Biometrika (2010), **97**, 4, *pp*. 997–1001
doi: 10.1093/biomet/asq049
Advance Access publication 31 July 2010

A note on overadjustment in inverse probability weighted estimation

BY ANDREA ROTNITZKY

Di Tella University, Sáenz Valiente 1010, Buenos Aires, Argentina
andrea@hsph.harvard.edu

LINGLING LI

Department of Population Medicine, Harvard Medical School, Harvard Pilgrim Health Care Institute, Boston, Massachusetts 02115, U.S.A.
lingling_li@post.harvard.edu

AND XIAOCHUN LI

Division of Biostatistics, Indiana University School of Medicine, Regenstrief Institute, Indianapolis, Indiana 46202, U.S.A.
xiaochun@iupui.edu

SUMMARY

Standardized means, commonly used in observational studies in epidemiology to adjust for potential confounders, are equal to inverse probability weighted means with inverse weights equal to the empirical propensity scores. More refined standardization corresponds with empirical propensity scores computed under more flexible models. Unnecessary standardization induces efficiency loss. However, according to the theory of inverse probability weighted estimation, propensity scores estimated under more flexible models induce improvement in the precision of inverse probability weighted means. This apparent contradiction is clarified by explicitly stating the assumptions under which the improvement in precision is attained.

Some key words: Causal inference; Propensity score; Standardized mean.

1. INTRODUCTION

Often, epidemiological studies aim to evaluate the causal effect of a discrete exposure on an outcome. In observational studies systematic bias due to confounding is a serious concern. For this reason, investigators routinely collect and adjust for a large number of confounding factors in data analyses. A common analytic strategy is to categorize the confounders and then to compare the exposure group-specific standardized means. These are exposure group-specific weighted means of the outcome across levels of the categorized confounders with weights equal to the empirical probabilities of the categorized confounders in the entire sample. It is well known that overcategorization, i.e. unnecessary categorization, may induce efficiency losses. This issue is essentially the same as the well-understood increase in variance induced by adding in a linear regression model covariates that have no partial correlation with the outcome (Cochran, 1968). It has been studied in a number of nonlinear regression settings, e.g. Mantel & Haenszel (1959), Breslow (1982), Gail (1988), Robinson & Jewell (1991), Neuhauhaser & Becher (1997) and De Stavola & Cox (2008), and has been empirically analyzed for standardized means in Brookhart et al. (2006).

The issue, however, appears to contradict well-known facts in the theory of inverse probability weighted estimation. Specifically, a standardized mean is equal to a so-called inverse probability of treatment

weighted mean. More precisely, it is equal to a group-specific mean of the outcome weighted by the inverse of the empirical propensity score. An empirical propensity score is the maximum likelihood estimate of the true propensity score, i.e. of the probability of being in the exposure group given the confounders, under a saturated model for the probability of exposure given the categorized confounder. The apparent contradiction is that more refined categorization corresponds to more flexible models for the propensity score, and according to the theory of inverse probability estimation, the use of more flexible propensity score models induces an improvement in the precision of inverse probability means, and not a decrease in precision as regression theory indicates.

The purpose of this note is to clarify this apparent contradiction showing that indeed, efficiency losses induced by unnecessarily refined categorizations do not contradict, and indeed are a consequence of, the theory of inverse probability estimation.

2. THE APPARENT CONTRADICTION

Consider a cohort study in which a discrete exposure variable A, an outcome Y and a vector of pre-exposure covariates X are measured for each of n subjects drawn at random from a study population. Although the typical goal of such a study is the evaluation of the exposure effect on the outcome, i.e. a comparison across exposure levels, the issues in this note are best understood by considering inference about the outcome mean at one specific exposure level. Thus, we will assume that A is binary and that the goal is to estimate the outcome mean at exposure level $A = 1$. Consider a categorization of X into J strata and let L denote the polytomous variable that records the stratum, a subject with covariates X belongs to. The standardized mean at exposure level $A = 1$ and with categorized variable L is

$$\hat{\mu} \equiv E_n\{E_n(Y \mid A = 1, L)\}, \tag{1}$$

where throughout for any U and V,

$$E_n(U) \equiv n^{-1} \sum_{i=1}^{n} U_i, \quad E_n(U \mid A = 1, V) \equiv \left(\sum_{i:A_i=1, V_i=V} U_i \right) \Big/ \left(\sum_{i:A_i=1, V_i=V}^{n} 1 \right).$$

For standardized means to be informative about the causal effects certain assumptions need to hold. The issue is best articulated within the potential outcomes framework. Let Y_a be the subject's potential outcome if, perhaps contrary to fact, he is exposed to $A = a$. Contrasts comparing $E(Y_1)$ and $E(Y_0)$ quantify the causal effect of exposure. The standardized mean $\hat{\mu}$ is consistent for $\mu \equiv E(Y_1)$ under the following assumptions.

