The Stata Journal (2008)
8, Number 4, pp. 532–539

Erratum and discussion of propensity-score reweighting

Austin Nichols
Urban Institute
Washington, DC
austinnichols@gmail.com

Keywords: st0136_1, xtreg, psmatch2, nnmatch, ivreg, ivreg2, ivregress, rd, lpoly, xtoverid, ranktest, causal inference, match, matching, reweighting, propensity score, panel, instrumental variables, excluded instrument, weak identification, regression, discontinuity, local polynomial

1 Erratum

Nichols (2007) described estimating the probability that an observation receives a binary "treatment" as a function of observable variables X (by using, e.g., `logit` or `probit`) and described using the estimated probabilities of treatment, or "propensity scores", $\widehat{\lambda}$, to reweight the data (as an alternative to matching). Section 3.4 of that article neglects to mention that the weights $\widehat{\lambda}/\left(1-\widehat{\lambda}\right)$ should be applied only to make the control group's outcomes represent the counterfactual outcomes of the treatment group by making the groups similar with respect to observable characteristics. The reweighting makes the mean of each variable in matrix X (i.e., those variables included in the propensity-score model) approximately equal across the treatment and control groups. The examples in section 3.5 of that article also neglect this restriction on propensity-based weights.

That is, the line

```
generate w=_ps/(1-_ps)
```

which generates a weight equal to $\widehat{\lambda}/\left(1-\widehat{\lambda}\right)$, should be followed by the command

```
replace w=1 if _tr
```

(where `_tr` is a treatment indicator and `_ps` is the propensity score). This makes the weight equal to 1 for observations receiving treatment (e.g., those belonging to a union or having completed college), i.e., those having `_tr==1`. It is also advisable to scale weights within the treatment and control groups so that the reweighted proportions are similar to those observed in the original sample. In fact, the reweighting of the control group to resemble the treatment group is only one of several plausible reweighting schemes, and a regression of outcomes on a treatment indicator using this weight can be considered an estimate of the average treatment effect (ATE) on the treated.

2 Alternative weighting schemes

The above pair of commands generating weights can be replaced by the single command

```
generate w=cond(_tr,1,_ps/(1-_ps))
```

with the same result. A rescaled weight to approximately preserve proportions in treatment and control groups would be

```
summarize _tr
generate w1=cond(_tr,r(mean)/(1-r(mean)),_ps/(1-_ps))
```

Multiplying treatment weights by $p/(1-p)$, where p is the proportion of the sample receiving treatment; multiplying control weights by $(1-p)/p$; or multiplying treatment weights by p and control weights by $(1-p)$ all produce identical results if weights themselves are rescaled to sum to N (Stata internally rescales `aweight`s to sum to N). The weight $\widehat{\lambda}/\left(1-\widehat{\lambda}\right)$ for untreated "control" observations reweights the distribution of observable characteristics included in the `logit` or `probit` model to be like that of the treated group. A weighted regression of outcome on treatment is thus a comparison of means across treatment and control groups, but the control group is reweighted to represent the average outcome that the treatment group would have exhibited in the absence of treatment. That is, every control group observation is contributing to an estimate of the mean counterfactual outcome for all treated observations (rather than specific observations being matched).

An alternative weighting scheme of the form

```
summarize _tr
generate w2=cond(_tr,(1-_ps)/_ps*r(mean)/(1-r(mean)),1)
```

reweights the distribution of observables in the treatment group to be like that of the control group. A comparison of means across (reweighted) treatment and control groups, for example, by using a weighted regression of an outcome variable on the treatment indicator, is then an estimate of the ATE on the controls. The treatment group is reweighted to represent the average outcome that the control group would have exhibited in the presence of treatment.

One method of computing an estimate of the ATE for the population is to take the weighted mean of these two estimates, with the weight attached to the ATE on the treated equal to the proportion receiving treatment, and with the weight attached to the ATE on the controls equal to one minus the proportion receiving treatment.

