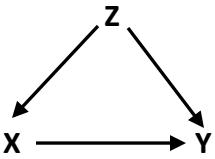
Techniques for Robustness & Threats to Inference: Inverse Probability Weighted Regression Adjustment

Selena Caldera

March 12, 2019

Using Inverse Probability Weighted Regression Adjustment to Estimate Unbiased Treatment Effects

IPWRA is one approach to estimate unbiased treatment effects when we have confounding.



We find this often with observational data – we observe some treatment but no randomization of assignment to treatment.

- Confounding due to selection bias
- Are selection characteristics observed in the data? If so, we can condition treatment on those characteristics to get an unbiased estimate of treatment effect

Conceptually, IP weighting:

1. Estimates selection to treatment (treatment model)
2. Predicts treatment for all observations
3. Assigns the inverse of probability of treatment for treated individuals AND the inverse probability of not being treated for control individuals
4. Re-estimates the outcome model using these new weights

The IP weights magnify treatment individuals who otherwise look like they would not have selected treatment and magnify control individuals who otherwise look like they would have selected treatment. We create counterfactuals where they are not observed in the data.

One important feature of IPWRA is ***double robustness***. Even if one of the models (treatment or outcome) is mis-specified, the estimator is still consistent. You can get one wrong and still be right!

The examples use an example Health Cost and Utilization Project dataset from Cattaneo (2010) Journal of Econometrics 155: 138-154.

We look at how mother's smoking affects a baby's birth weight. Theory tells us that the following covariates are also associated with birth weight:

- mother's age
- whether mother had a prenatal visit in the 1st trimester
- marital status of mother
- whether this is her first baby

We include these as covariates in the model of smoking status on baby's birth weight.

```
. set more off

. global homedir  "C:\Users\selen\OneDrive\2018_19 PRC Stats Consulting"

. global logdir   "$homedir\log files"

. global datadir  "$homedir\data"

. global output   "$homedir\output"

. use "$datadir\cattaneo2.dta", clear
(Excerpt from Cattaneo (2010) Journal of Econometrics 155: 138-154)
```

Our descriptive analysis of the data shows that mothers who smoke tend to be:

- younger
- have lower levels of educational attainment
- a smaller share of the mother's who smoke are having their first baby
- a smaller share of the mother's who smoke are married

Many of these selection characteristics might also influence baby's weight at birth (confounding).

Estimate treatment model, generate predicted conditional probabilities, and generate IP weights separately (based on code from Hernan & Robins)

In this example we use a probit model that includes all the covariates in our outcome model plus mother's age squared & mother's education. Mother's smoking status is the outcome.

```
. probit mbsmoke i.mmarried c.mage##c.mage i.fbaby medu

Iteration 0:   log likelihood = -2230.7484
Iteration 1:   log likelihood = -2042.6734
Iteration 2:   log likelihood = -2040.5088
Iteration 3:   log likelihood = -2040.5061
Iteration 4:   log likelihood = -2040.5061

Probit regression                                 Number of obs     =      4,642
                                                  LR chi2(5)        =     380.48
                                                  Prob > chi2       =     0.0000
Log likelihood = -2040.5061                       Pseudo R2         =     0.0853

------------------------------------------------------------------------------
      mbsmoke |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
     mmarried |
     married  |  -.6484821   .0526991   -12.31   0.000    -.7517705   -.5451938
        mage  |   .1744327   .0352437     4.95   0.000     .1053562    .2435092
              |
c.mage#c.mage |  -.0032559   .0006462    -5.04   0.000    -.0045224   -.0019894
              |
        fbaby |
         Yes  |  -.2175962   .0491066    -4.43   0.000    -.3138433    -.121349
        medu  |  -.0863631   .0098692    -8.75   0.000    -.1057064   -.0670198
       _cons  |  -1.558255   .4511589    -3.45   0.001     -2.44251       -.674
------------------------------------------------------------------------------
```

Predict the conditional probability of smoking for each mother in the sample

```
. predict p_mbsmoke, pr
```

Now we generate the inverse probability weights as P(T=1|covariates) if T = 1 (mother is a smoker), and 1-P(T=1|covariates) if T = 0 (mother is a nonsmoker)

```
. gen w=.
(4,642 missing values generated)

. replace w=1/p_mbsmoke if mbsmoke==1
(864 real changes made)

. replace w=1/(1-p_mbsmoke) if mbsmoke==0
(3,778 real changes made)
```

Check the balance of the covariates after weighting:

Check the mean of the weights; we expect it to be close to 2.0:

```
. summarize w

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
           w |      4,642    1.980605     2.11765   1.007511   29.91177
```

Fit the outcome model using the inverse probability weights:

This creates a pseudo-population by averaging individual heterogeneity across the treatment and control groups.

