

|  |  |
| --- | --- |
| **MEMORANDUM** | 1100 1st Street, NE, 12th Floor  Washington, DC 20002-4221  Telephone (202) 484-9220  Fax (202) 863-1763  www.mathematica-mpr.com |

**TO:** CPC+ Impacts Team

**FROM:** Keith Kranker **DATE:** [Publish Date]

**SUBJECT**: Technical description of matching methods used for reweighting the Round 1 internal comparison group

To address selection bias, the CPC+ impact analysis team used comparison group design where CPC+ practices were matched to comparison practices. The resulting matched comparison group was similar to the treatment group on observable characteristics (or “balanced”), reducing the likelihood that underlying differences are responsible for estimated effects of CPC+ (see, for example, Rosenbaum and Rubin 1983; Dehejia and Wahba 2002; Imbens and Rubin 2015; Stuart 2010). Matching can address concerns that regression-based impact estimates could be biased when functional form assumptions do not hold and the treatment and comparison groups are not balanced (Imbens and Wooldridge 2009; Imbens 2015).

We constructed the Round 1 internal comparison group using novel methods that allowed us balance the tradeoff between balance and statistical power. Compared to comparison groups constructed using other methods, our comparison group will have more statistical power (smaller standard errors) in the impact analyses. The improved power comes at the expense of (slightly) worse balance between the treatment group and comparison group, although measures of imbalance remained within pre-specified tolerance levels (cite report).

The methods used for constructing the comparison group were related to the *covariate-balancing propensity score* (CBPS) from Imai and Ratkovic (2014). The CBPS methods are, in turn, based on older and more familiar method—inverse propensity weighting (IPW). In this memo, I briefly introduce the IPW and CBPS methods, and document the modifications made to the methods in our new approach. Then I describe the programming code I wrote to implement the new method in Stata.

A. Overview of the matching methods

1. Inverse propensity weighting (IPW)

IPW re-weights observations in the potential comparison group to reduce the imbalance between the treatment and comparison group on one or more observed covariates (or “matching variables”). Each comparison observation is weighted by the inverse of the estimated probability that the observation is in the treatment group—that is, the inverse of the observation’s propensity score.[[1]](#endnote-1) In most cases, after reweighting with IPW, the sample to exhibit better covariate balance between the treatment and comparison groups than the original (unweighted). IPW tends to improve balance on variables associated with treatment status.

IPW involves several straightforward steps. The first step involves estimating a propensity score model—treatment status is modeled as a function of the matching variables, . The propensity score model is traditionally estimated by logistic regression, although other estimation methods are available and can be superior in some settings (for example, see Lee, Lessler, and Stuart 2009). Second, estimated propensity scores are calculated for each observation, *i*, using the model: . In the case of logistic regression, . Third, a formula is used to convert the propensity score for each observation into a corresponding weight. When computing weights for estimating the average treatment effect on the treated (ATET), the treatment group observations receive a matching weight of , and comparison group observations receive a weight equal to .[[2]](#endnote-2) To bound the influence function, some researchers then normalize the matching weights to have mean equal 1.[[3]](#endnote-3) That is, . Fourth, the weights are applied to the data, allowing researches to (1) assess balance and (2) estimate treatment effects with the reweighted sample. The treatment effect is estimated either as the difference in weighted means between the treatment and comparison groups or estimated by weighted regression.[[4]](#endnote-4)

The matched comparison group includes all the observations in the comparison group, with each observation receiving a continuous weight that are greater than zero, but are theoretically unbounded (they could approach infinity).[[5]](#endnote-5) IPW estimators can be unstable when the treatment and comparison groups have poor overlap. Poor overlap causes some IPW weights to become very large. A potential workaround is to restrict the propensity scores to a region of common empirical support—that is, remove observations with extremely high or low propensity scores (Crump et al. 2009; StataCorp 2017).

