Abstract Title Page

Title: Propensity Score Weighting with Error-prone Covariates

Author(s): Daniel F. McCaffrey, J.R. Lockwood, Claude M. Setodji

Abstract Body

Limit 4 pages single spaced.

Background / Context:

Inverse probability weighting (IPW) estimates are widely used in applications where data are missing due to nonresponse or censoring (Kang and Schafer, 2007; Lunceford and Davidian, 2004; Scharfstein et al., 1999; Robins and Rotnitzky, 1995; Robins et al., 1995) or in observational studies of causal effects where the counterfactuals cannot be observed (Schafer and Kang, 2008; Bang and Robins, 2005; McCaffrey et al., 2004; Robins et al., 2000). This extensive literature has shown the estimators to be consistent and asymptotically normal under very general conditions, and combining IPW with modeling for the mean function yields "doubly robust" estimates which are consistent and asymptotically normal (Kang and Schafer, 2007; Bang and Robins, 2005; Lunceford and Davidian, 2004; van der Laan and Robins, 2003; Scharfstein et al., 1999; Robins and Rotnitzky, 1995), (Robins et al.,1995). Recent studies have considered estimation of the response or treatment assignment functions (Hill, 2010; Harder et al., 2011; Lee et al., 2009; McCaffrey et al., 2004; Hirano et al., 2003) and have shown nonparametric and boosting type estimators work well in simulations and applications.

The consistency and asymptotically normality of IPW estimates generally are guaranteed to hold only with data where response or treatment assignment is independent of the outcomes of interest conditional on a set of observed covariates. The extensive IPW literature only considers settings where the covariates are free of measurement error. However, covariates measured with error are common in applications that might use IPW estimation. For instance, psychological scales from surveys are imperfect measures of the underlying constructs and it is likely that outcomes of interest and treatment assignment or response to a survey depend on the individual's underlying psychological state not on the error-prone measure. Similarly, achievement tests for school students can have very large errors for some students and again it is clear that future achievement depends on a students' true level on achievement and not on the error-prone test scores. Ignoring the measurement error in the covariates can result in bias in IPW estimates (Steiner et al., 2011).

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

The purposes of this study are to:

1. Develop an analytic form for a weighting function that can be used to estimate weighted means that consistently estimate the population mean from incomplete data when covariates are measured with error;
2. Develop a consistent estimator of the weighting function and the population mean when covariates are measured with error;
3. Evaluate the small sample properties of the estimator through a simulation study.

Significance / Novelty of study:

Description of what is missing in previous work and the contribution the study makes.

Measurement error is extremely common in observational evaluations of educational interventions. Many observational studies involve groups which did or did not receive the intervention and differ on their pre-existing levels of achievement and risk factors for low achievement. Such differences between groups typically arise because of the large differences in background characteristics among students from different schools or different classes within schools. Often standardize achievement test scores are used to account for these differences in estimation of the intervention effects. However, achievement tests often have very large measurement error for students with very high or low achievement and large errors even for students just moderately above or below the mean. For instance, in a large urban school district the average squared standard error of measure on state English language arts tests for students in grades 6, 7, and 8 equals 20 percent of the estimated variance of an error-free achievement measure. For reading the average squared standard error of measure is 26 percent of the estimated variance of an error-free achievement measure and the percentages are 15, 11, and 12 for mathematics, science and social studies. For students with very high levels of achievement the measurement error variance can exceed 50 percent of the variance in the error-free achievement measure.

IPW estimators allow for estimation of group means without parametric assumptions about the mean function. In addition, tuning of the weighting function can be conducted without use of the outcome variables, so its effect on the final estimate of the intervention impact estimate cannot influence the selection of the function. In addition the estimators can be robust to errors in either the weighting or mean function, if both are estimated. Thus, such IPW weights are very valuable for observational studies and quickly gaining in usage.

However, weights based on error-prone covariates can result in biased estimates of the mean. Given the large measurement error in achievement tests the bias in estimated treatment effects potentially could be large for education evaluation that use achievement tests to adjust for pre-existing group differences.