We want heteroskedasticity-consistent SEs for our weighted estimators. Stata automatically calls the robust option when pweights are specified.

```
. regress bweight mbsmoke mage prenatal1 mmarried fbaby [pweight=w]
(sum of wgt is   9.1940e+03)

Linear regression                               Number of obs     =      4,642
                                                F(5, 4636)        =      51.29
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0549
                                                Root MSE          =     568.81

------------------------------------------------------------------------------
             |               Robust
     bweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mbsmoke |  -228.3259   26.22851    -8.71   0.000    -279.7462   -176.9055
       mage  |  -1.167128   3.298776    -0.35   0.724    -7.634299    5.300043
   prenatal1 |   53.68661   26.84983     2.00   0.046     1.048178     106.325
    mmarried |   143.6948   24.83004     5.79   0.000     95.01612    192.3735
       fbaby |  -18.15393   30.27812    -0.60   0.549    -77.51345    41.20559
       _cons |    3298.67   90.18191    36.58   0.000     3121.871     3475.47
------------------------------------------------------------------------------


. regress bweight mbsmoke mage prenatal1 mmarried fbaby

      Source |       SS           df       MS      Number of obs   =     4,642
-------------+----------------------------------   F(5, 4636)      =     56.62
       Model |  89487999.5         5  17897599.9   Prob > F        =    0.0000
    Residual |  1.4654e+09     4,636  316090.646   R-squared       =    0.0576
-------------+----------------------------------   Adj R-squared   =    0.0565
       Total |  1.5549e+09     4,641  335032.156   Root MSE        =    562.22

------------------------------------------------------------------------------
     bweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mbsmoke |  -226.9851   21.95345   -10.34   0.000    -270.0243   -183.9459
       mage  |   1.018963   1.736228     0.59   0.557     -2.38487    4.422796
   prenatal1 |   57.59001   22.24885     2.59   0.010     13.97169    101.2083
    mmarried |   154.4452   21.14817     7.30   0.000     112.9848    195.9057
       fbaby |  -52.07058    17.6856    -2.94   0.003    -86.74277   -17.39839
       _cons |   3245.509   46.50306    69.79   0.000     3154.341    3336.677
------------------------------------------------------------------------------
```

Use Stata's teffects

Stata's teffects ipwra command makes all this even easier and the post-estimation command, tebalance, includes several easy checks for balance for IP weighted estimators. Here's the syntax:

teffects ipwra (ovar omvarlist [, omodel noconstant]) /// (tvar tmvarlist [, tmodel noconstant]) [if] [in] [weight] [, stat options]

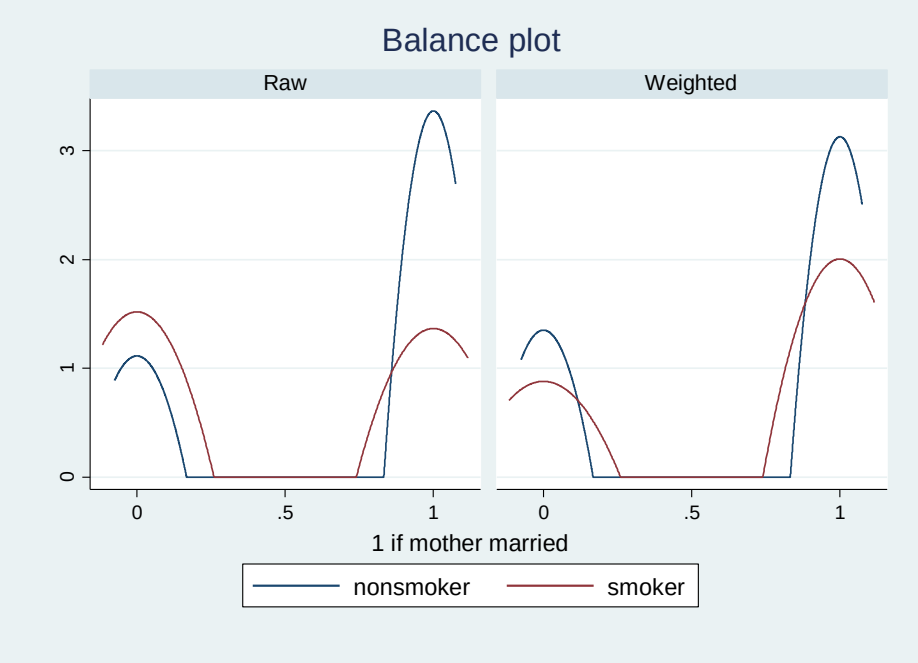
Outcome models may be linear (default), logit, probit, poisson, heteroskedastic probit, or fractional logit/probit. Treatment models may be logit, probit, heteroskedastic probit.