2. Covariate-balancing propensity score (CBPS)

Imai and Ratkovic (2014) introduced a new IPW-type reweighting methodology, which they named a *covariate-balancing propensity score* (CBPS). The key insight that led to the CBPS paper was that researchers may want to find a set of propensity scores (and hence ATET weights) that simply achieve the best balance possible. This differs from an IPW methods, where the propensity score models is estimated (by maximum likelihood) to best predict treatment status. Starting with the IPW approach; it would be possible to “tweaking” the coefficients in the (logit) propensity score model up or down and improve balance on the matching variables. As a trivial example, Exhibit 1 (from Kranker 2015) shows how CBPS can (slightly modify the coefficients in the logit propensity score model and, by doing so, significantly improve covariate balance.

CBPS methods combine steps in the IPW workflow (above) into a single model. Numerical optimization methods are used to find the set of coefficients () in the propensity score model (step 1) that result in the least imbalance (in step 4). The CBPS method can be thought of as searching for the propensity score model coefficients that minimize the following function:

where , the loss function, is a measure of imbalance between the treatment and weighted comparison groups given the data and a vector of weights , covariate-balancing propensity scores are defined as , and the weight for observation is calculated as a function of using the formulas from IPW.[[6]](#endnote-6) Imai and Ratkovic (2014) measured imbalance using a multivariate analogue to standardized bias; it measures the overall imbalance across the matching variables and is expressed in standardized units. The Stata code discussed below implements this measure and several alternative measures of imbalance as well.

3. Alternative CBPS method used for CPC+

With the CPC+ data, we found that CBPS produced excellent balance between the CPC+ and comparison practices. However, some of the comparison practices were receiving very large weights (say, over 25 and sometimes as large as 50 or 75) and power calculations revealed that the wide distribution in weights were likely to substantively reduce the statistical power in the outcome models.[[7]](#endnote-7),[[8]](#endnote-8) The team suspected that the CBPS approach could be “tweaked” in a way that would result in fewer “extreme” weights—leading to improved statistical power—with only modest reductions in balance. This hypothesis gave rise to model that was similar to CBPS, but modified the loss function as follows:

where , , and are the coefficient of variation (CV), skewness, and kurtosis of the vector of weights (), respectively; the tuning parameters are specified by the user; and the remaining functions, variables, and parameters are defined the same as in the CBPS model. This numerical optimization problem still minimizes the original CBPS loss function, but a penalty is applied if the distribution of the weights differences from the distribution “targeted” by the user (specifically, if the CV, skewness, or kurtosis was different than specified through the parameters ). The user has control over the size of the penalty. This alternative model is not particularly grounded in statistical theory. Rather, it (1) was a pragmatic approach based on understanding the inner workings of the CBPS model and (2) was found to work well in practice with some test datasets and, ultimately, the CPC+ data.

A key question that arises when implementing the model is how to choose the tuning parameters. We found it practical to use a grid search. The procedure began by estimating the CBPS model and noting the CV, skewness, and kurtosis of the matching weights. Then, we systematically tried a range of tuning parameters , with the aim of lowering the CV and skewness of the matching weights to a more desirable level.[[9]](#endnote-9) With each set of candidate parameters, we estimated the modified CBPS model, constructed matching weights, measured covariate imbalance (in multiple ways), measured the distribution of the matching weights, and calculated statistical power. In some cases, initial results helped us add tuning parameters to the grid. The models that yielded better covariate balance tended to have weights with a larger CV or skewness (and hence worse power). The team then tabulated the various results, and identified a set of parameters that we thought gave the best tradeoff between covariate balance and power, based on a broad understanding of the goals of the CPC+ evaluation.

4. Sampling weights

It is straightforward to adopt all three methods to cases where the original data are weighted. For IPW, weighted models are used to estimate the propensity scores. For both IPW and CBPS, the data are weighted using the product of the original weights and Matching weights, , are normalized so that *weighted* mean of the matching weights equals one (see footnote 3).

B. Programming implementation

I wrote programming code to implement this new approach. The code was written in Stata’s matrix programming language, Mata, using object-oriented (class) programming methods. Users need to understand and use Mata to have full access to all of the capabilities available in the code. To make the core methods more accessible to some users, a “wraparound” Stata command is provided to give easy access to the main techniques from Stata’s main programming language. This section first describes the Stata command, and then describes the underlying Mata implementation.

