Practice

Diwen

2023-12-12

Table of contents

1 Introduction

This quarto file is used to practice my formatting skills as much as possible.

Thus, the *content* may **be** displayed in an ***esoteric*** pattern like this.

Chapter 2

Figure 3.1

2 Stata and R

2.1 Robust Standard Errors

In probit estimation, Stata and R will produce difference robust standard errors.

```
use "E:/Curriculum001/UG3/EC338/Assignment2/Smoking.dta"
probit smoker i.smkban age i.hsdrop i.hsgrad i.colsome i.colgrad i.black i.hispanic if femal
```

```
Iteration 0:   log pseudolikelihood = -2489.3259
Iteration 1:   log pseudolikelihood =  -2346.727
Iteration 2:   log pseudolikelihood = -2345.1465
Iteration 3:   log pseudolikelihood = -2345.1464

Probit regression                                     Number of obs =  4,363
                                                      Wald chi2(8)  = 263.16
                                                      Prob > chi2   = 0.0000
Log pseudolikelihood = -2345.1464                     Pseudo R2     = 0.0579

------------------------------------------------------------------------------
             |               Robust
      smoker | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    1.smkban |   -.197581   .0427786    -4.62   0.000    -.2814255   -.1137366
         age |  -.0032241   .0017416    -1.85   0.064    -.0066377    .0001894
    1.hsdrop |   1.095518   .0992209    11.04   0.000     .9010486    1.289987
    1.hsgrad |   .9249239   .0836137    11.06   0.000      .761044    1.088804
   1.colsome |   .6861135   .0851945     8.05   0.000     .5191354    .8530916
   1.colgrad |   .3216797   .0894715     3.60   0.000     .1463188    .4970405
     1.black |  -.0003162   .0840926    -0.00   0.997    -.1651348    .1645023
  1.hispanic |  -.2377863   .0691798    -3.44   0.001    -.3733763   -.1021963
       _cons |  -1.067899   .1068883    -9.99   0.000    -1.277396    -.858402
------------------------------------------------------------------------------
```

```r
reg91 <- glm(smoker ~ as.factor(smkban) + age + hsdrop + hsgrad + colsome+ colgrad +
               black + hispanic,data = data01,family = binomial(link="probit"),
             subset = female==0)
coeftest(reg91, vcov = vcovHC(reg91, type="HC1"))
```

```
z test of coefficients:

                     Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)        -1.06789853  0.10634581 -10.0418 < 2.2e-16 ***
as.factor(smkban)1 -0.19758088  0.04287431  -4.6084 4.058e-06 ***
age                -0.00322415  0.00172811  -1.8657  0.062083 .
hsdrop              1.09551672  0.09983792  10.9730 < 2.2e-16 ***
hsgrad              0.92492334  0.08379179  11.0384 < 2.2e-16 ***
colsome             0.68611331  0.08541225   8.0330 9.515e-16 ***
colgrad             0.32167916  0.08968817   3.5866  0.000335 ***
black              -0.00031533  0.08447431  -0.0037  0.997022
hispanic           -0.23778551  0.07042242  -3.3766  0.000734 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.2 Plot

A arbitary graph draw from EC338 Assignment 1

