 \mid R = 1] - E[X \mid R = 1]^2 + (1 + \lambda)\sigma^2}.$$

We let $\nu^2 = E[X^2 \mid R = 1] - E[X \mid R = 1]^2$ and turn to $\mu - \delta_0 - \delta_1 E[W_{1+\lambda}] = 0$ to obtain

$$\mu_{1+\lambda} = \delta_{*0} + \delta_{*1}\left(1 - \frac{\nu^2}{\nu^2 + (1 + \lambda)\sigma^2}\right) E[X \mid R = 1] + \frac{\delta_{*1}\nu^2}{\nu^2 + (1 + \lambda)\sigma^2} E[X],$$

where the last term holds because $E[W_{1+\lambda}] = E[X]$.

Demonstration of Consistency of SIMEX

This section demonstrates extrapolation functions for other outcomes and selection models, analogous to Figure 2 in Section 4.4.

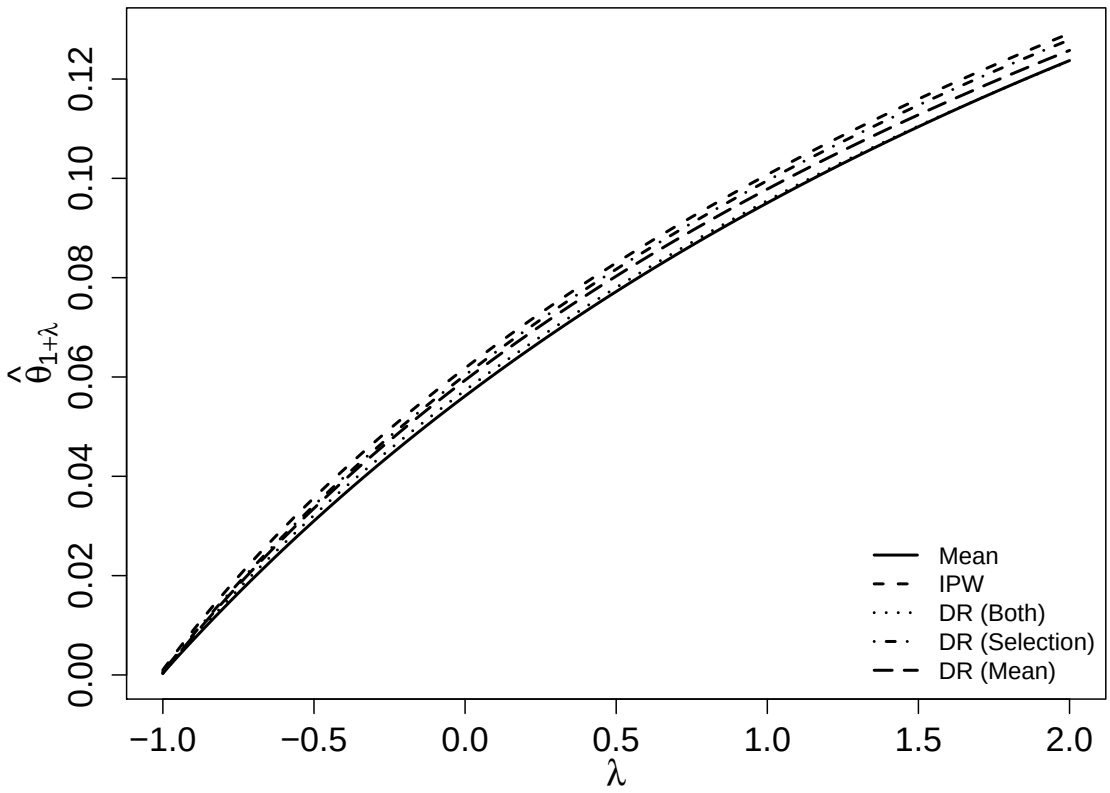

Figure 8: Extrapolation functions for simulated example analogous to Figure 2, but for the linear outcome.

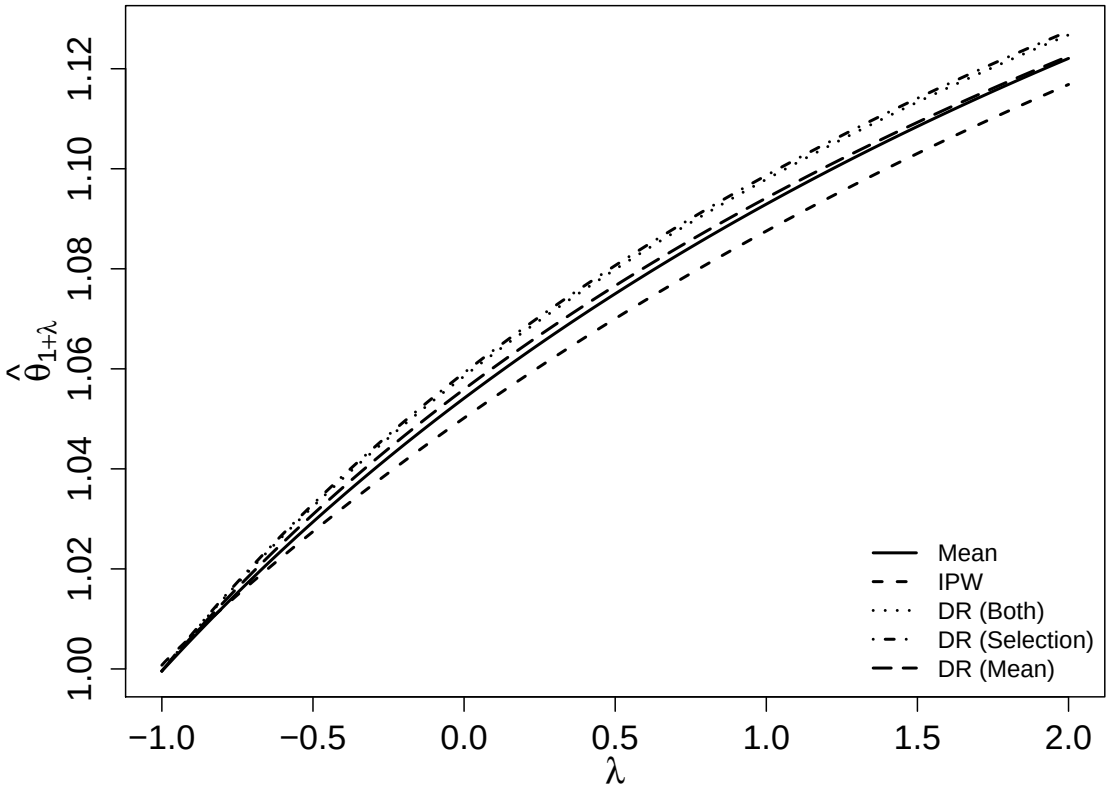

Figure 9: Extrapolation functions for simulated example analogous to Figure 2, but for the logistic propensity score function.