1. Stata implementation

a. Syntax

b. Description

c. Options

d. Examples

2. Mata class

Exhibit . Example of IPW and CBPS methods

|  |
| --- |
| . webuse cattaneo2  . logit mbsmoke mmarried c.mage##c.mage fbaby medu  . cbps  mbsmoke mmarried c.mage##c.mage fbaby medu, att logit  --------------------------------------------------------------------------                                                         (1)          (2)  Covariate                     Variable name            IPW         CBPS  --------------------------------------------------------------------------  Mother married                mmarried              -1.146       -1.188                                                    (-12.47)     (-12.47)  Mother's age (years)          mage                   0.322        0.327                                                      (5.04)       (5.11)  Mother's age, squared         c.mage#c.mage       -0.00604     -0.00620                                                     (-5.09)      (-5.17)  First baby                    fbaby                 -0.386       -0.403                                                     (-4.39)      (-4.54)  Mother's education (years)    medu                  -0.142       -0.105                                                     (-8.20)      (-8.01)  Constant                      \_cons                 -2.951       -3.400                                                     (-3.64)      (-4.19)  --------------------------------------------------------------------------  Sample size                   N                       4642         4642  --------------------------------------------------------------------------  Imbalance                     imbalance\_atet         0.203        0.000  --------------------------------------------------------------------------  t statistics in parentheses. The data is from Cattaneo (2010). |

Endnotes

References

Busso, M., J. DiNardo, and J. McCrary. 2014. “New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators.” *Review of Economics and Statistics*, 96(5): 885–897, doi:10.1162/REST\_a\_00431.

Cattaneo, M. D. 2010. “Efficient Semiparametric Estimation of Multi-Valued Treatment Effects under Ignorability.” *Journal of Econometrics*, 155(2): 138–154, doi:10.1016/j.jeconom.2009.09.023.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik. 2009. “Dealing with Limited Overlap in Estimation of Average Treatment Effects.” *Biometrika*, 96(1): 187–199, doi:10.1093/biomet/asn055.

Dehejia, R. H. and S. Wahba. 2002. “Propensity Score-Matching Methods for Nonexperimental Causal Studies.” *Review of Economics and Statistics*, 84(1): 151–161, doi:10.1162/003465302317331982.

Guo, S. and M. W. Fraser. 2010. Propensity Score Analysis: Statistical Methods and Applications, Thousand Oaks, CA: SAGE Publications.

Hansen, L. P. 1982. “Large Sample Properties of Generalized Method of Moments Estimators.” *Econometrica*, 50(4): 1029–1054, doi:10.2307/1912775.

Hirano, K., G. W. Imbens, and G. Ridder. 2003. “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score.” *Econometrica*, 71(4): 1161–1189.

Horvitz, D. G. and D. J. Thompson. 1952. “A Generalization of Sampling Without Replacement from a Finite Universe.” *Journal of the American Statistical Association*, 47(260): 663–685, doi:10.1080/01621459.1952.10483446.

Huber, M., M. Lechner, and C. Wunsch. 2013. “The Performance of Estimators Based on the Propensity Score.” *Journal of Econometrics*, 175(1): 1–21, doi:10.1016/j.jeconom.2012.11.006.

Imai, K. and M. Ratkovic. 2014. “Covariate Balancing Propensity Score.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1): 243–263, doi:10.1111/rssb.12027.

Imbens, G. W. 2015. “Matching Methods in Practice: Three Examples.” *Journal of Human Resources*, 50(2): 373–419, doi:10.3368/jhr.50.2.373.

Imbens, G. W. and D. B. Rubin. 2015. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction, New York: Cambridge University Press.

Imbens, G. W. and J. M. Wooldridge. 2009. “Recent Developments in the Econometrics of Program Evaluation.” *Journal of Economic Literature*, 47(1): 5–86, doi:10.1257/jel.47.1.5.

Kranker, K. 2015. Multiple-Equation Models: An Introduction and Potential Applications to Our Work at Mathematica. Available at: http://mathnet.mathematica.net/topic/interest/designworkshop/.