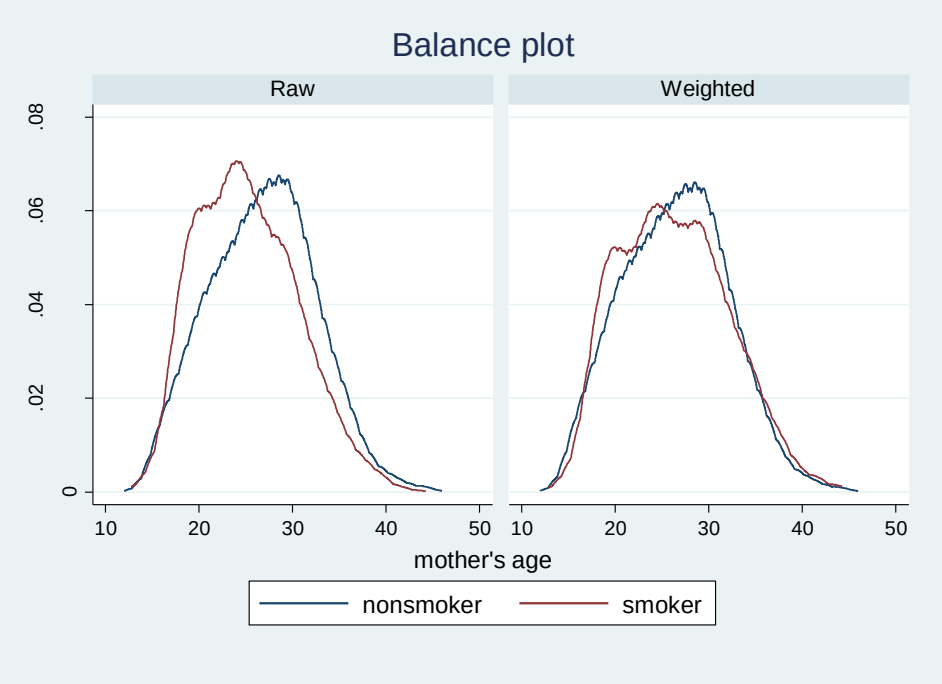
```
. teffects ipwra (bweight mage prenatal1 mmarried fbaby) ///
        (mbsmoke mmarried c.mage##c.mage fbaby medu, probit), aequations ate


Iteration 0:   EE criterion =  9.416e-21
Iteration 1:   EE criterion =  6.706e-26

Treatment-effects estimation                    Number of obs     =      4,642
Estimator      : IPW regression adjustment
Outcome model  : linear
Treatment model: probit
------------------------------------------------------------------------------
                             |               Robust
                     bweight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------------+------------------------------------------------
ATE                          |
                     mbsmoke |
(smoker vs nonsmoker)        |  -229.9671   26.62668    -8.64   0.000    -282.1544   -177.7798
-----------------------------+------------------------------------------------
POmean                       |
                     mbsmoke |
                   nonsmoker |   3403.336    9.57126   355.58   0.000     3384.576    3422.095
-----------------------------+------------------------------------------------
OME0                         |
                       mage |   2.893051   2.134788     1.36   0.175    -1.291056    7.077158
                   prenatal1 |   67.98549   28.78428     2.36   0.018     11.56933    124.4017
                    mmarried |   155.5893   26.46903     5.88   0.000      103.711    207.4677
                       fbaby |   -71.9215   20.39317    -3.53   0.000    -111.8914   -31.95162
                       _cons |   3194.808   55.04911    58.04   0.000     3086.913    3302.702
-----------------------------+------------------------------------------------
OME1                         |
                       mage |  -5.068833   5.954425    -0.85   0.395    -16.73929    6.601626
                   prenatal1 |   34.76923   43.18534     0.81   0.421    -49.87248    119.4109
                    mmarried |   124.0941   40.29775     3.08   0.002     45.11193    203.0762
                       fbaby |   39.89692   56.82072     0.70   0.483    -71.46966    151.2635
                       _cons |   3175.551   153.8312    20.64   0.000     2874.047    3477.054
-----------------------------+------------------------------------------------
TME1                         |
                    mmarried |  -.6484821   .0554173   -11.70   0.000     -.757098   -.5398663
                       mage |   .1744327   .0363718     4.80   0.000     .1031452    .2457202
                            |
                c.mage#c.mage |  -.0032559   .0006678    -4.88   0.000    -.0045647   -.0019471
                            |
                       fbaby |  -.2175962   .0495604    -4.39   0.000    -.3147328   -.1204595
                        medu |  -.0863631   .0100148    -8.62   0.000    -.1059917   -.0667345
                       _cons |  -1.558255   .4639691    -3.36   0.001    -2.467618   -.6488926
------------------------------------------------------------------------------
```