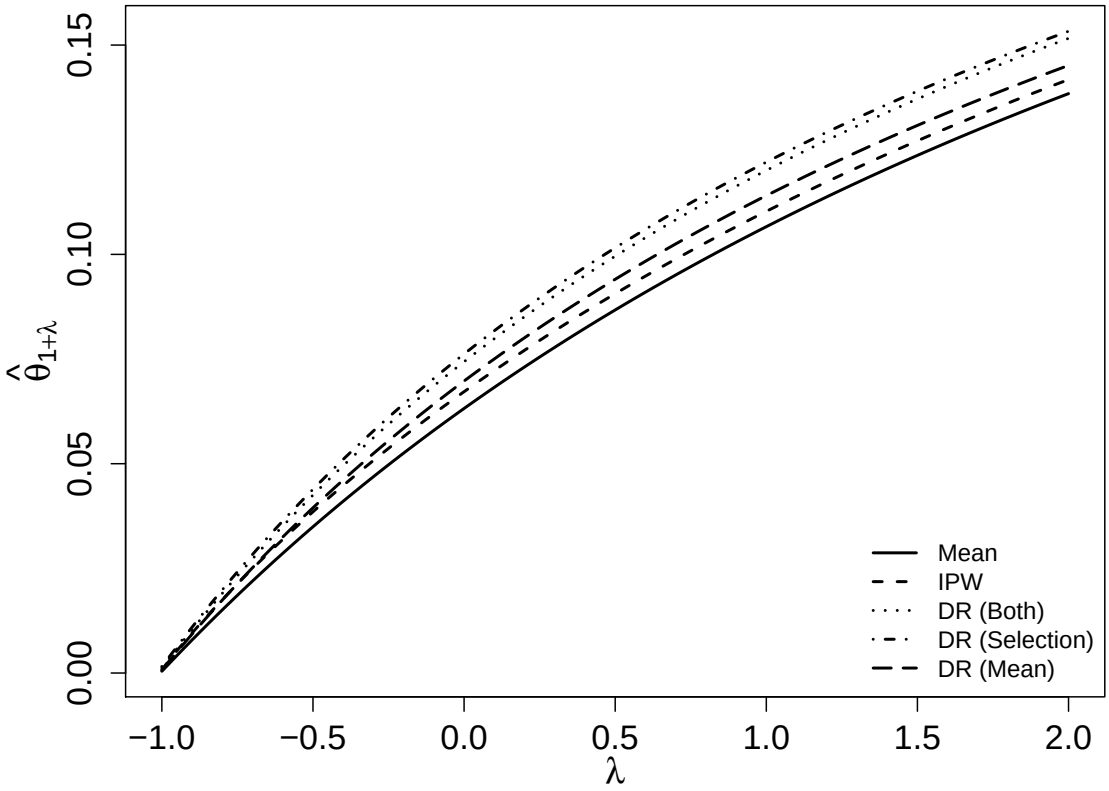

Figure 10: Extrapolation functions for simulated example analogous to Figure 2, but for the linear outcome and the logistic propensity score function.

Performance of Estimators Under Alternative Scenarios

This section presents summaries of estimator performance for alternative outcomes and covariate reliability, analogous to Figure 4 in Section 5.

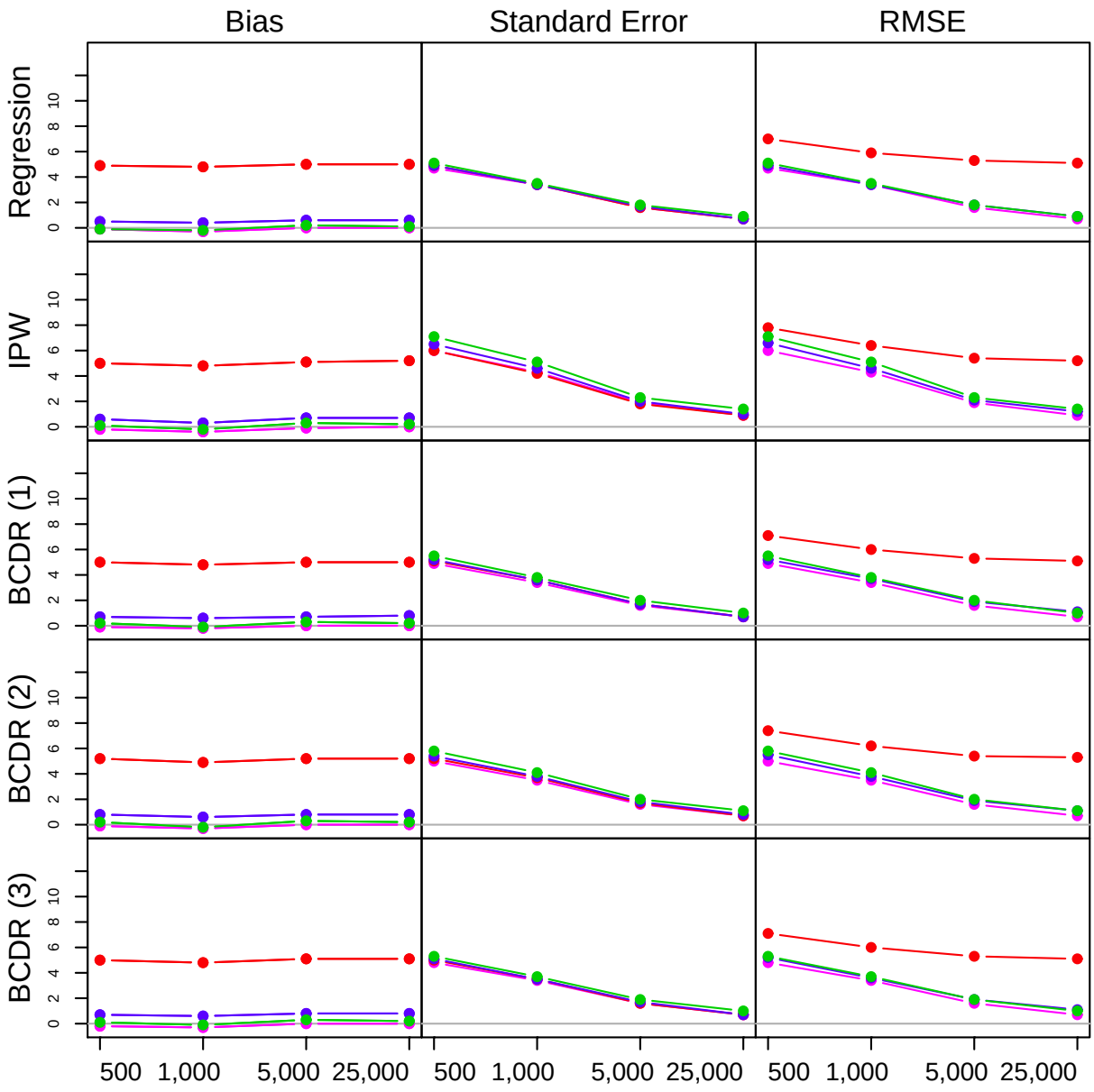

Figure 11: Summary of simulation results analogous to Figure 4, but for tobit outcome. Lines: magenta – "ideal"; red – naïve; blue – SIMEX(2); green – SIMEX(4).