Assumption 1. Consistency: $Y = Y_A$.

Assumption 2. Positivity: $\text{pr}(A = 1 \mid L) > 0$.

Assumption 3. No unmeasured confounders: Y_1 and A are conditionally independent given L, because in such a case

$$\mu = E\{E(Y \mid A = 1, L)\}. \tag{2}$$

The apparent contradiction discussed in this note refers to the asymptotic behaviour of $\hat{\mu}$ under two categorizations, one more refined than the other. The essence of the matter is best understood by considering the extreme case contrasting the asymptotic behaviour of the adjusted average $\hat{\mu}$ with that of the crude unadjusted average,

$$\widetilde{\mu} \equiv E_n(Y \mid A = 1).$$

Our discussion focusses on this comparison. The well-known risk of bias induced by underadjustment, i.e. by failure to adjust for an important confounder, is vividly unmasked in this extreme case: $\widetilde{\mu}$ does not generally converge in probability to $E(Y_1)$. Formally, $\widetilde{\mu}$ converges to $E(Y_1 \mid A = 1)$ which is not generally equal to $E(Y_1)$ because Y_1 and A may share the common determinant L. Consistency of $\widetilde{\mu}$ requires that, in addition to Assumptions 1–3, at least one of the following two independencies hold.

Assumption 4. The variables Y and L are conditionally independent given $A = 1$.

Assumption 5. The variables A and L are independent.

In the Appendix we show that $\hat{\mu}$ solves the inverse probability weighted estimating equation

$$E_n\left\{\frac{A}{E_n(A \mid L)}(Y-\mu)\right\} = 0, \tag{3}$$

whereas $\widetilde{\mu}$ solves the inverse probability weighted estimating equation

$$E_n\left\{\frac{A}{E_n(A)}(Y-\mu)\right\} = 0 \tag{4}$$

whence the apparent contradiction emerges. Specifically, both $E_n(A \mid L)$ and $E_n(a)$ can be regarded as efficient estimators of the propensity score $\pi(l) \equiv E(A \mid L)$, the former under a saturated model on L and the latter under the smaller model that assumes independence of A and L. According to the theory of inverse probability estimation, inclusion of covariate L in an efficiently estimated model for the propensity score should not be detrimental to the efficiency with which $E(Y_1)$ is estimated even if the covariate is not needed for bias correction. This appears to contradict the fact that under Assumptions 1–3, $\widetilde{\mu}$ is more efficient than $\hat{\mu}$ when Assumption 4 holds and Assumption 5 fails.

3. Explaining the apparent contradiction

The apparent contradiction arises because of the vagueness of the statement about the efficiency gains induced by including L in the propensity score estimators, which does not explicitly mention the assumptions required for its validity. To explain the contradiction, let $\mathcal{A}$ denote the model defined by Assumptions 1–3, let $\mathcal{B}$ denote the model defined by Assumptions 1–4 and let $\mathcal{C}$ denote Assumptions 1–3 and 5.

Both $\hat{\mu}$ and $\widetilde{\mu}$ are consistent for $E(Y_1)$ under model $\mathcal{B}$ or $\mathcal{C}$ but only $\hat{\mu}$ is consistent for $E(Y_1)$ under model $\mathcal{A}$.

The estimator $\hat{\mu}$ is asymptotically efficient under model $\mathcal{A}$ and under model $\mathcal{C}$ but $\widetilde{\mu}$ is asymptotically efficient under model $\mathcal{B}$. These efficiency results are best understood by examining the likelihood

$$\mathcal{L}_n(f_{A,Y,L}) = \mathcal{L}_{1,n}(f_L, f_{Y|A,L})\mathcal{L}_{2,n}(f_{A|L}), \tag{5}$$

where

$$\mathcal{L}_{1,n}(f_L, f_{Y|A,L}) = \prod_{i=1}^{n} f_L(L_i) f_{Y|A,L}(Y_i \mid A_i, L_i), \quad \mathcal{L}_{2,n}(f_{A|L}) = \prod_{i=1}^{n} f_{A|L}(A_i \mid L_i).$$

Model $\mathcal{A}$ imposes restrictions on the law of (Y_1, L, A) but not on the distribution $f_{A,Y,L}$ of the observed data (Y, L, A) (Gill et al., 1997) and hence is a nonparametric model for the observables. Because the estimator $\hat{\mu}$ is the plug-in estimator of $\mu = E\{E(Y \mid A = 1, L)\}$, it is the maximum likelihood estimator of μ under the nonparametric model $\mathcal{A}$.