An alternative estimate of the ATE is available. First, the outcome under treatment for the whole population, i.e., the mean outcome if every unit received treatment, can be estimated by a weighted mean of outcomes in the treatment group with the weights $1/\widehat{\lambda}$ (Brunell and DiNardo 2004). Similarly, the outcome under control for the whole population, i.e., the mean outcome if every unit received no treatment, can be estimated by a weighted mean of outcomes in the control group with the weights $1/\left(1-\widehat{\lambda}\right)$. The weights for both groups are given by

```
summarize _tr
generate w3=cond(_tr,1/_ps*r(mean)/(1-r(mean)),1/(1-_ps))
```

An ATE estimate is then simply a weighted comparison of mean outcomes in the treatment and control groups (e.g., via a weighted regression of the outcome on a treatment indicator, and possibly covariates for the so-called "double robust estimator"). One problem that is exacerbated in this scheme is measurement error in the estimated propensity score; as DiNardo (2002) writes, "Small errors in estimating $\rho(x)$ can produce potentially large errors in the weights. Since the weight is a nuisance parameter from the viewpoint of estimating a density or a specific moment of the distribution, this is not a straightforward problem."

A fourth reweighting scheme,

```
generate w4=cond(_tr,(1-_ps),_ps)
```

minimizes the observable distance between the treatment and control groups in the sense that a test statistic for the difference in means (the Hotelling test) is zero (and the weighted groups are of equal size, so the mean of the treatment indicator is one half), but a difference in means using this weight is not so readily interpreted as an ATE. Nevertheless, simulation evidence not presented here indicates that it can be very effective (in the sense of having small bias and mean squared error) in estimating the ATE, especially when the estimated propensities are near zero or one.[1] It also exhibits good robustness to omitted variables in the selection equation (the first-stage `logit` or `probit`).

See Lunceford and Davidian (2004) and Busso, DiNardo, and McCrary (2008) for additional discussions of the construction of weights and rescaling, including an asymptotically variance-minimizing choice.

3 Results of reweighting

The results of reweighting are clear in a Hotelling test or an equivalent linear discriminant model. The example below,[2] using `hotelling` and `regress`, gives identical F statistics, but the `regress` approach allows relaxing of the assumption of equal variance across groups via the `vce(robust)` option.

1. Estimated propensities near zero or one represent a possible violation of the condition required for matching or reweighting that the probability of treatment is bounded away from zero and one. Here it is advisable to restrict to a subpopulation in which estimated propensities are never near zero or one and reestimate. The densities of propensities near the zero and one boundaries should be estimated by using `kdens` with boundary correction options, available from the Statistical Software Components archive. Note also that the probability of treatment must be strictly bounded away from zero and one to satisfy the assumptions, implying not only that the density should be zero at the boundaries but that the derivative of the density function should be zero at the boundaries.
2. This extract of the 1968 National Longitudinal Survey of Young Women 14–26 years of age does not include sample weights, but in general, we would prefer to convolve the weights by multiplying our reweighting factor by the sample weights.

```
webuse nlswork, clear
keep if year==77
local x "collgrad age tenure not_smsa c_city south nev_mar"
hotelling `x´, by(union)
regress union `x´
regress union `x´, vce(robust)
logit union `x´
predict _ps if e(sample)
summarize union if e(sample)
local p=r(mean)
generate w3=cond(union,`p´/_ps/(1-`p´),1/(1-_ps))
hotelling `x´ [aw=w3], by(union)
regress union `x´ [aw=w3]
regress union `x´ [aw=w3], vce(robust)
regress ln_wage union `x´ [aw=w3], vce(robust)
```

The F statistic drops from 20 to 0.1 after reweighting (18 to 0.1 when using heteroskedasticity-robust statistics), and the weighted means of each individual variable look much closer. The last regression of log(`wage`) on `union` using the inverse-probability weights based on propensity scores gives an estimate of the effect of union membership on wages, over both union and nonunion workers, suggesting that an individual would earn 14% more in 1977 as a union member than as a nonunion worker, on average. Instead, using weights generated by

```
generate w1=cond(union,`p´/(1-`p´),_ps/(1-_ps))
regress ln_wage union `x´ [aw=w1], vce(robust)
```

gives an estimate of the effect of union membership on wages for union members, suggesting that union members earned 14.5% more in 1977 than they would have as nonunion workers.