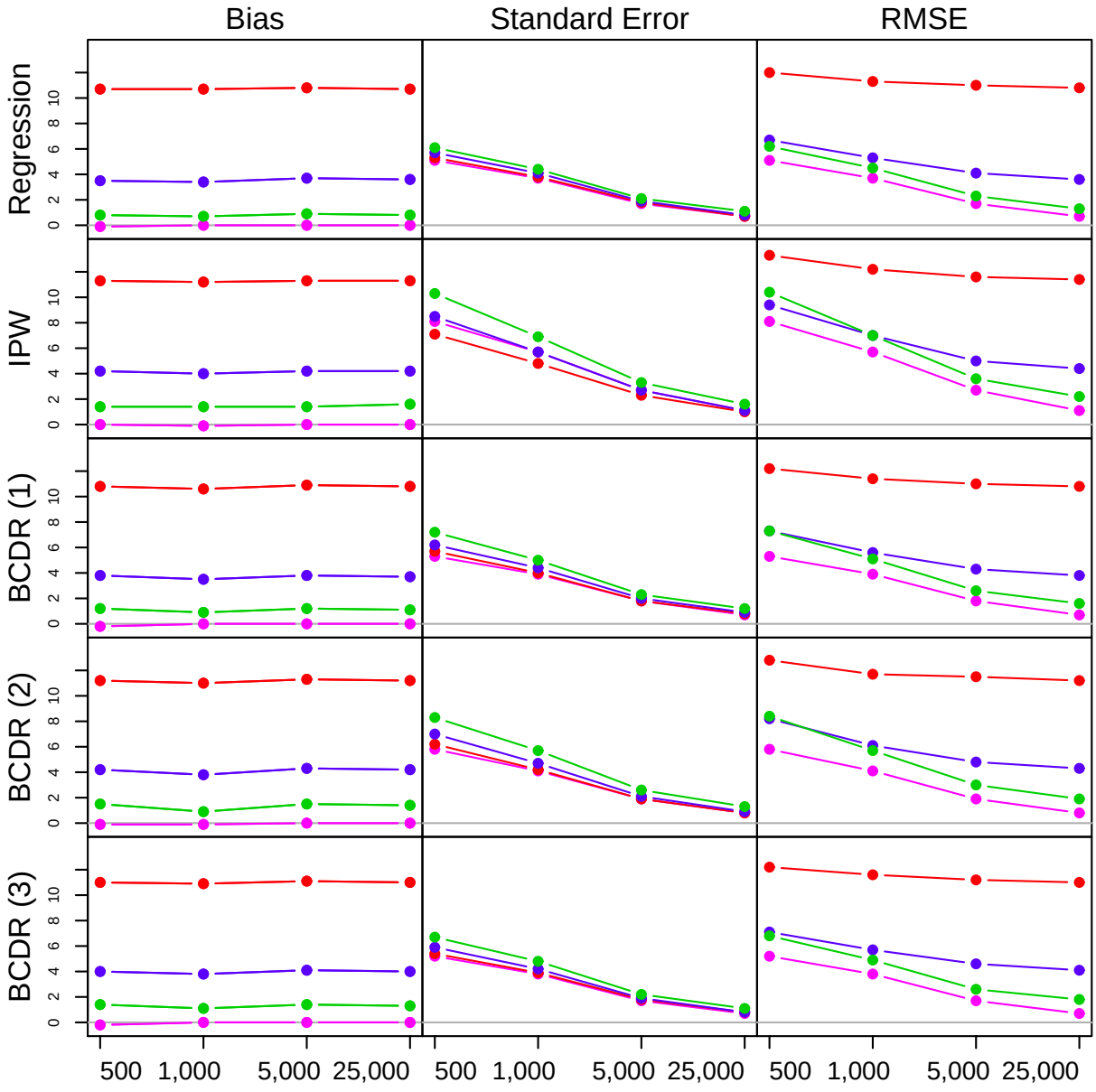

Figure 12: Summary of simulation results analogous to Figure 4, but with reliability of $W_{(1)}$ equal to 0.70. Lines: magenta – "ideal"; red – naïve; blue – SIMEX(2); green – SIMEX(4).

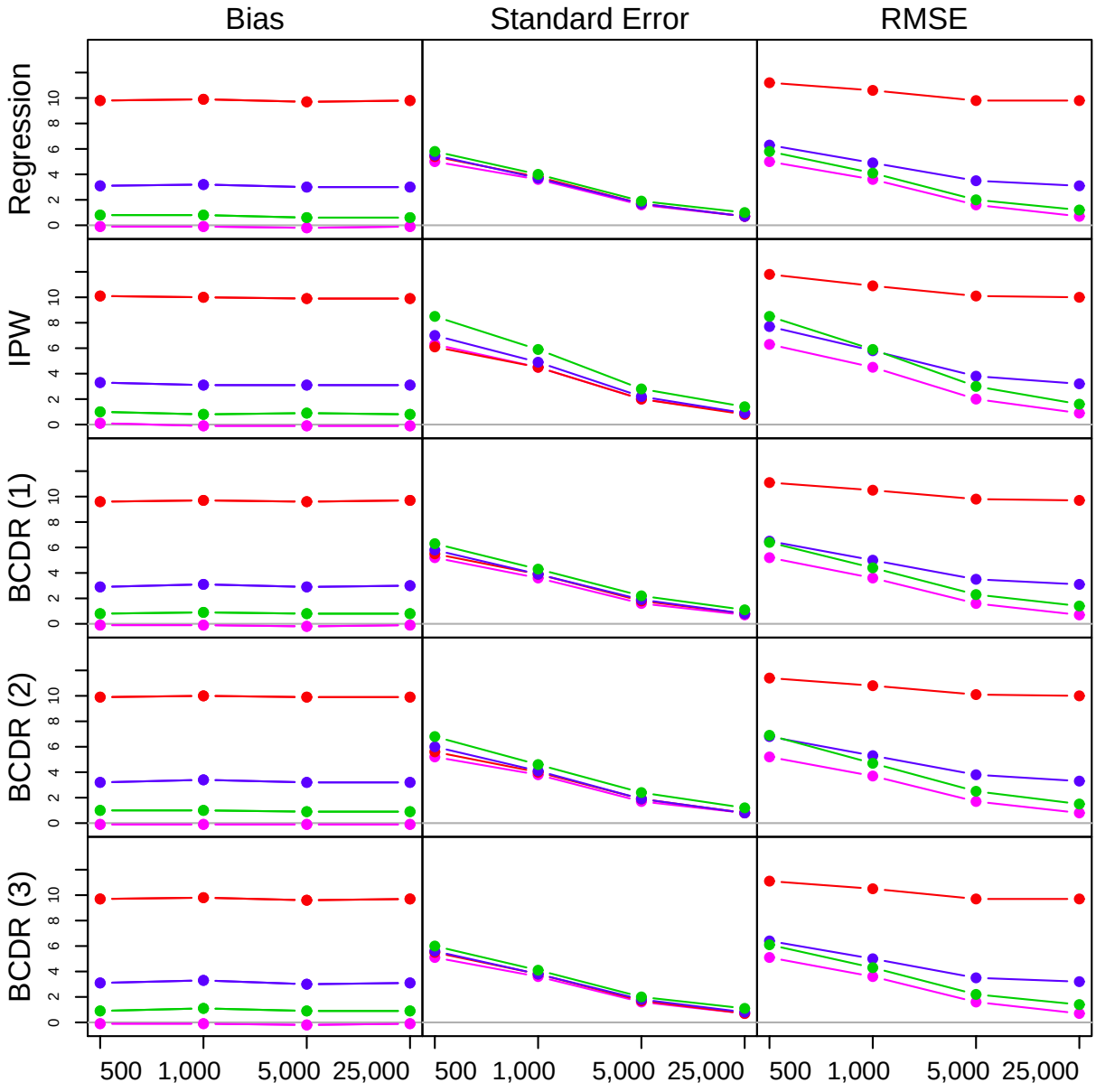

Figure 13: Summary of simulation results analogous to Figure 4, but with tobit outcome and with reliability of $W_{(1)}$ equal to 0.70. Lines: magenta – "ideal"; red – naïve; blue – SIMEX(2); green – SIMEX(4).

Details for Sandwich Standard Error Computations

In this section we use standard results for estimating equations to develop variance estimators for each of the population mean estimators. These values can then be inserted into the projection formula in Equation 6 to obtain standard error estimators for the SIMEX estimators. As noted in Section 3, $\widehat{\theta}_n$ solves $n^{-1}\sum_{i=1}^n \psi(Y_i, R_i, Z_i, X_i, \theta) = 0$, for the appropriate vector of functions ψ. For an estimator $\widehat{\theta}_n$ that solves $n^{-1}\sum_{i=1}^n \psi(Y_i, R_i, Z_i, X_i, \theta) = 0$, the asymptotic variance can be estimated consistently by $n^{-1}\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{A}^{-1}\right)'$, where $\mathbf{A} = n^{-1}\sum_{i=1}^n -\dot{\psi}(Y_i, R_i, Z_i, X_i, \widehat{\theta}_n)$, with $\dot{\psi}(Y, R, Z, X, \theta) = \partial\psi(Y, R, Z, X, \theta)/\partial\theta'$, and $\mathbf{B} = n^{-1}\sum_{i=1}^n \psi(Y_i, R_i, Z_i, X_i, \widehat{\theta}_n)\psi(Y_i, R_i, Z_i, X_i, \widehat{\theta}_n)'$. With error-prone measurements, including the measurements with additional error used during SIMEX, the same formulas can be used to estimate the asymptotic variance $\widehat{V}(\widehat{\theta}_{(1+\lambda_\ell),b,n})$ of $\widehat{\theta}_{(1+\lambda_\ell),b,n}$ by replacing $\widehat{\theta}_n$ with $\widehat{\theta}_{(1+\lambda_\ell),b,n}$ and X_i with $W^*_{(1+\lambda_\ell)i,b}$ in the expressions for $\mathbf{A}$ and $\mathbf{B}$.