Model $\mathcal{C}$ restricts the law $f_{A|L}$ entering the second term on the right-hand side of (5) since Assumption 5 postulates that $f_{A|L} = f_A$. Because by (2), μ depends only on the components of the law entering in the $\mathcal{L}_{1,n}$-part of the likelihood (5), the maximum likelihood estimators of μ under models $\mathcal{A}$ and $\mathcal{C}$ must agree. Thus, $\hat{\mu}$ is the maximum likelihood estimator of μ under model $\mathcal{C}$ and consequently asymptotically efficient, i.e. avar$(\hat{\mu})$ is equal to the semiparametric variance bound for μ under the model. We let avar$(\cdot)$ denote the variance of the limiting distribution, hereafter.

Model $\mathcal{B}$ imposes the restriction $f_{Y|A=1,L} = f_{Y|A=1}$ and hence it restricts the law $f_{Y|A,L}$ in $\mathcal{L}_{1,n}$. The estimator $\hat{\mu}$ is not the maximum likelihood estimator under model $\mathcal{B}$ because it does not exploit this restriction. In fact, under model $\mathcal{B}$, $\widetilde{\mu}$ is asymptotically efficient. Furthermore, $\widetilde{\mu}$ is asymptotically strictly more efficient than $\hat{\mu}$ unless Assumption 5 also holds. Proof of these results can be found in the online Supplementary Material. We are now ready to explain the contradiction.

Given an arbitrary function $d(l)$ and any $\pi(l)$, let $\hat{\mu}_d(\pi)$ denote the solution to

$$E_n\left\{\frac{A}{\pi(L)}d(L)(Y-\mu)\right\}=0. \tag{6}$$

The following Lemma, a corollary of the theory laid out in Robins et al. (1994), states the precise result of the theory of inverse probability weighted estimation that the gain in efficiency of $\widetilde{\mu}$ over $\hat{\mu}$ appears to contradict.

LEMMA 1. *Given one of the models $\mathcal{A}$, $\mathcal{B}$ or $\mathcal{C}$ for the observables, let $\hat{\pi}(l)$ and $\widetilde{\pi}(l)$ be the maximum likelihood estimators of $f_{A|L}(1 \mid l)$ under two nested models for $f_{A|L}$ that are correctly specified under the assumptions of the given model. Then $\sqrt{n}\{\hat{\mu}_d(\hat{\pi})-\mu\}$ and $\sqrt{n}\{\hat{\mu}_d(\widetilde{\pi})-\mu\}$ converge to mean zero normal distributions. If $\hat{\pi}(l)$ is the estimator of $f_{A|L}(1 \mid l)$ under the larger model, then*

$$\text{avar}\{\hat{\mu}_d(\hat{\pi})\} \leqslant \text{avar}\{\hat{\mu}_d(\widetilde{\pi})\}.$$

Observe that because $\hat{\mu}$ solves (3) and $\widetilde{\mu}$ solves (4) we can write $\hat{\mu}=\hat{\mu}_{d_1}(\hat{\pi})$ and $\widetilde{\mu}=\hat{\mu}_{d_1}(\widetilde{\pi})$ with $d_1(l)=1$, $\hat{\pi}(l)=E_n(A \mid L=l)$ and $\widetilde{\pi}(l)=E_n(A)$. The improved efficiency of $\widetilde{\mu}$ over $\hat{\mu}$, i.e. the fact that generally avar$(\widetilde{\mu})$ is strictly smaller than avar$(\hat{\mu})$, under model $\mathcal{B}$ does not contradict Lemma 1 because $\widetilde{\pi}(l)$ does not meet its premise. Specifically, Lemma 1 makes the premise that $\widetilde{\pi}(l)$ is computed under a model for $f_{A|L}$ that is correctly specified under the given model, in the case of our concern, model $\mathcal{B}$. However, $\widetilde{\pi}(l)=E_n(A)$ is the fitted value under a model for $f_{A|L}$ that assumes that A and L are independent, an assumption not made by model $\mathcal{B}$.