```r
s3t1f[[1]]
s3t1f[[2]]
```

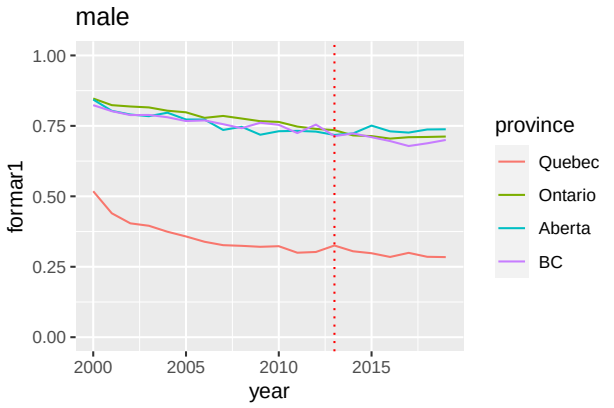


Figure 2.1: Arbitary

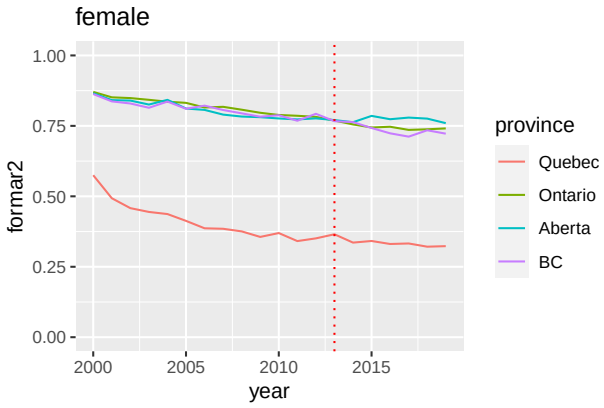


Figure 2.2: Arbitary

3 Mathtype

$$
\begin{aligned}
for D_i &= \{0,1\} \\
Var(D_i|x_i = x_j) &= \Pr(D_i|x_i = x_j)(1 - \Pr(D_i|x_i = x_j)) \\
\Sigma_{j=1}^{m} Var(D_i|x_i = x_j)\Pr(D_i|x_i = x_j) &= \Pr(D_i|x_i = x_j)(1 - \Pr(D_i|x_i = x_j))\Pr(x_i = x_j) \\
for x_i &= \{1, 2...m\}
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_{j=1}^{m} Var(D_i|x_i = x_j)\Pr(D_i|x_i = x_j) &= \mathbb{E}[Var(D_i|x_i = x_j)] \\
&= \mathbb{E}[\Pr(D_i|x_i = x_j)(1 - \Pr(D_i|x_i = x_j))] \\
&= \mathbb{E}[e(x_i) - (1 - e(x_i))]
\end{aligned}
$$

3.1 Proofreader

Austen-Smith, D. & Banks, J.S. 2005, Positive political theory II: strategy and structure, University of Michigan Press, Ann Arbor, [Mich.].

Example 6.4 Let $N = \{1, 2, 3\}$, $\delta_i = \delta$ all i and $q = 2$. By Theorem 6.2, any individual $i \in N$ recognized in the first period proposes an allocation giving a strictly positive amount δV_j to exactly one other committee member $j \neq i$ and nothing to the remaining individual. Consequently, the following three equations must hold in any stationary equilibrium:

$$\begin{aligned} V_1 &= p_1[1 - r_{12}\delta V_2 - r_{13}\delta V_3] + \delta V_1[p_2 r_{21} + p_3 r_{31}]; \\ V_2 &= p_2[1 - r_{21}\delta V_1 - r_{23}\delta V_3] + \delta V_2[p_1 r_{12} + p_3 r_{32}]; \\ V_3 &= p_3[1 - r_{31}\delta V_1 - r_{32}\delta V_2] + \delta V_3[p_1 r_{13} + p_2 r_{23}]. \end{aligned}$$

Noting $r_{21} = 1 - r_{12}$ etc, this system can be written in matrix notation,

$$H(V)r = p \qquad (*)$$

where

$$H(V) = \begin{bmatrix} p_1\delta(V_2 - V_3) & p_2\delta V_1 & -p_3\delta V_1 \\ -p_1\delta V_2 & p_2\delta(V_3 - V_1) & p_3\delta V_2 \\ p_1\delta V_3 & -p_2\delta V_3 & p_3\delta(V_1 - V_2) \end{bmatrix}$$

$V_1 = p_1[1 - r_{12}\delta_2 V_2 - r_{13}\delta_3 V_3] + \delta V_1[p_2 r_{21} + p_3 r_{31}]$

$V_1 = p_1 - p_1 r_1 \delta V_1 - p_1(1 - r_1)\delta V_3 + \delta V_1 p_2(1 - r_2) + p_3 r_{31}\delta V_1$

$-p_1 = -r_1(p_1\delta V_1 - p_1\delta V_3) - p_1\delta V_3 + \delta V_1 p_2 - r_2\delta V_1 p_2 + r_3\delta V_1 p_3 - V_1$

$p_1 = r_1 p_1 \delta(V_1 - V_3) + r_2\delta V_1 p_2 - r_3\delta V_1 p_3 + V_1 + p_1\delta V_3 - \delta V_1 p_2$

Figure 3.1: Linear

4 Seminar 4:Clustering Standard Errors and Models with Binary Outcomes

4.1 Application 1: Clustering Standard Errors

4.1.1 General setup

Our first application will use two simulated datasets on student achievement with the following variables:

- y is the outcome variable: a measure of student achievement
- i is an individual-level identifier
- *classroom* is a classroom-level identifier
- *aircon* is a binary variable equal to one if the classroom has air conditioning

The goal in this application is to understand that, if we think that our data is characterised by group-level shocks, we will need to adjust the standard errors. In particular, basic standard errors are calculated on the presumption of iid error terms. However, in many applications we may think of our data being organised in a number of groups (called clusters), and the unobservables may be correlated within these groups.

Why is that?

Think about the example from this application. Suppose you want to estimate the effect that air conditioning has on student achievement: $y = \alpha + \beta aircon_i + \eta_{(c)i} + \epsilon_i$. We have two unobservables here: a classroom-level shock (c)i and an individual shock ϵ_i . You can think of the regression unobservables as being denoted by $u_i = \eta_{(c)i} + \epsilon_i$. The key issue in this example

(and with clustering in general) is that, since two students in the same class will be exposed to the same classroom-level shock, our error term ui will no longer satisfy the iid assumption. If we do not adjust our standard errors accordingly, inference on the parameters in our model can go badly wrong.

$$\Omega = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

Clustering of Standard Error In the first part, we will look at the dataset aircon1 where, by construction, the **iid assumption of our error terms holds**.