We now derive ψ and $\dot{\psi}$ for each of the estimators. For these derivations we let we $D = (1, Z', X')'$. A similar development is provided by Hirano and Imbens (2001), and a summary of general results on estimating standard errors of parameters estimated via estimating equations is given by Stefanski and Boos (2002).

Regression

As shown in the article, for the regression estimator, with a linear model for the mean, the estimating functions for $\theta = (\mu, \delta')'$ are

$$\psi(Y, R, D, \theta) = \begin{pmatrix} \mu - D'\delta \\ R(Y - D'\delta)D \end{pmatrix}.$$

This yields

$$\dot{\psi}(Y, R, D, \theta) = \begin{pmatrix} 1 & -D' \\ 0 & -RDD' \end{pmatrix}.$$

Inverse Probability-of-Response Weighting

For IPW, we assume that the propensity scores are modeled by a generalized linear model with a Cauchy link. If $G(u)$ is the CDF of the Cauchy distribution, then $G(u) = \pi^{-1}\arctan(u) + 0.5$, $\dot{G}(u) = \pi^{-1}(1 + u^2)^{-1}$. The estimating functions for $\theta = (\mu, \alpha')'$ are

$$\psi(Y, R, D, \theta) = \begin{pmatrix} R(Y - \mu)[G(D'\alpha)^{-1}] \\ (R - G(D'\alpha))\dot{G}(D'\alpha)\left[G(D'\alpha)(1 - G(D'\alpha))\right]^{-1} D \end{pmatrix}.$$

This yields

$$\dot{\psi}(Y, R, D, \theta) = \begin{pmatrix} -RG(D'\alpha)^{-1} & -R(Y - \mu)\dot{G}(D'\alpha)[G(D'\alpha)^{-2}]D' \\ 0 & \mathcal{Q}(R, D, \alpha)DD' \end{pmatrix},$$

where

$$\mathcal{Q}(R, D, \alpha) = -\left\{R\left[\frac{1}{G(D'\alpha)^2} + \frac{2\pi(D'\alpha)}{G(D'\alpha)}\right] + (1 - R)\left[\frac{1}{(1 - G(D'\alpha))^2} - \frac{2\pi(D'\alpha)}{1 - G(D'\alpha)}\right]\right\}\dot{G}(D'\alpha)^2.$$

Doubly Robust

For the DR estimators, we again assume that the propensity scores are modeled by a generalized linear model with a Cauchy link, and that a linear model is used for the mean. We also allow for the covariates (or functions of the covariates, such as interactions) used to model the mean to differ from those used to model the propensity score. We let D_1 denote the vector of variables used in modeling the mean, including 1 for the intercept, and we let D_2 denote the vector of variables used in modeling the propensity score, again including 1 for the intercept. The estimating functions for $\theta = (\mu, \delta', \alpha')'$ are then given by

$$\psi(Y, R, D_1, D_2, \theta) = \begin{pmatrix} R(Y - D_1'\delta)[G(D_2'\alpha)^{-1}] - \tau(\mu - D_1'\delta) \\ R(Y - D_1'\delta)D_1 \\ (R - G(D_2'\alpha))\dot{G}(D_2'\alpha)\left[G(D_2'\alpha)(1 - G(D_2'\alpha))\right]^{-1} D_2 \\ R[G(D_2'\alpha)^{-1}] - \tau \end{pmatrix}.$$

This yields

$$\dot{\psi}(Y, R, D_1, D_2, \theta) = \begin{pmatrix} -\tau & (\tau - R[G(D_2'\alpha)^{-1}])D_1' & -R(Y - D_1'\delta)\dot{G}(D_2'\alpha)[G(D_2'\alpha)^{-2}]D_2 & D_1'\delta - \mu \\ 0 & -RD_1D_1' & 0 & 0 \\ 0 & 0 & \mathcal{Q}(R, D_2, \alpha)D_2D_2' & 0 \\ 0 & 0 & -R\dot{G}(D_2'\alpha)[G(D_2'\alpha)^{-2}]D_2' & -1 \end{pmatrix}.$$

Performance of Standard Error Estimators Under Alternative Scenarios

This section presents summaries of performance of standard error estimators for alternative outcomes, sample sizes and covariate reliability, analogous to Figure 5 in Section 5. Note that for the tobit outcome, in Figures 17 to 20, standard errors using the projection method are not presented because we did not compute the analytical sandwich standard errors for the tobit outcome mean estimators.