The efficiency gains conferred by $\widetilde{\mu}$ over $\hat{\mu}$ under model $\mathcal{B}$ can be deduced from the general theory of efficient inverse probability estimation in semiparametric models for missing data (Robins et al., proposition 8.1, 1994). In the Supplementary Material we apply this theory to show that: (a) $\widetilde{\mu}$ is asymptotically equivalent to $\hat{\mu}_{d_2}(\hat{\pi})$ with $d_2(l)=E(A \mid L=l)$ and (b) $\hat{\mu}_{d_2}(\hat{\pi})$, and therefore $\widetilde{\mu}$, is semiparametric efficient under $\mathcal{B}$.

In conclusion, the fallacy arises because the claim about efficiency gains assumes an explicit model for the law of (A, L, Y) and it requires that both propensity score models be correct under the given model. However, $E_n(A)$ is the efficient propensity score estimator under a model not implied by model $\mathcal{B}$, so the efficiency claim does not apply.

4. CONCLUDING REMARKS

Our analysis extends to inference in marginal structural mean models for the effect of a, possibly polytomous, exposure A given, a possibly strict, subset Z of the confounders L. These models assume that $E(Y_a \mid Z)=m(a, Z; \beta)$, where $m(\cdot)$ is known and β unknown. Estimators of β are obtained by solving (6) with $A/\pi(l)$ replaced by an estimator of $1/f_{A|L}(A \mid L)$, μ replaced by $m(A, Z; \beta)$ and with $d(l)$ of the dimension of β. When Assumption 5 holds, using $1/\widetilde{f}_A(A)$ where $\widetilde{f}_A(A)=E_n\{I_a(A)\}$ and $I_a(A)$ is the indicator that $A=a$ yields consistent and asymptotically normal estimators of β that are generally more efficient than those obtained using $1/\hat{f}_{A|L}(A \mid L)$ where $\hat{f}_{A|L}(A \mid l)=E_n\{I_a(A) \mid L=l\}$. Once again, this raises an apparent contradiction with inverse probability weighted estimation which can be explained as in § 3.

ACKNOWLEDGEMENT

Andrea Rotnitzky was funded by a grant from the National Institutes of Health, U.S.A. The authors wish to thank two referees and the associate editors for helpful comments.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *Biometrika* online.

Appendix

For any given law $f(l, a, y)$, define the new law $f^*(l, a, y) = f(l)I_1(a)f(y \mid a, l)$. Then $E\{E(Y \mid A = 1, L)\} = E^*(Y)$ where $E(\cdot)$ and $E^*(\cdot)$ denote expectations under f and f^* respectively. But, $f^*(l, a, y)/f(l, a, y) = I_1(a)/f(a \mid l)$, so $E^*(Y) = E\{I_1(a)Y/f(A \mid L)\}$ thus proving that $E\{E(Y \mid A = 1, L)\} = E\{AY/f(1 \mid L)\}$ for any f and A binary. That $\hat{\mu}$ solving (3) also admits the representation (1) follows by applying this result when f is the empirical law.

References

Brookhart, M. A., Schneeweiss, A. L., Rothman, K. J., Glynn, R. J., Avorn, J. & Sturmer, T. (2006). Variable selection for propensity score models. *Am. J. Epidemiol.* **163**, 1149–56.

Breslow, N. (1982). Design and analysis of case control studies. *Annual Rev. Public Health* **3**, 29–54.

Cochran, W. C. (1968). The effectiveness of adjusting by subclassification in removing bias in observational studies. *Biometrics* **24**, 295–313.

De Stavola, H. L. & Cox, D. R. (2008). On the consequences of overstratification. *Biometrika* **95**, 992–6.

Gail, M. (1988). The effect of pooling across strata in perfectly balanced studies. *Biometrics* **44**, 151–62.

Gill, R., van der Laan, M. & Robins, J. M. (1997). Coarsening at random: characterizations, conjectures and counterexamples. In *Proc. 1st Seattle Symp. Biostatist.* Ed. D. Lin and T. Fleming, pp. 255–94. New York: Springer.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.* **22**, 719–48.

Neuhausaer, M. & Becher, H. (1997). Improved odds ratio estimation by post-hoc stratification of case-control data. *Statist. Med.* **16**, 993–1004.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.

Robinson L. and Jewell, N. P. (1991) Some surprising results about covariate adjustment in logistic regression models. *Int. Statist. Rev.* **59,** 227–40.

[*Received November* 2009. *Revised April* 2010]