001 Data analysis project - carry out an IPTW causal analysis

Iuliia Allaiarova

2024-03-01

For this project, I use data from Lalonde, that aimed to evaluate the impact of National Supported Work (NSW) Demonstration, which is a labor training program, on post-intervention income levels. Interest is in estimating the causal effect of this training program on income.The data have n=614 subjects and 10 variables.

Variable	Description
`age`	Age in years.
`educ`	Years of schooling.
`black`	Indicator variable for Blacks.
`hispan`	Indicator variable for Hispanics.
`married`	Indicator variable for marital status.
`nodegree`	Indicator variable for high school diploma.
`re74`	Real earnings in 1974.
`re75`	Real earnings in 1975.
`re78`	Real earnings in 1978.
`treat`	Indicator variable for treatment status.

The outcome variable is `re78` – post-intervention income.

The treatment variable is `treat` – equal to 1 if the subject received labor training and 0 otherwise.

The potential confounding variables are: `age`, `educ`, `black`, `hispan`, `married`, `nodegree`, `re74`, `re75`.

What I am going to do:

1. Fit a propensity score model.
2. Use a logistic regression model, where the outcome is treatment.
3. Include the 8 confounding variables in the model as predictors, with no interaction terms or non-linear terms (such as squared terms).
4. Obtain the propensity score for each subject.
5. Obtain the inverse probability of treatment weights for each subject.

This is the Table 1 for the data:

```
table1_m = CreateTableOne(vars=xvars_m,strata="treat", data=mydf, test=FALSE)
print(table1_m)
```

```
##                       Stratified by treat
##                        0                1
##   n                        429              185
##   age (mean (SD))        28.03 (10.79)    25.82 (7.16)
```

```
##   educ (mean (SD))       10.24 (2.86)      10.35 (2.01)
##   black (mean (SD))       0.20 (0.40)       0.84 (0.36)
##   hispan (mean (SD))      0.14 (0.35)       0.06 (0.24)
##   married (mean (SD))     0.51 (0.50)       0.19 (0.39)
##   nodegree (mean (SD))    0.60 (0.49)       0.71 (0.46)
##   re74 (mean (SD))     5619.24 (6788.75) 2095.57 (4886.62)
##   re75 (mean (SD))     2466.48 (3292.00) 1532.06 (3219.25)
```

Let's fit the propensity score model:

```
psmodel_m <- glm(treat ~ age + educ + black + hispan + married + nodegree + re74 + re75,
               family  = binomial(link ="logit"), data=mydf)
```

The minimum weight is 1.01 and the maximum weight is 40.08.

Let's find the standardized differences for each confounder on the weighted (pseudo) population. What is the standardized difference for nodegree?

```
##                       Stratified by treat
##                        0                 1                 SMD
##   n                     616.00            553.63
##   age (mean (SD))        27.10 (10.80)     25.57 (6.53)     0.172
##   educ (mean (SD))       10.29 (2.74)      10.61 (2.05)     0.132
##   black (mean (SD))       0.40 (0.49)       0.45 (0.50)     0.101
##   hispan (mean (SD))      0.12 (0.32)       0.12 (0.33)     0.014
##   married (mean (SD))     0.41 (0.49)       0.31 (0.47)     0.197
##   nodegree (mean (SD))    0.62 (0.48)       0.57 (0.50)     0.112
##   re74 (mean (SD))     4552.74 (6337.09) 2932.18 (5709.42)  0.269
##   re75 (mean (SD))     2172.04 (3160.14) 1658.07 (3072.89)  0.165
```

The standardized difference for nodegree is 0.11.

Using IPTW, I find the estimate and 95% confidence interval for the average causal effect. This can be obtained from svyglm:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.009   1.052   1.170   1.905   1.623  40.077

## (Intercept)       treat
##   6422.8390    224.6763

##                 2.5 %   97.5 %
## (Intercept)  5705.529 7140.149
## treat       -1562.856 2012.208
```

We received `Est: 224.68 and 95% CI: (-1562.86, 2012.21)`.

Now truncate the weights at the 1st and 99th percentiles. This can be done with the trunc=0.01 option in svyglm.

```
weightmodel_m2<-ipwpoint(exposure= treat, family = "binomial", link ="logit",
                  denominator= ~ age + educ + black + hispan + married + nodegree + re74 + re75,
                  data=mydf, trunc=.01)
```

Using IPTW with the truncated weights, let's find the estimate and 95% confidence interval for the average causal effect:

```
coef(msm_m2)
```

```
## (Intercept)       treat
##   6422.9362    486.9336
```

```
confint(msm_m2)
```

```
##                 2.5 %   97.5 %
## (Intercept)  5705.614 7140.258
## treat       -1093.765 2067.632
```

The estimate is 486.93 and 95% CI is (-1093.77, 2067.63).