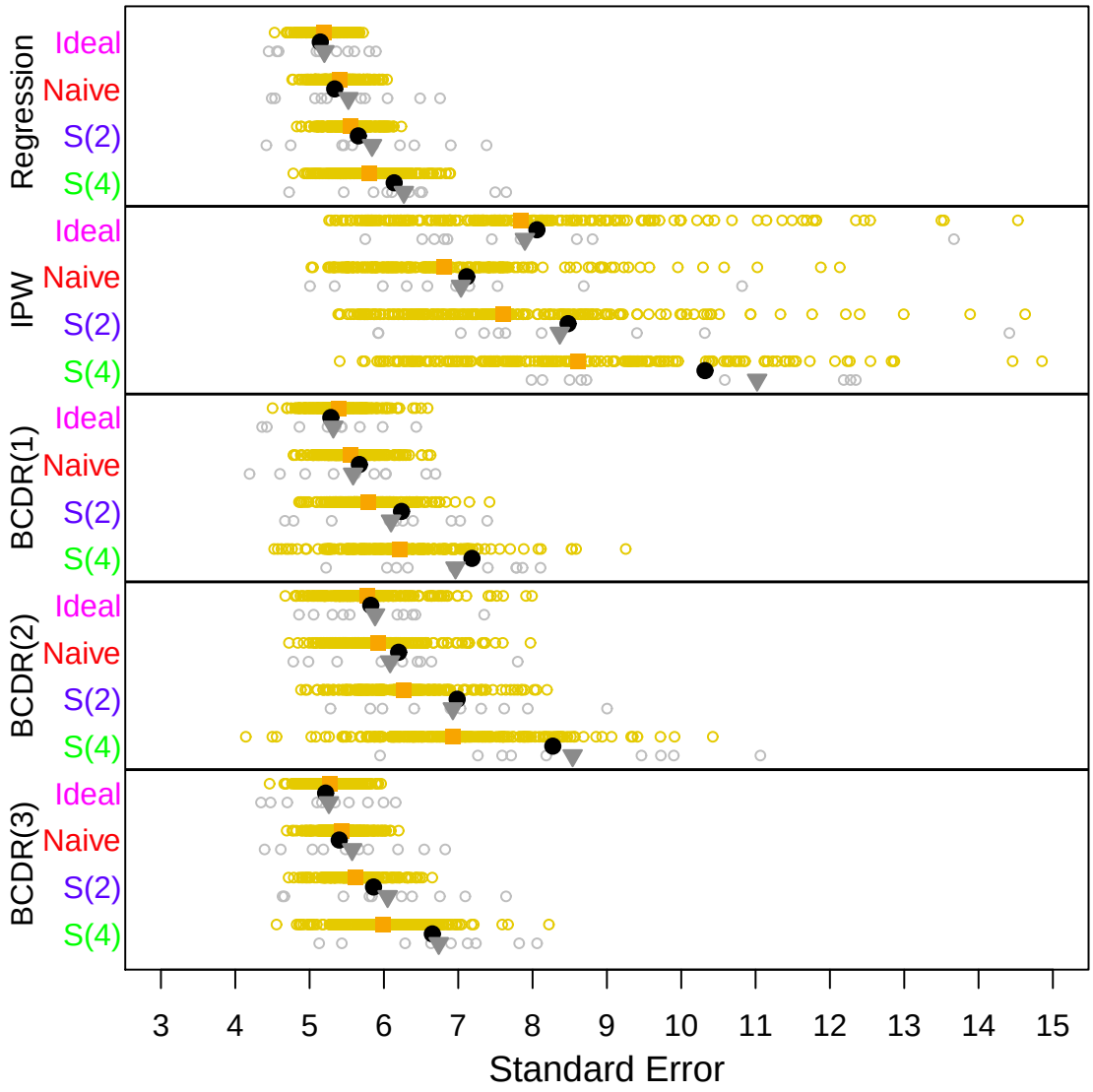

Figure 14: Summary of estimated standard errors for different estimators of the mean of the linear outcome, with $n = 500$ and reliability of $W_{(1)}$ equal to 0.70. Analogous to Figure 5.

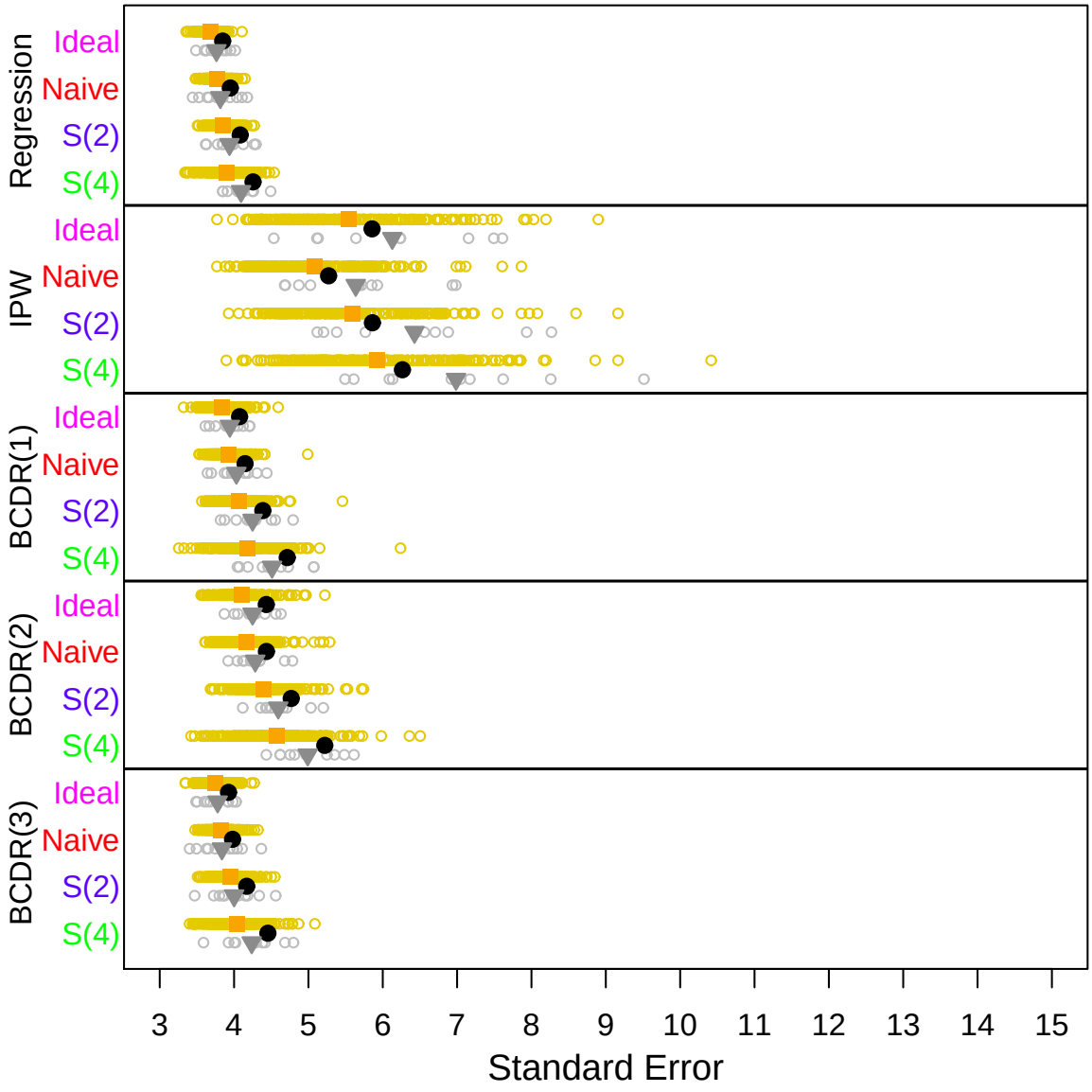

Figure 15: Summary of estimated standard errors for different estimators of the mean of the linear outcome, with $n = 1000$ and reliability of $W_{(1)}$ equal to 0.85. Analogous to Figure 5.

