BLP Assignment

Kwaku Atuahene

10-08-2024

Propsensity Score Matching

For this problem, you will analyze the data from:

Chirstopher Blattman and J Annan. 2010. “The consequences of child soldiering” *Review of Economics and Statistics* 92 (4):882-898

The data are from a panel survey of male youth in war-afflicted regions of Uganda. The authors want to estimate the impact of forced military service on various outcomes. They focus on Uganda because there were a significant number of abductions of young men into Lord’s Resistance Army.

Blattman and Annan describe the abductions as follows:

Abduction was large-scale and seemingly indiscriminate; 60,000 to 80,000 youth are estimated to have been abducted and more than a quarter of males currently aged 14 to 30 in our study region were abducted for at least two weeks. Most were abducted after 1996 and from one of the Acholi districts of Gulu, Kitgum, and Pader.

Youth were typically taken by roving groups of 10 to 20 rebels during night raids on rural homes. Adolescent males appear to have been the most pliable, reliable and effective forced recruits, and so were disproportionately targeted by the LRA. Youth under age 11 and over 24 tended to be avoided and had a high probability of immediate release. Lengths of abduction ranged from a day to ten years, averaging 8.9 months in our sample. Youth who failed toe escape were trained as fighters and, after a few months, received a gun. Two thirds of abductees were forced to perpetrate a crime or violence. A third eventually became fighters and a fifth were forced to murder soldiers, civilians, or even family members in order to bind them to the group, to reduce their fear of killing, and to discourage disobedience.

In this problem we will look at the effect of abduction on *educ* (years of education). The *abd* variable is the treatment in this case. Note that *educ, distress, and logwage* are all outcomes/post-treatment variables.

Variables	Description
abd	abducted by the LRA (the treatment)
c_ach - c_pal	Location indicators (each abbreviation corresponds to a subdistrict; i.e. ach = Acholibur, etc.)
age	age in years
fthr_ed	father’s education (years)
mthr_ed	mother’s education (years)
orphan96	indicator if parent’s died before 1996
hh_fthr_frm	indicator if father is a farmer
hh_size96	household size in 1996
educ	years of education
distress	index of emotional distress (0-15)
logwage	log of average daily wage earned in last 4 weeks

1. Calculate the naive Average Treatment Effect (ATE) of abduction on education (educ), distress (distress), and wages (logwage). Do this by running three separate regressions.

```
library(readr)
blattman <- read_csv("blattman.csv")
```

```
## Rows: 741 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (18): abd, C_ach, C_akw, C_ata, C_kma, C_oro, C_pad, C_paj, C_pal, age, ...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
library(modelsummary)
```

```
## ‘modelsummary‘ 2.0.0 now uses ‘tinytable‘ as its default table-drawing
##   backend. Learn more at: https://vincentarelbundock.github.io/tinytable/
##
## Revert to ‘kableExtra‘ for one session:
##
##   options(modelsummary_factory_default = ’kableExtra’)
##   options(modelsummary_factory_latex = ’kableExtra’)
##   options(modelsummary_factory_html = ’kableExtra’)
##
## Silence this message forever:
##
##   config_modelsummary(startup_message = FALSE)
```

```
reg1<- lm(educ ~ abd, data= blattman)
reg2<- lm(distress ~ abd, data= blattman)
reg3<- lm(log.wage~ abd, data= blattman)
modelsummary(list("Education"=reg1,"Distress"=reg2,"Log wage"=reg3),stars = TRUE, coef_rename = c("abd"=
```

2. Use a parametric model (Probit/Logit) to calculate the propensity scores for each person in the data to be abducted. Include whatever covariates or functions of covariates you think may be important.

```
logit1<-glm(abd~ hh_size96+hh_fthr_frm+orphan96+mthr_ed+fthr_ed+age+C_ach+C_akw+C_ata+C_kma+C_oro+C_pad+
blattman$pscore<-predict(logit1,type="response")
#hist(blattman$pscore)
#hist(blattman$pscore)
```