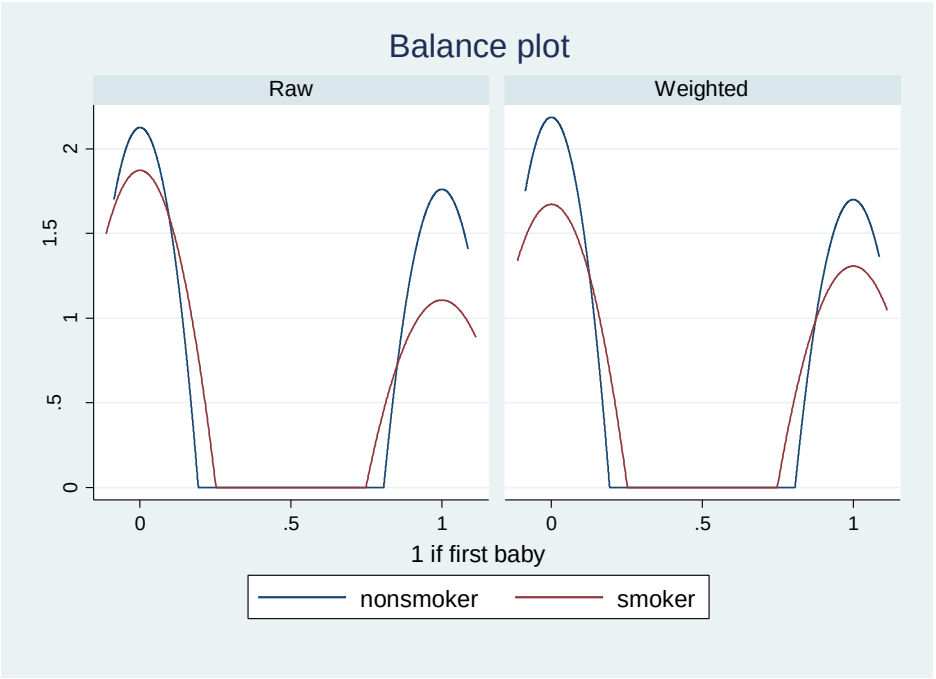
These results are close but differ slightly from the ones obtained above using Hernan & Robins's code. Why? Teffects estimates treatment-specific predicted outcomes (POs) for each subject then computes the means of these POs. These are contrasted to estimate the average treatment effect and average treatment effect on the treated.