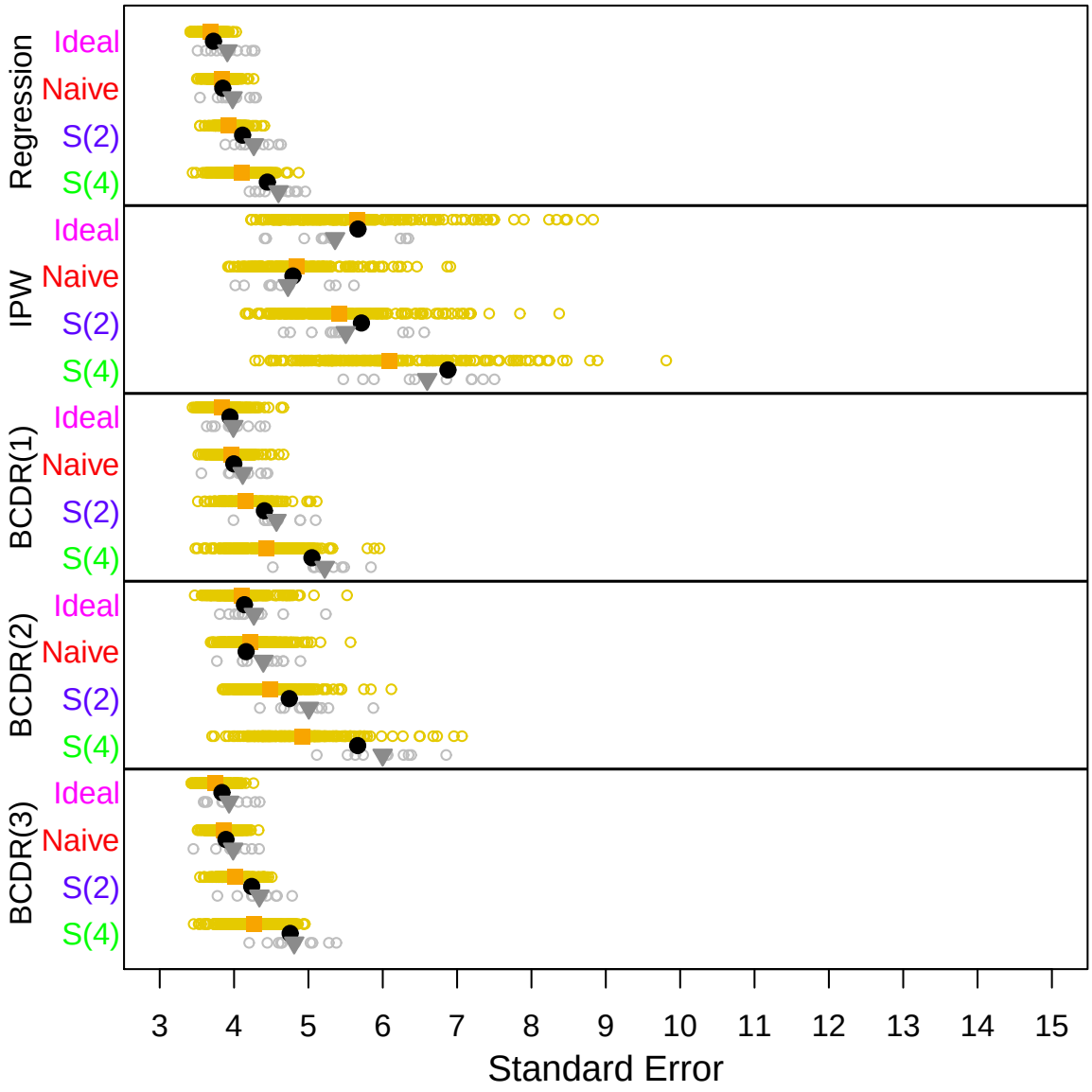

Figure 16: Summary of estimated standard errors for different estimators of the mean of the linear outcome, with $n = 1000$ and reliability of $W_{(1)}$ equal to 0.70. Analogous to Figure 5.

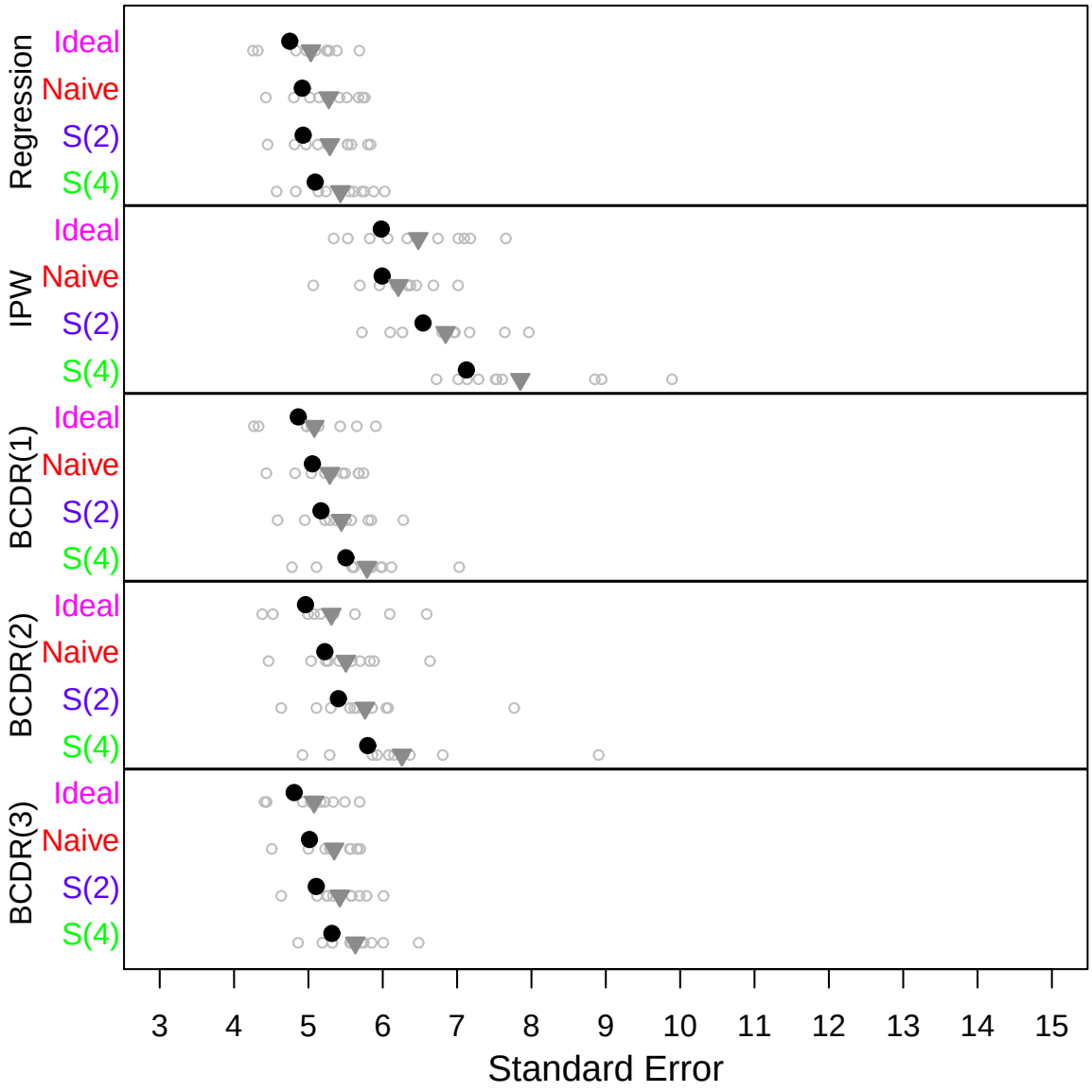

Figure 17: Summary of estimated standard errors for different estimators of the mean of the tobit outcome, with $n = 500$ and reliability of $W_{(1)}$ equal to 0.85. Analogous to Figure 5.