3. Use optimal match over the whole data set to estimate the ATE using propensity score matching. Do this for all three dependent variables.

```
library(MatchIt)
v2<- abd~ hh_size96+hh_fthr_frm+orphan96+mthr_ed+fthr_ed+age+C_ach+C_akw+C_ata+C_kma+C_oro+C_pad+C_paj+C
m.nn<-matchit(v2 , data = blattman, ratio=1, method ="optimal")
```

```
## Warning: Fewer control units than treated units; not all treated units will get
## a match.
```

	Education	Distress	Log wage
(Intercept)	7.416***	3.759***	4.205***
	(0.172)	(0.144)	(0.219)
Abduction	−0.595**	0.593**	0.272
	(0.218)	(0.182)	(0.277)
Num.Obs.	741	741	741
R2	0.010	0.014	0.001
R2 Adj.	0.009	0.013	0.000
AIC	3672.4	3406.0	4028.4
BIC	3686.2	3419.8	4042.2
Log.Lik.	−1833.183	−1699.990	−2011.182
RMSE	2.87	2.40	3.65

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

	Education	Distress	Log Wages
(Intercept)	7.416***	3.759***	4.205***
	(0.173)	(0.138)	(0.221)
Abduction	−0.563*	0.530**	−0.060
	(0.245)	(0.195)	(0.312)
Num.Obs.	558	558	558
R2	0.009	0.013	0.000
R2 Adj.	0.008	0.011	−0.002
AIC	2774.4	2521.2	3043.6
BIC	2787.4	2534.2	3056.6
Log.Lik.	−1384.205	−1257.595	−1518.796
RMSE	2.89	2.30	3.68

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

```r
#The New Matched Dataset
nn.match<-match.data(m.nn)

reg4 <- lm(educ ~ abd, data = nn.match)
reg5 <- lm(distress ~ abd, data = nn.match)
reg6 <- lm(log.wage ~ abd, data = nn.match)

modelsummary(list("Education"=reg4,"Distress"=reg5,"Log Wages"=reg6),stars = TRUE, coef_rename = c("abd
```

4. Use the cobalt package to make a “Love plot”. You can find information of the cobalt package here

```r
library(cobalt)
```

```
##  cobalt (Version 4.5.5, Build Date: 2024-04-02)
```

```
##
## Attaching package: 'cobalt'

## The following object is masked from 'package:MatchIt':
##
##     lalonde
```