What is not clear from `hotelling` or `regress` is if the distributions of variables are similar; even if the means of X variables are equal in the reweighted sample, that does not imply that their distributions are similar. As long as treatment status can be inferred from higher moments of the X variables, we have not fully controlled for the observable differences across treatment and control groups. In practice, however, reweighting to make means match seems to make the distributions of observables very similar, for the same reason that matching on the propensity score does.

The difference in distributions is most clearly observable in the distribution of estimated propensity scores but can be seen in the individual variables (e.g., `tenure` in the example below; `kdens` is available from the Statistical Software Components archive).

(*Continued on next page*)

```
webuse nlswork, clear
keep if year==77
local x "collgrad age tenure not_smsa c_city south nev_mar"
logit union `x´
predict _ps if e(sample)
kdens _ps if union, bw(.03) ll(0) ul(1) gen(f1 x) nogr
kdens _ps if !union, bw(.03) ll(0) ul(1) gen(f0) at(x) nogr
label var f0 "pdf of propensities for unweighted non-union (control) obs"
label var f1 "pdf of propensities for unweighted union (treatment) obs"
line f1 f0 x, leg(col(1)) name(unwtd, replace)
summarize union if e(sample)
local p=r(mean)
generate w3=cond(union,`p´/(1-`p´),_ps/(1-_ps))
kdens _ps if union [aw=w3], bw(.03) ll(0) ul(1) gen(g1 x1) nogr
kdens _ps if !union [aw=w3], bw(.03) ll(0) ul(1) gen(g0) at(x1) nogr
label var g0 "pdf of propensities for reweighted non-union (control) obs"
label var g1 "pdf of propensities for reweighted union (treatment) obs"
line g1 g0 x1, leg(col(1)) name(rewtd, replace)
kdens tenure if union [aw=w3], bw(1.5) ll(0) gen(td) at(tenure) nogr
label var td "Density for union members"
kdens tenure if !union [aw=w3], bw(1.5) ll(0) gen(cd) at(tenure) nogr
label var cd "Density for nonunion reweighted to resemble union members"
line td cd tenure, sort leg(col(1))
```

Matching on the propensity score ensures that the distributions of estimated propensity scores are virtually identical in (matched) treatment and control groups, especially if matching models are iterated until balance is achieved, but reweighting does not. For example, if the distribution of some variable (including propensity scores) is bimodal in the control group and single-peaked in the treatment group, those properties will typically still be observable in the reweighted data. Nevertheless, reweighting achieves much of the balancing achievable via matching on the propensity score.

The last approach to reweighting always achieves the smallest difference in means, with an F statistic as close to zero as is feasible given machine precision, but the distributions of observable characteristics and estimated propensity scores are very similar under all these approaches to reweighting. See Iacus, King, and Porro (2008) for an alternative matching method that controls, up to user-chosen specified levels, "for all imbalances in central absolute moments, comoments, coskewness, interactions, nonlinearities, and other multidimensional distributional differences between treated and control groups".

4 Uses of reweighting

The propensity-based reweighting approach is at the heart of the method proposed by DiNardo, Fortin, and Lemieux (1996). The paradigmatic example of that approach uses two years of data, estimates the probability that an observation is in the first year or the second, then reweights the second year's observations by $\widehat{\lambda}/\left(1-\widehat{\lambda}\right)$ so that the distributions are nearly equal across the two years. Changes in means or distributions (of some outcome variables) in the reweighted data are then interpreted as estimates of change had the means of the X variables not changed over time.[3]

A similar method could be applied to estimate the proportion of a wage gap observed across men and women or white and nonwhite workers that is attributable to characteristics, along the lines of the `oaxaca` method described by Jann (2008) and related methods referenced there. The connections between these methods are discussed by, e.g., DiNardo (2002) and Lemieux (2002).