```
*** import dataset 1
import excel using "aircon1", first

*** estimating the effect of air conditioning on achievement

* start with the simplest model, where assume iid errors
reg y aircon
```

```
(4 vars, 5,000 obs)

      Source |       SS           df       MS      Number of obs   =     5,000
-------------+----------------------------------   F(1, 4998)      =   1298.54
       Model |  1276.01233         1  1276.01233   Prob > F        =    0.0000
    Residual |  4911.29211     4,998  .982651483   R-squared       =    0.2062
-------------+----------------------------------   Adj R-squared   =    0.2061
       Total |  6187.30445     4,999  1.23770843   Root MSE        =    .99129

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      aircon |   1.010351   .0280379    36.04   0.000     .9553849    1.065318
       _cons |   4.989557   .0198258   251.67   0.000      4.95069    5.028424
------------------------------------------------------------------------------
```

```r
library(readxl)
data4 <- read_excel("aircon1.xls")
summary(lm(y ~ aircon, data = data4))
```

5 Appendix

[**definition**] Variance equation: A is a non-stochastic matrix, y is a stochastic matrix

$$
\begin{aligned}
Var(Ay) &= A * Var(y) * A' \\
Var(\hat{\beta}) &= Var((X'X)^{-1}X'y) \\
&= (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}
\end{aligned}
$$

In iid, where all observations are independent to each other, the variance-covariance looks like.

$$
\Omega = \begin{bmatrix}
\sigma^2 & 0 & 0 & \cdots & 0 \\
0 & \sigma^2 & 0 & \cdots & 0 \\
0 & 0 & \sigma^2 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & \sigma^2
\end{bmatrix} = \sigma^2 I
$$

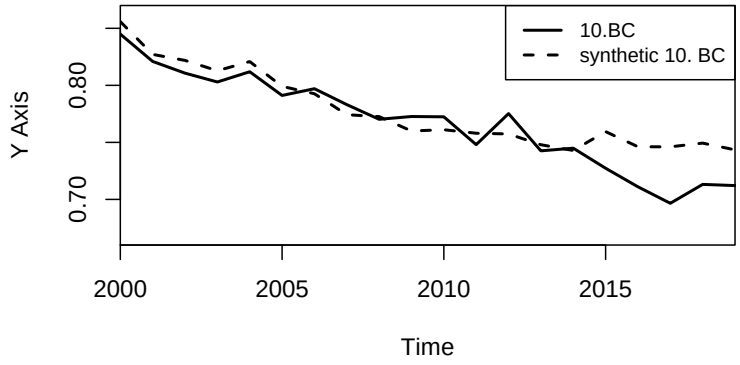
Synthetic control with R

```
df341<-read_dta('sectionC_task4_collapseddata.dta')
df34 <- df341 %>%
  group_by(province, year) %>%
  summarise(across(starts_with("edu"), ~mean(.x, na.rm = TRUE)),
            formar = mean(formar, na.rm = TRUE),
            cohab = mean(cohab, na.rm = TRUE))
df34$name <- ""
proname <- c("Newfoundland & Labrador", "Prince Edward Island", "Nova Scotia", "New Brunswic
for (i in 1:10){
df34$name <- ifelse(df34$province == i, proname[i],df34$name)
}
```

```r
df34 <- as.data.frame(df34)
df34$province <- as.numeric(df34$province)
# Prepare the data
t3synp1 <- dataprep(
   foo = df34,
   predictors    = c("edu12", "edu13", "edu21", "edu22", "edu23", "edu31", "edu32", "edu33")
   predictors.op = "mean",
   dependent     = "formar",
   unit.variable = "province",
   time.variable = "year",
   unit.names.variable = "name",
   treatment.identifier  = 10,
   controls.identifier   = c(1:9),
   time.predictors.prior = c(2000:2012),
   time.optimize.ssr     = 2000:2013,
   time.plot             = 2000:2019
)

t3syn1 <- synth(t3synp1)
```

```
X1, X0, Z1, Z0 all come directly from dataprep object.