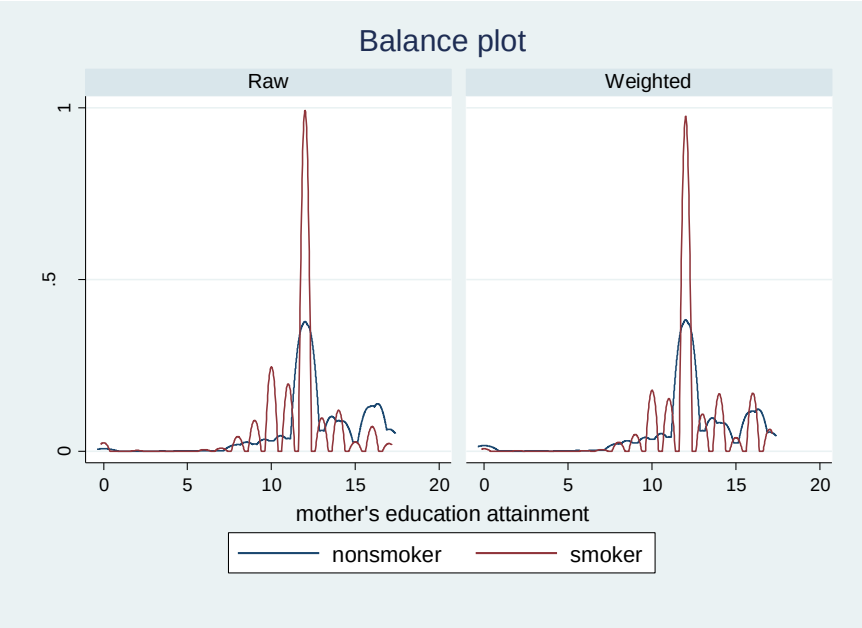
Let's make sure the treatment model balanced the covariates. Our treatment effects are only accurate if balance is achieved.

```
. tebalance summarize

  Covariate balance summary
                                                      Raw     Weighted
                          ----------------------------------------------
                          Number of obs =         4,642      4,642.0
                          Treated obs   =           864      2,290.8
                          Control obs   =         3,778      2,351.2
                          ----------------------------------------------

  ------------------------------------------------------------------------
                  |Standardized differences          Variance ratio
                  |        Raw    Weighted            Raw   Weighted
  ----------------+-------------------------------------------------------
         mmarried |  -.5953009   -.0073683      1.335944   1.006339
             mage |   -.300179   -.0363272      .8818025   1.050069
                  |
             mage#|
             mage |  -.3028275   -.0300786      .8274389    1.07782
                  |
            fbaby |  -.1663271    .0027075      .9430944   1.000687
             medu |  -.5474357   -.1042143      .7315846   .5192651
  ----------------+-------------------------------------------------------

. foreach var of varlist mmarried mage fbaby medu {
.         tebalance density `var', saving("$output\balance_`var'", replace)
.         }
(file C:\Users\selen\OneDrive\2018_19 PRC Stats Consulting\output\balance_mmarried.gph saved)
(file C:\Users\selen\OneDrive\2018_19 PRC Stats Consulting\output\balance_mage.gph saved)
(file C:\Users\selen\OneDrive\2018_19 PRC Stats Consulting\output\balance_fbaby.gph saved)
(file C:\Users\selen\OneDrive\2018_19 PRC Stats Consulting\output\balance_medu.gph saved)
```




Balance plot
Raw
Weighted
0
.02
.04
.06
.08
10
20
30
40
50
mother's age
nonsmoker
smoker


Balance plot
Raw
Weighted
0
.5
1
1.5
2
0
.5
1
1 if first baby
nonsmoker
smoker



Finally, we can run an overidentification test to check our findings from the diagnostics above.

```
. tebalance overid, nolog

Overidentification test for covariate balance
         H0: Covariates are balanced:

         chi2(6)      =  43.3799
         Prob > chi2  =   0.0000
```

It looks like we need to revisit our treatment model. There are options for using stabilized and trimmed IP weights that can account for the influence of outlier observations in your data. This should, however, get you started with exploring IPWRA.

Resources: A pre-publication version of *Causal Inference* plus SAS, Stata, R, and Python code for all the examples can be found here: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/. This on-line version is just generally an amazing methods resource!

See Morgan & Winship, *Counterfactuals and Causal Inference*, Ch. 6 for a more detailed discussion of double robustness and IPWRA.