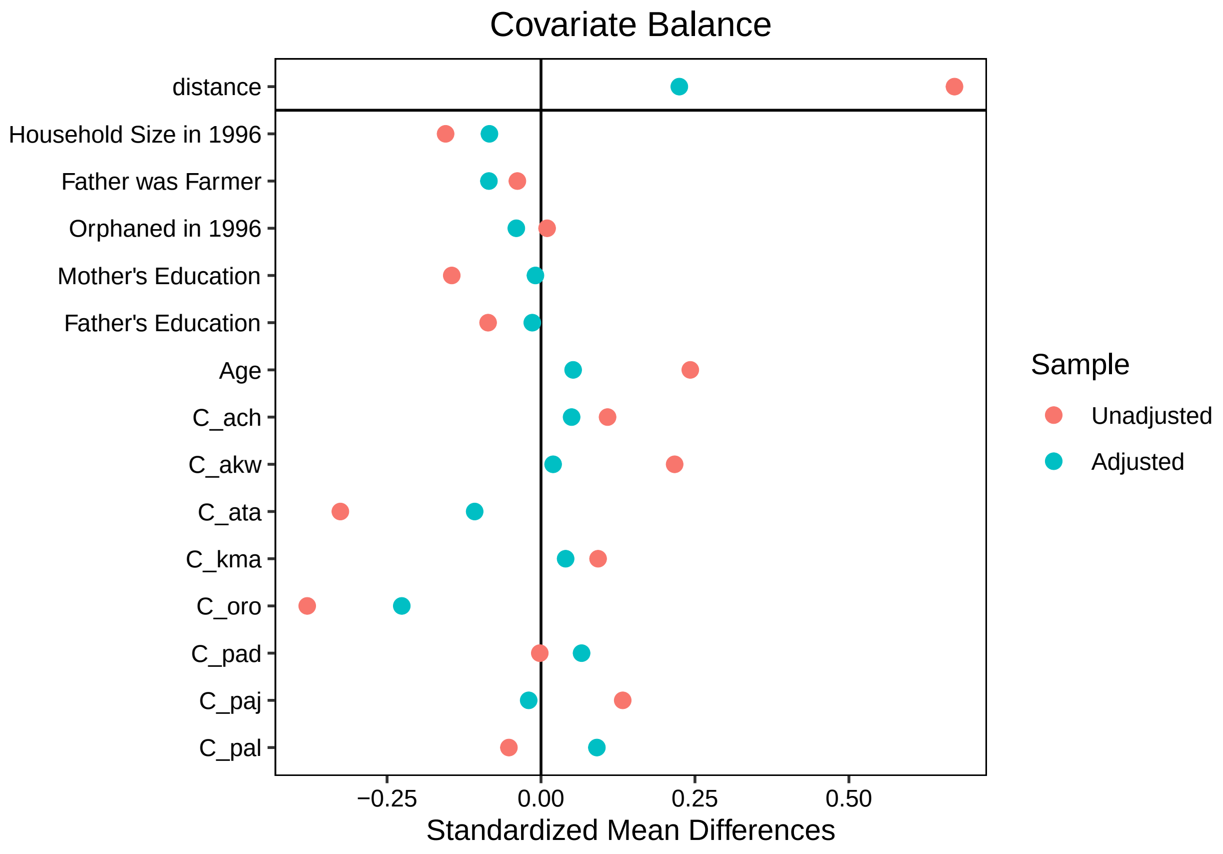
```
b1<-bal.tab(v2,data=blattman,int = TRUE)
```

```
## Note: `s.d.denom` not specified; assuming "pooled".
```

```
v1<-var.names(b1, type = "vec", minimal = TRUE)
v1["hh_size96"]<-"Household Size in 1996"
v1["hh_fthr_frm"]<-"Father was Farmer"
v1["orphan96"]<-"Orphaned in 1996"
v1["mthr_ed"]<-"Mother's Education"
v1["fthr_ed"]<-"Father's Education"
v1["age"]<-"Age"

#love.plot(b1,std) #use v1 for your variable names

love.plot(m.nn, binary ="std", var.names = v1 )
```



```
#library(RItools)
#xBalance(v2, data = blattman, report=c("chisquare.test"))
#xBalance(v2, data = nn.match, report=c("chisquare.test"))
```

Problem Set: BLP Methodology In this problem you will perform demand estimation using market level data. Run the following code in R

```
#install.packages("BLPestimatoR")
library(BLPestimatoR)
data(productData_cereal)
```

A table of market shares, prices, and characteristics of the top-selling brands of cereal in 1992 across several markets is now available in your environment. The data are aggregated from household-level scanner data (collected at supermarket checkout counters). We observe the following variables

price = price paid for the cereal const = just a column of 1's that you can ignore. sugar = how much sugar is in the cereal mushy = how mushy the cereal becomes with milk. share = market share of the cereal in that particular market. This number is between 0 and 1. cdid = tells you which market you are in. product_id = tells you which cereal is captured. IV1-IV20 = 20 constructed instrumental variables.

1. Find the market share of the outside good in every market. That is, sum all of the shares across all of the cereals for each market. You will notice that this number is less than 1. The market share of the outside option is equal to 1 - total cereal market share in each market. (Hint: you can use the aggregate to sum up the cereal shares by market)

```
# We can use the function ave(variable, grouping variable, FUN = function(x) 1-sum(x))

productData_cereal$outside_share <- ave(productData_cereal$share,productData_cereal$cdid,FUN = function
# this will create your new depedent variable (i.e. log(sj)-log(so))
productData_cereal$y <- log(productData_cereal$share)-log(productData_cereal$outside_share)
```

2. Estimate the share regression using sugar, mushy and price as explantory variables using OLS

```
## In this section we are just going to run OLS on the linear demand curve
blp.reg.1<-lm(y~ price+sugar+mushy, data=productData_cereal)
library(fixest)
blp.reg.2 <-feols(y~price+sugar+mushy|cdid, data=productData_cereal) # Include market fixed effects
```