Currently there are no weighted estimators or other estimators that make use of the propensity score in the presence of measurement error in the covariates. The literature in this area includes several simulation studies that demonstrate the potential for error when weighting or matching with propensity scores fit to error-prone measures, but they do not provide methods for consistent estimation. Our method fills this gap. We provide an analytic form for a weighting function that can provide consistent estimates of a population means from sample with missing data and consistent estimates of the treatment effects under standard assumptions. We also develop a method for estimating the weighting function and population means or treatment effects.

Statistical, Measurement, or Econometric Model:

Let Y_i, i=1,…, n, be the outcome of of primary interest obtained from a sample of units from a population, where interest is in the population mean of Y, μ. IPW estimation commonly is applied to two scenarios where the outcomes are observed for only portion of the sample. The first scenario is missing data due survey nonresponse, loss-to-follow-up, or censoring in which sampled units fail to provide requested data. The second scenario involves the estimation of the causal effect of a treatment or treatments in which only one of the possible potential outcomes for each study unit is observed, the outcome corresponding to the unit's assigned treatment, and all other potential outcomes are unobserved. Let R_i be a "response" indicator, i.e., $R_i = 1$ if Y_i is observed and $R_i = 0$ is Y_i is unobserved or missing. For observational studies, let T_i be the treatment indicator with $T_i = 1$ if study unit *i* received the treatment and $T_i = 0$ if the unit received the control condition. We set $R_i = T_i$ when estimating the mean of the potential outcomes for treatment, and $R_i = 1 - T_\text{i}$ when estimating the mean of the potential outcomes for control. We will use the generic term response indicator but the results apply to both nonresponse and observational studies.

For each unit, there exists a error-free covariate U_i which is unobserved and observed covariates $X_i = U_i + \xi_i$ and Z_i, where the measurement error, ξ_i, has a known distribution and is independent of Y_i, R_i and Z_i, which is observed without error.

Assumption 1. $0 < P(R_i = 1 \mid U_i, Z_i) < 1$ for all sampled units.
Assumption 2. Y_i is independent of R_i conditional on U_i and Z_i.

The assumptions are similar to strong ignorability (Rosenbam & Rubin, 1983), both require Assumption 1. However, in the context of causal effect estimation, strong ignorability requires the conditional joint distribution of the potential outcomes, (Y_{i0}, Y_{i1}), to be independent of treatment. We only require each potential outcome be marginally independent of treatment assignment conditional on U_i and Z_i or *weak unconfoundness* (Imbens, 2000). More importantly, independence is conditional on Z_i and the error-free variable U_i not the observed error-prone covariate X_i.

Usefulness / Applicability of Method:

Theorem 1. Given Assumptions 1 and 2 and a weighting function $W(x,z)$ that satisfies

1. $\mathrm{E}(W(X, z)| U) = P(R = 1 \mid U, Z=z)^{-1}$, for any z in the support of Z, and
2. $R\,Y\,W(X)$ have finite first moment,

then

$$\hat{\mu} = \frac{\sum_{i=1}^{n} R_i Y_i W(X_i, Z_i)}{\sum_{i=1}^{n} R_i W(X_i, Z_i)}$$

is a consistent estimator of μ.

Remark 1. Theorem 1 naturally extends to settings with multiple error-prone and error-free covariates.

Corrollary 1. Let $\mu_t = \mathrm{E}(Y_t)$, t=0, 1, where expectation is for the entire population, and $\delta = \mu_1 - \mu_0$ equal the average treatment effect. Let $W_1(x, z)$ satisfy Conditions 1 and 2 of Theorem 1 with $R = T$ and Y_{i1} and $W_0(x, z)$ satisfy the conditions with $R = 1 - T$ and $Y_{i0.}$ Let $0 < P(T_i = 1) < 1$ for all sampled units and Y_{it}, t=0, 1, be independent of T_i conditional on U_i and Z_i then

$$\hat{\delta} = \frac{\sum_{i=1}^{n} T_i Y_{i1} W_1(X_i, Z_i)}{\sum_{i=1}^{n} T_i W_1(X_i, Z_i)} - \frac{\sum_{i=1}^{n} (1 - T_i) Y_{i0} W_0(X_i, Z_i)}{\sum_{i=1}^{n} (1 - T_i) W_0(X_i, Z_i)}$$

is a consistent estimate of δ.