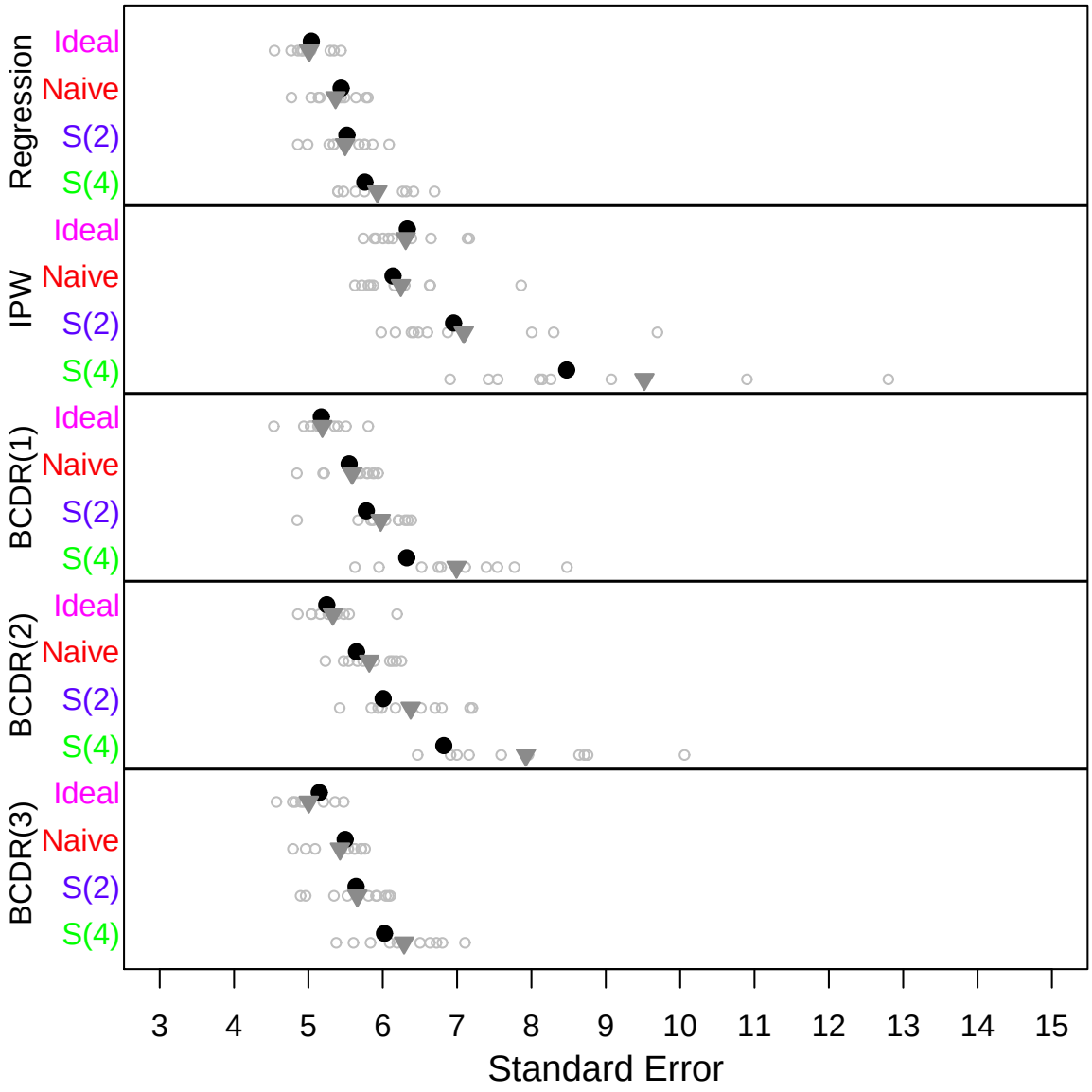

Figure 18: Summary of estimated standard errors for different estimators of the mean of the tobit outcome, with $n = 500$ and reliability of $W_{(1)}$ equal to 0.70. Analogous to Figure 5.

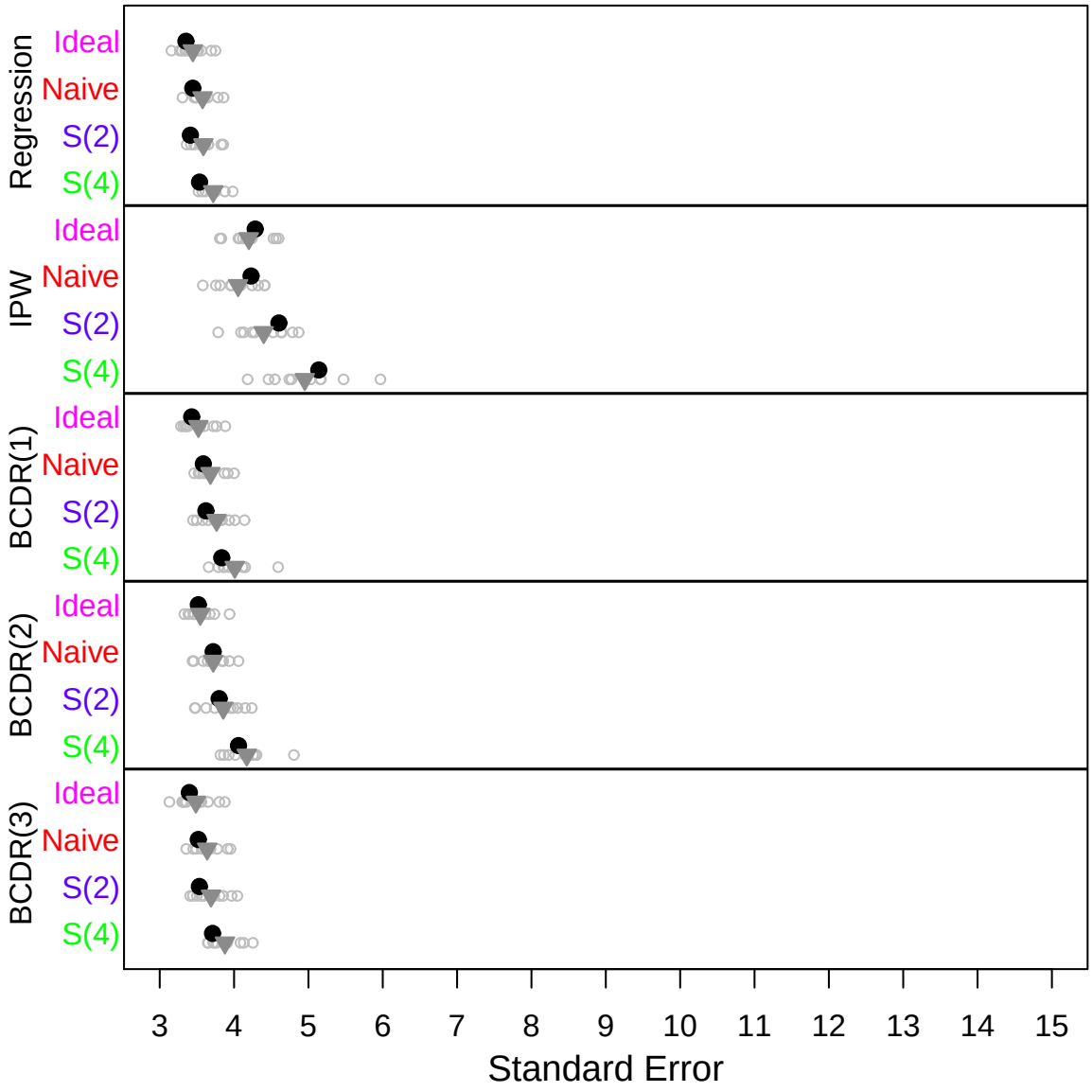

Figure 19: Summary of estimated standard errors for different estimators of the mean of the tobit outcome, with $n = 1000$ and reliability of $W_{(1)}$ equal to 0.85. Analogous to Figure 5.