3. 2SLS: Use the instrumental variables IV1 - IV10 to instrument for price

```
library(fixest)
blp.reg.3 <- feols(y~sugar+mushy|cdid| price ~ IV1+IV2+IV3+IV4+IV5+IV6+IV7+IV8+IV9+IV10, data=productDa
modelsummary::modelsummary(list("OLS"=blp.reg.1,"Fixed Effect"=blp.reg.2,"IV"=blp.reg.3),stars = TRUE,c
```

4. 2SLS: perform the first stage F-stat test to judge the strength of your instruments and the second stage sargent test to see if these instruments are independent of the error term.

```
# Hint use the summary function with object bls.reg.3
summary(blp.reg.3)
```

	OLS	Fixed Effect	IV
(Intercept)	−2.993***		
	(0.112)		
price	−10.120***	−9.319***	−9.236***
	(0.880)	(0.760)	(0.792)
sugar	0.046***	0.045***	0.045***
	(0.004)	(0.005)	(0.005)
mushy	0.052	0.056	0.057
	(0.052)	(0.043)	(0.043)
Num.Obs.	2256	2256	2256
R2	0.079	0.210	0.210
R2 Adj.	0.078	0.175	0.175
R2 Within		0.081	0.081
R2 Within Adj.		0.080	0.080
AIC	7068.7	6906.3	6906.3
BIC	7097.3	7461.3	7461.3
Log.Lik.	−3529.339		
RMSE	1.16	1.07	1.07
Std.Errors		by: cdid	by: cdid
FE: cdid		X	X

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

```
## TSLS estimation - Dep. Var.: y
##                   Endo.    : price
##                   Instr.   : IV1, IV2, IV3, IV4, IV5, IV6, IV7, IV8, IV9, IV10
## Second stage: Dep. Var.: y
## Observations: 2,256
## Fixed-effects: cdid: 94
## Standard-errors: Clustered (cdid)
##            Estimate Std. Error   t value   Pr(>|t|)
## fit_price -9.235583   0.791762 -11.66459  < 2.2e-16 ***
## sugar      0.044858   0.005051   8.88123 4.7572e-14 ***
## mushy      0.056966   0.043007   1.32458 1.8856e-01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.07114     Adj. R2: 0.175154
##                 Within R2: 0.080861
## F-test (1st stage), price: stat = 7,401.1    , p < 2.2e-16 , on 10 and 2,243 DoF.
##                Wu-Hausman: stat =     0.31764, p = 0.573088, on 1 and 2,158 DoF.
##                    Sargan: stat =    23.3    , p = 0.005464, on 9 DoF.
```

5. 2SLS: can you use a smaller set of instruments to get a better result? If so, then what instruments did you include? Report your results including the first stage F-stats and the overidentification test.