Research Design:

In practice the propensity score, $p(u, z) = P(R = 1 \mid U=u\ , Z=z)$, and must be estimated using the observed Z_i, and X_i variables rather than U_i and methods for fitting nonlinear models with measurement error (Carroll, Ruppert, Stefanski, and Craniceanu, 2006). Also, the weighting function must be calculated. If $p(u, z)^{-1}$ has a Fourier transform then there exist analytic solutions for $W(x,z)$ using the inverse Fourier transform of $p(u, z)^{-1}$ and the characteristic function of the known distribution of the measurement error, ξ_i. Alternatively, $W(x, z)$ can be approximated by an additive function in x and z and the coefficients of this function can be estimated using simulated data. For instance, the following algorithm can be used to approximate $W(x, z)$. For each unit in the sample with $R_i = 1$,

1. Assume $W(x, z) = \sum_{k=1}^{K} \beta_{zk} \varphi_k(x)$ for a finite K and known basis functions such as polynomials (i.e., $\varphi_k(x_{jb}) = x^k$) or B-splines
2. Estimate the $p(u, z)$ by $\hat{p}(u, z)$
3. Select a sample of u_j, j = 1,..,J, values from the range of U
4. Generate a simulated sample of ξ_b, b = 1,.., B from the distribution of ξ_i
5. For each u_j,
   a. Generate $x_{jb} = u_j + \xi_b$, b = 1,.., B
   b. Calculate $\varphi_k(x_{jb})$ for every simulated x_{jb} and all basis functions
   c. Calculate the$\bar{\varphi}_{jk} = B^{-1} \sum_{b=1}^{B} \varphi_k(x_{jb})$
   d. Calculate $\hat{p}(u_j, z_i)$
6. Estimate β_{zk} through a linear regression of $\bar{\varphi}_{jk}$ on $\hat{p}(u_j, z_i)$

We test our estimation procedures using simulated data for an observational evaluation of an intervention designed to match distributions of achievement data of middle school students in a large urban school system. We explore alternative functional forms for the propensity score (additive linear and nonlinear logistic models in x and z), different assumptions about the distribution of U (known or unkown), different estimators for $p(u, z)$, and different values for B and J. Methods are evaluated by the bias and mean-square error of the resulting treatment effect estimates.

Conclusions:

Inverse probability weighting estimates are valuable and gaining wide usage. They can be biased by measurement error in the covariates used to adjust for differences among cases with observed and unobserved outcomes and no current methods exist for consistent estimation in these cases. We provide a consistent estimator and methods for implementing it in applications. We evaluate the estimator via simulation to discuss its feasibility and small sample properties.

Appendices

Appendix A. References

References are to be in APA version 6 format.

Bang, H., & Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics, 61*, 962–972. doi:10.1111/j-1541.0420-2005.00377.x

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd Edition). Boca Raton, FL: Chapman & Hall/CRC Press.

Harder, V., Stuart, E., & Anthony, J. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*, 234–249. doi:10.1037/a0019623

Hill, J. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*, 217–240. doi:10.1198/jcgs.2010.08162

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*, 1161–1189. doi:10.1111/1468-0262.00442

Imbens, G. W. (2000). The role of the propensity score in estimating dose- response functions. *Biometrika*, *87*, 706–710. doi:10.1093/biomet/87.3.706

Kang, J., & Schafer, J. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*, 523 – 539. doi:10.1214/07-STS227

Lee, B., Lessler, J., & Stuart, E. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*, 337–346. doi:10.1002/sim.3782

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*, 2937–2960. doi:10.1002/sim.1903

McCaffrey, D., Ridgeway, G., & Morral, A. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*, 403–425. doi:10.1037/1082-989X.9.4.403

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*, 550–560.

Robins, J., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*, 122–129.

Robins, J., Rotnitzky, A., & Zhao, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, *90*, 106–121.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55. doi:10.1093/biomet/70.1.41

Schafer, J., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279–313. doi:10.1037/a0014268

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (C/R: P1121-1146). *Journal of the American Statistical Association*, *94*, 1096–1120.

Steiner, P., Cook, T., & 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*, 213–236. doi:10.3102/1076998610375835

van der Laan, M., & Robins, J. (2003). *Unified methods for censored longitudinal data and causality*. New York: Springer.