The reweighting approach extends easily to a polytomous categorical treatment variable, by considering the analogy to the DiNardo, Fortin, and Lemieux (1996) approach applied to multiple years. For example, each subsequent year's data can be reweighted to have observable characteristics similar to the first year, or each year can be reweighted to match some other base year's distribution. In the same way, observations receiving various levels of treatment can be reweighted to match some base category (the choice of base category can affect the interpretation of results).

Extensions of the reweighting approach to the case of a continuous treatment are also possible by using the generalized propensity-score approach of Hirano and Imbens (2004), described by Bia and Mattei (2008). The generalized propensity score, $r(t \mid x)$, is the density of treatment conditional on $X = x$, estimated as $\widehat{r}(t, x)$.

5 Missing data and a conjecture

This kind of reweighting could also be employed to correct for bias because of missing data. For example, the distributions of variables observed for both survey respondents and nonrespondents (i.e., potential stratification variables) can be adjusted via reweighting to look similar. Then the hope is that the unobservable survey responses of nonrespondents will be suitably captured by reweighted respondents. An alternative approach, imputing responses to nonrespondents, is a form of matching. Depending on the type of imputation, this can be propensity-score matching, nearest-neighbor matching, or exact matches on observables (also known as hotdeck imputation).

3. See also Altonji, Bharadwaj, and Lange (2008) for a recent article dealing not only with differences in distributions in samples but also with sample attrition and missing values.

The standard approach to missing data is to multiply impute responses (see, e.g., Carlin, Galati, and Royston [2008]). It is natural to wonder whether multiple imputation could also be fruitfully applied to the imputation of hypothetical counterfactual outcomes (the unobserved outcomes of treatment cases when in the control group, or the outcomes of control cases when in the treatment group).

6 Acknowledgments

I thank without implicating John DiNardo and Ben Jann for helpful conversations.

7 References

Altonji, J. G., P. Bharadwaj, and F. Lange. 2008. Changes in the characteristics of American youth: Implications for adult outcomes. Working paper, Yale University. http://www.econ.yale.edu/~fl88/SkillComposition.pdf.

Bia, M., and A. Mattei. 2008. A Stata package for the estimation of the dose–response function through adjustment for the generalized propensity score. *Stata Journal* 8: 354–373.

Brunell, T. L., and J. DiNardo. 2004. A propensity score reweighting approach to estimating the partisan effects of full turnout in American presidential elections. *Political Analysis* 12: 28–45.

Busso, M., J. DiNardo, and J. McCrary. 2008. Finite sample properties of semiparametric estimators of average treatment effects. Working paper, University of Michigan. http://www-personal.umich.edu/~jdinardo/BDM2008_v11.pdf.

Carlin, J. B., J. C. Galati, and P. Royston. 2008. A new framework for managing and analyzing multiply imputed data in Stata. *Stata Journal* 8: 49–67.

DiNardo, J. 2002. Propensity score reweighting and changes in wage distributions. Working Paper, University of Michigan. http://www-personal.umich.edu/~jdinardo/bztalk5.pdf.

DiNardo, J., N. M. Fortin, and T. Lemieux. 1996. Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach. *Econometrica* 64: 1001–1044.

Hirano, K., and G. W. Imbens. 2004. The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, ed. A. Gelman and X.-L. Meng, 73–84. Chichester, UK: Wiley.

Iacus, S. M., G. King, and G. Porro. 2008. Matching for causal inference without balance checking. Downloadable from http://gking.harvard.edu/cem.

Jann, B. 2008. The Blinder–Oaxaca decomposition for linear regression models. *Stata Journal* 8: 453–479.

Lemieux, T. 2002. Decomposing changes in wage distributions: A unified approach. *Canadian Journal of Economics* 35: 646–688.

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

Nichols, A. 2007. Causal inference with observational data. *Stata Journal* 7: 507–541.

About the author

Austin Nichols is an economist at the Urban Institute, a nonprofit, nonpartisan think tank. He occasionally teaches statistics and econometrics, and he has used Stata almost daily since 1995. His research interests include poverty, social insurance, tax policy, and demographic outcomes such as fertility, marital status, health, and education.