Lee, B. K., J. Lessler, and E. A. Stuart. 2009. “Improving Propensity Score Weighting Using Machine Learning.” *Statistics in Medicine*, 29(3): 337–346, doi:10.1002/sim.3782.

Nichols, A. 2008. “Erratum and Discussion of Propensity-Score Reweighting.” *Stata Journal*, 8(4): 532–539.

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky. 2007. “Comment: Performance of Double-Robust Estimators When ‘Inverse Probability’ Weights Are Highly Variable.” *Statistical Science*, 22(4): 544–559, doi:10.1214/07-STS227D.

Rosenbaum, P. R. and D. B. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika*, 70(1): 41–55, doi:10.1093/biomet/70.1.41.

StataCorp. 2017. Stata: Release 15. Statistical Software, College Station, TX: StataCorp LLC.

Stuart, E. A. 2010. “Matching Methods for Causal Inference: A Review and a Look Forward.” *Statistical Science*, 25(1): 1–21, doi:10.1214/09-STS313.

1. For more details on IPW, see Cattaneo (2010); Guo and Fraser (2010); Hirano et al. (2003); Huber et al. (2013); Imbens and Rubin (2015); Nichols (2008) and Stuart (2010). Basic IPW methods can be implemented with Stata’s the teffects ipw command (StataCorp 2017). IPW matching weights are related to nonresponse weights used for survey data analyses (Horvitz and Thompson 1952). [↑](#endnote-ref-1)
2. When estimating causal impacts, the literature calls this estimate the *average treatment effects on the treated (ATET)*. That is, it produces estimated associations between treatment and outcomes among the observations in the treatment group. This should be distinguished from the *average treatment effect (ATE)*, which would alternatively estimate associations between treatment and outcomes in the hypothetical scenario in which all observations received the treatment, or *average treatment effects on the untreated (ATEU)*, which would estimate associations between treatment and outcomes in the hypothetical scenario in which the comparison observations received the treatment. Weights for ATE and ATEU estimation are similar to the weights used for ATE. Specifically, ATE matching weights are set to for treatment observations and for comparison observations. ATEU matching weights are set to for treatment observations and for comparison observations. [↑](#endnote-ref-2)
3. If the weights are normalized to have a mean equal 1, then the sum of the weights equals the number of observations. The weights can be normalized across the pooled sample or separately by treatment status. See Busso, DiNardo, and McCrary (2014), Robins et al. (2007), and StataCorp (2017) for more details on normalizing the weights. [↑](#endnote-ref-3)
4. For purpose of estimating standard errors, the propensity score model and outcome model can be estimated jointly (as a system of equations), so that the standard errors in the outcome model account for the fact that that weights were estimated (Cattaneo 2010; Hansen 1982; Kranker 2015). Stata’s teffects ipw and teffects ipwra commands, for example, estimate standard errors this way (StataCorp 2017). [↑](#endnote-ref-4)
5. In this regard, IPW differs from other matching methods, which typically give matched comparison group observations a weight that is an integer or reciprocal of an integer (and give unmatched comparison group observations a weight of zero). [↑](#endnote-ref-5)
6. Imai and Ratkovic (2014) estimate model by generalized method of moments (GMM). They provide two versions of the model; the “just identified” model minimizes overall imbalance while the “over identified” model minimizes overall imbalance and the fit the propensity score model. This section of the memo describes the “just identified” model; the Stata code discussed below implements both CBPS models. [↑](#endnote-ref-6)
7. A related concern was that the large weights could negatively affect the survey, since effective nonresponse rates would be large if one or more of the survey respondents in comparison practices with large weights did not respond. [↑](#endnote-ref-7)
8. Another issue, not discussed in this memo, was that IPW and CBPS gave a large number of comparison extremely small weights, indicating these practices were unlike the CPC+ practices. These practices were trimmed from the sample before matching. [↑](#endnote-ref-8)
9. With the CPC+ data, we set to zero since the weights had acceptable kurtosis in all iteratioins. [↑](#endnote-ref-9)