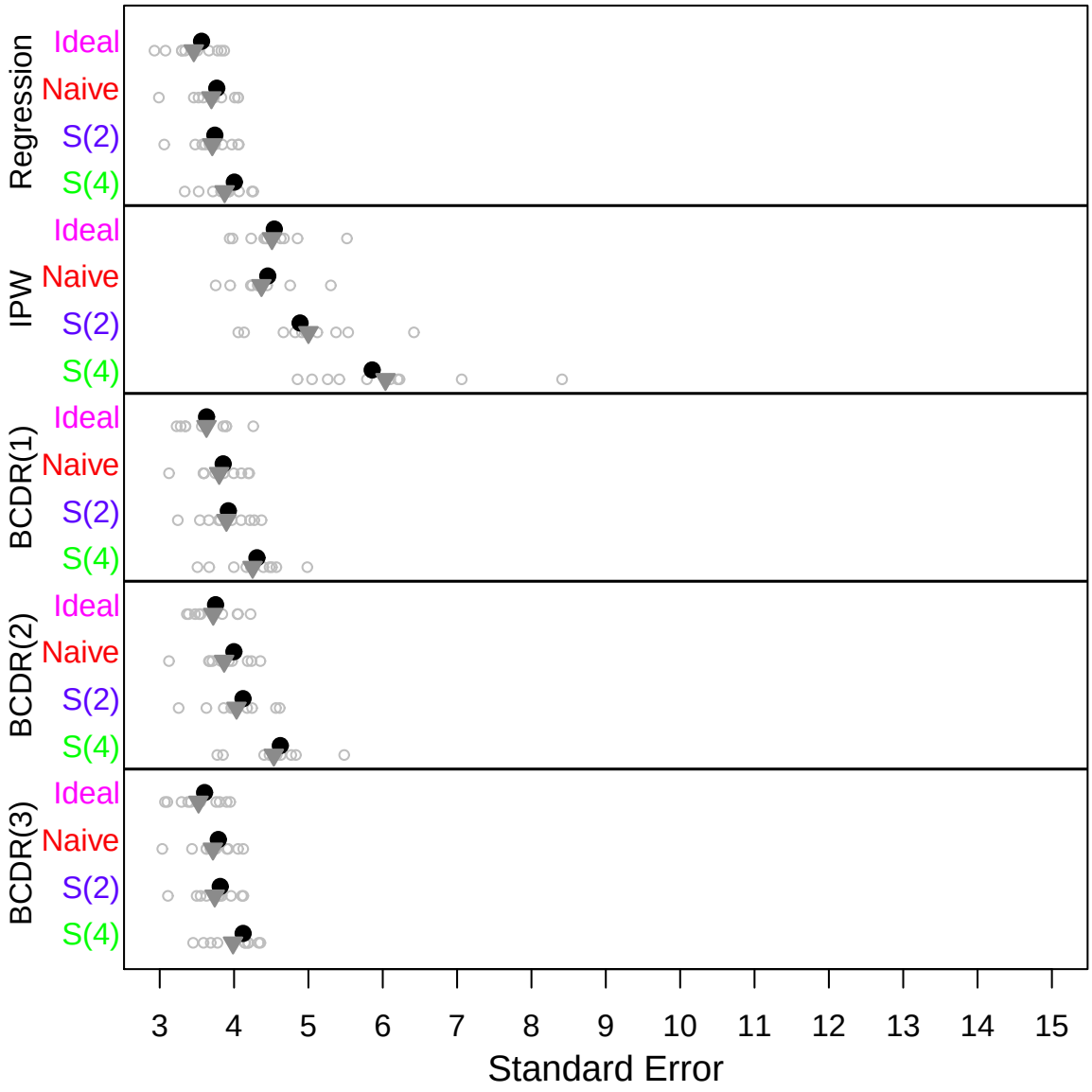

Figure 20: Summary of estimated standard errors for different estimators of the mean of the tobit outcome, with $n = 1000$ and reliability of $W_{(1)}$ equal to 0.70. Analogous to Figure 5.

References

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

Battauz, M. and Bellio, R. (2011). Structural modeling of measurement error in generalized linear models with Rasch measures as covariates. *Psychometrika*, 76(1):40–56.

Braun, H. (2005). Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models. Princeton, NJ: Educational Testing Service, Policy Information Center.

Carroll, R., Küchenhoff, H., Lombard, F., and Stefanski, L. (1996). Asymptotics for the simex estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, 91(433):242–250.

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective (2nd ed.)*. Chapman and Hall, London.

Carroll, R. and Wang, Y. (2008). Nonparametric variance estimation in the analysis of microarray data: A measurement error approach. *Biometrika*, 95(2):437–449.

Cook, J. and Stefanski, L. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328.

Crocker, L. and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Holt, Rinehart and Winston, Orlando, FL.

D'Agostino, Jr., R. and Rubin, D. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, 95(451):749–759.

Delaigle, A. and Meister, A. (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *Journal of the American Statistical Association*, 102(480):1416–1426.

Devanarayan, V. and Stefanski, L. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics & Probability Letters*, 59:219–225.

Freier, M., Bell, R., and Ellickson, P. (1991). *Do Teens Tell the Truth?* RAND Corporation.

Fuller, W. (2006). *Measurement Error Models (2nd ed)*. John Wiley & Sons, New York.

Fung, K. and Krewski, D. (1999). Evaluation of regression calibration and SIMEX methods in logistic regression when one of the predictors is subject to additive measurement error. *Journal of Epidemiology and Biostatistics*, 4(2):65–74.

Haberman, S. (1984). Adjustment by minimum discriminant information. *The Annals of Statistics*, 12:971–988.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20:25–46.

Harris, D. (2011). *Value-Added Measures in Education: What Every Educator Needs to Know*. Harvard Education Press, Cambridge, MA.

Hirano, K. and Imbens, G. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology*, 2(3-4):259–278.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Imbens, G. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Kane, T. and Staiger, D. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, 16(4):91–114.

Kang, J. and Schafer, J. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Kim, J. (2010). Calibration estimation using exponential tilting in sample surveys. *Survey Methodology*, 36(2):145–155.

Kleiber, C. and Zeileis, A. (2008). *Applied Econometrics with R*. Springer-Verlag, New York. ISBN 978-0-387-77316-2.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Lockwood, J. and McCaffrey, D. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics*, 39(1):22–52.

Lockwood, J. and McCaffrey, D. (2015a). Matching estimators for causal inference with error-prone covariates. Conditional acceptance, *Journal of the American Statistical Association*.

Lockwood, J. and McCaffrey, D. (2015b). Simulation-extrapolation with latent heteroskedastic error variance. Unpublished manuscript.

Lord, F. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Lunceford, J. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Martino, S., McCaffrey, D., Klein, D., and Ellickson, P. (2009). Recanting of life-time inhalant use: How big a problem and what to make of it. *Addiction*, 104(8):1373–1381.

McCaffrey, D., Liccardo Pacula, R., Han, B., and Ellickson, P. (2010). Marijuana use and high school dropout: the influence of unobservables. *Health Economics*, 19(11):1281–1299.

McCaffrey, D. and Lockwood, J. (2014). Doubly robust estimation with error-prone covariates. Presented at Atlantic Causal Inference Conference, Providence, RI, May 2014.

McCaffrey, D., Lockwood, J., Koretz, D., Louis, T., and Hamilton, L. (2004a). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1):67–101.

McCaffrey, D., Lockwood, J., and Setodji, C. (2013). Inverse probability weighting with error-prone covariates. *Biometrika*, 100(3):671–680.

McCaffrey, D., Ridgeway, G., and Morral, A. (2004b). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403–425.

Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion). *JASA*, 78(381):47–55.

Novick, S. and Stefanski, L. (2002). Corrected score estimation via complex variable simulation extrapolation. *Journal of the American Statistical Association*, 97(458):472–481.

Pane, J., Griffin, B., McCaffrey, D., and Karam, R. (2014). Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis*, 36(2):127–144.

Pearl, J. (2010). On measurement bias in causal inference. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 425–432, Corvallis, Oregon. AUAI Press.

Pirracchio, R., Petersen, M., and van der Laan, M. (2015). Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology*, 181(2):108–119.

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raykov, T. (2012). Propensity score analysis with fallible covariates : A note on a latent variable modeling approach. *Educational and Psychological Measurement*, 72(5):715–733.

Regier, M., Moodie, E., and Platt, R. (2014). The effect of error-in-confounders on the estimation of the causal parameter when using marginal structural models and inverse probability-of-treatment weights: A simulation study. *The International Journal of Biostatistics*, 10(1):1–15.

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

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

Rosenbaum, P. and Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. (1980). Discussion of "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3-4):169–188.

Scharfstein, D., Rotnitzky, A., and Robins, J. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94:1096–1120.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.

Stefanski, L. (1989). Unbiased estimation of a nonlinear function a normal mean with application to measurement error models. *Communications in Statistics-Theory and Methods*, 18(12):4335–4358.

Stefanski, L. and Boos, D. (2002). The calculus of M-estimation. *The American Statistician*, 56(1):29–38.

Stefanski, L. and Cook, J. (1995). Simulation-extrapolation: The measurement error jackknife. *Journal of the American Statistical Association*, 90(432):1247–1256.

Steiner, P., Cook, T., and Shadish, W. (2011). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Journal of Educational and Behavioral Statistics*, 36(2):213–236.

Stuart, E. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21.

Sullivan, D. (2001). A note on the estimation of linear regression models with heteroskedastic measurement errors. Federal Reserve Bank of Chicago working paper #2001-23.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer, New York.

Valeri, L., Lin, X., and VanderWeele, T. (2014). Mediation analysis when a continuous mediator is measured with error and the outcome follows a generalized linear model. *Statistics in medicine*, 33(28):4875–4890.

van der Linden, W. and Hambleton, R. (1997). *Handbook of Modern Item Response Theory*. Springer-Verlag, New York, NY.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, pages 595–601.

Wang, Y., Ma, Y., and Carroll, R. (2009). Variance estimation in the analysis of microarray data. *Journal of the Royal Statistical Society: Series B*, 71(2):425–445.

Yi, G., Ma, Y., and Carroll, R. (2012). A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika*, 99(1):151–165.