Hint: You will need to run the first stage regression and identify which instruments are significant. Try using only the significant instruments.

```r
#First Stage regression to identify which instruments are significant.
summary(feols(price~sugar+mushy+ IV1+IV2+IV3+IV4+IV5+IV6+IV7+IV8+IV9+IV10|cdid, data=productData_cereal
```

```
## OLS estimation, Dep. Var.: price
## Observations: 2,256
## Fixed-effects: cdid: 94
## Standard-errors: Clustered (cdid)
##           Estimate Std. Error    t value  Pr(>|t|)
## sugar   0.00007886   0.000133   0.592729 0.5548012
## mushy   0.00427523   0.001557   2.745063 0.0072618 **
## IV1    -0.00199834   0.003945  -0.506520 0.6136906
## IV2     0.01019922   0.017096   0.596570 0.5522437
## IV3     0.00000258   0.000067   0.038422 0.9694336
## IV4    -0.00122058   0.001007  -1.212337 0.2284543
## IV5     0.14367312   0.011649  12.333598 < 2.2e-16 ***
## IV6   -13.56726280   0.426325 -31.823737 < 2.2e-16 ***
## IV7     0.00027946   0.000147   1.904907 0.0598822 .
## IV8     0.00438086   0.001589   2.757308 0.0070154 **
## IV9     0.68108857   0.023737  28.692560 < 2.2e-16 ***
## IV10    0.00260159   0.004423   0.588139 0.5578648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.004673     Adj. R2: 0.972816
##                  Within R2: 0.973277
```

```r
#First Stage regression shows only IV5,IV6,IV7,IV8 & IV9 are significant.
#Rerun the regression with only IV5,IV6,IV7,IV8 & IV9 which are significant.
blp.reg.4 <- feols(y~sugar+mushy|cdid| price ~ IV5+IV6+IV7+IV8+IV9, data=productData_cereal)
summary(blp.reg.4)
```

```
## TSLS estimation - Dep. Var.: y
##                   Endo.    : price
##                   Instr.   : IV5, IV6, IV7, IV8, IV9
## Second stage: Dep. Var.: y
## Observations: 2,256
## Fixed-effects: cdid: 94
## Standard-errors: Clustered (cdid)
##            Estimate Std. Error   t value   Pr(>|t|)
## fit_price -9.224012   0.792248 -11.64283  < 2.2e-16 ***
## sugar      0.044842   0.005054   8.87271 4.9586e-14 ***
## mushy      0.057031   0.043000   1.32630 1.8799e-01
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## RMSE: 1.07114     Adj. R2: 0.175153
##                 Within R2: 0.08086
## F-test (1st stage), price: stat = 14,702.0     , p < 2.2e-16 , on 5 and 2,248 DoF.
##                Wu-Hausman: stat =      0.408664, p = 0.522716, on 1 and 2,158 DoF.
##                    Sargan: stat =     15.5     , p = 0.003743, on 4 DoF.
```

```
#The regression fails the Sagan Test.
productData_cereal$resid<-blp.reg.4$residuals
summary(feols(resid~sugar+mushy+ IV5+IV6+IV7+IV8+IV9|cdid, data=productData_cereal))
```

```
## OLS estimation, Dep. Var.: resid
## Observations: 2,256
## Fixed-effects: cdid: 94
## Standard-errors: Clustered (cdid)
##        Estimate Std. Error   t value   Pr(>|t|)
## sugar  0.051245   0.016942  3.024822 0.00321624 **
## mushy -0.289070   0.133357 -2.167645 0.03273925 *
## IV5   -1.208342   0.810688 -1.490514 0.13947302
## IV6   -9.734625  23.486811 -0.414472 0.67948296
## IV7    0.076124   0.021328  3.569186 0.00056875 ***
## IV8   -0.357908   0.165119 -2.167581 0.03274425 *
## IV9    0.548980   1.282684  0.427993 0.66964588
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## RMSE: 1.06745     Adj. R2: -0.039207
##                 Within R2:  0.006877
```

```
#Second Stage regression shows only IV7,IV8 are correlated with the errors.
#Rerun the regression with only IV5,IV6 & IV9.
blp.reg.5 <- feols(y~sugar+mushy|cdid| price ~ IV5+IV6+IV9, data=productData_cereal)
summary(blp.reg.5)
```

```
## TSLS estimation - Dep. Var.: y
##                   Endo.    : price
##                   Instr.   : IV5, IV6, IV9
## Second stage: Dep. Var.: y
## Observations: 2,256
## Fixed-effects: cdid: 94
## Standard-errors: Clustered (cdid)
##            Estimate Std. Error   t value   Pr(>|t|)
## fit_price -9.231385   0.791146 -11.66837  < 2.2e-16 ***
## sugar      0.044852   0.005053   8.87672 4.8626e-14 ***
## mushy      0.056989   0.042995   1.32549 1.8826e-01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.07114     Adj. R2: 0.175154
##                 Within R2: 0.080861
## F-test (1st stage), price: stat = 23,883.8     , p < 2.2e-16 , on 3 and 2,250 DoF.
##                Wu-Hausman: stat =      0.338346, p = 0.560846, on 1 and 2,158 DoF.
##                    Sargan: stat =      2.14665 , p = 0.341869, on 2 DoF.
```