****************
 searching for synthetic control unit


****************
****************
****************

MSPE (LOSS V): 9.723109e-05

solution.v:
 6.1906e-06 0.01370176 3.0543e-06 0.04109318 0.3385545 0.298815 0.3066765 0.001149889

solution.w:
 4.5251e-06 0.004146942 0.00410196 0.005293371 0.0009749935 0.2324437 0.04271549 0.003839381 0
```

```
t3syng1 <- path.plot(dataprep.res = t3synp1,synth.res = t3syn1,Ylim = c(0.66,0.87), Legend =
```

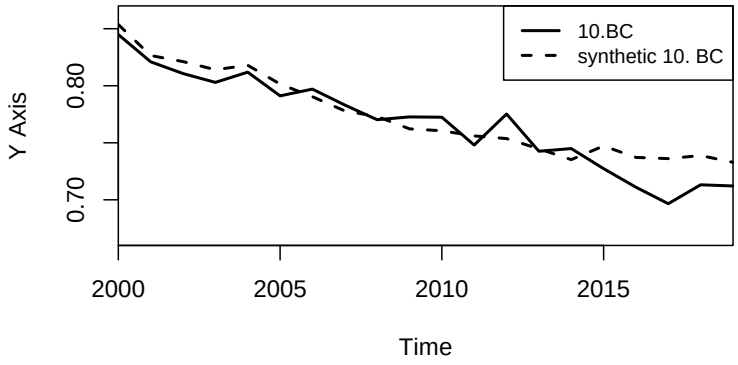



```
t3synp2 <- dataprep(
   foo = df34,
   predictors    = c("cohab","edu12", "edu13", "edu21", "edu22", "edu23", "edu31", "edu32",
   predictors.op = "mean",
   special.predictors = list(
      list("formar", 2009, "mean")),
   dependent     = "formar",
   unit.variable = "province",
   time.variable = "year",
   unit.names.variable = "name",
   treatment.identifier  = 10,
   controls.identifier   = c(1:9),
   time.predictors.prior = c(2000:2012),
   time.optimize.ssr     = 2000:2013,
   time.plot             = 2000:2019
)
t3syn2 <- synth(t3synp2)
```

X1, X0, Z1, Z0 all come directly from dataprep object.

```
****************
 searching for synthetic control unit


****************
****************
****************

MSPE (LOSS V): 9.616666e-05

solution.v:
 0.200451 0.03583939 0.04740207 0.08822844 0.01968502 0.1645045 0.1040342 0.1253762 0.1179669 0

solution.w:
 6.826e-07 2.5922e-06 0.07121743 1.4201e-06 3.981e-06 0.4011764 3.69393e-05 3.9048e-06 0.52755
```

```
t3syng2 <- path.plot(dataprep.res = t3synp2,synth.res = t3syn2, Ylim = c(0.66,0.87), Legend
```



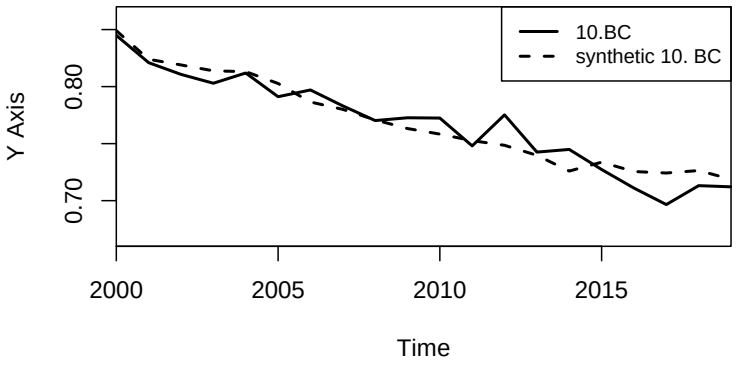

```
t3synp3 <- dataprep(
   foo = df34,
   predictors    = c("cohab", "edu12", "edu13", "edu21", "edu22", "edu23", "edu31", "edu32",
   predictors.op = "mean",
   special.predictors = list(
```

```
      list("formar", 2007, "mean"),
      list("formar", 2009, "mean"),
      list("formar", 2011, "mean")),
   dependent     = "formar",
   unit.variable = "province",
   time.variable = "year",
   unit.names.variable = "name",
   treatment.identifier  = 10,
   controls.identifier   = c(1:9),
   time.predictors.prior = c(2000:2012),
   time.optimize.ssr     = 2000:2011,
   time.plot             = 2000:2019
)
t3syn3 <- synth(t3synp3)
```

```
X1, X0, Z1, Z0 all come directly from dataprep object.


****************
 searching for synthetic control unit


****************
****************
****************

MSPE (LOSS V): 6.475819e-05

solution.v:
 0.1823591 0.1522035 0.01560817 0.129904 0.02137578 0.06901198 0.06846723 0.06128481 0.0979448
```

```
solution.w:
 7.232e-07 1.1139e-06 0.1728935 5.951e-07 1.5768e-06 0.5146288 3.97503e-05 1.4187e-06 0.312432
```

```
t3syng3 <- path.plot(dataprep.res = t3synp3,synth.res = t3syn3, Ylim = c(0.66,0.87), Legend
```


10.BC
synthetic 10. BC
Y Axis
0.80
0.70
2000
2005
2010
2015
Time

References