Tarea_3_Almonacid_Gonzalez

December 10, 2022

1 Parte 1: Experimentos

Deben conceptualizar un experimento con el objetivo de estudiar posibles incentivos o estrategias para incrementar la asistencia a clases en estudiantes universitarios de la UdeC. El outcome del tratamiento es la proporcion promedio de estudiantes que asisten a clases. Todos los elementos del experimento deben ser definidos, respondiendo a las siguientes preguntas:

1.1 Pregunta 1

Asumiendo la existencia de recursos disponibles e implementacion a nivel de estudiante, sugiera un tratamiento que pueda ser testeado a traves de un experimento aleatorizado controlado. Sea especifico en cuanto a los detalles del tratamiento (costos, materiales, duracion, etcetera).

Desde nuestro punto de vista, gran parte de la pérdida de interés en ir a clases se debe a la pérdida de sincronía, es decir, no saber cual es el estado actual del ramo, y que es lo que se verá en clase. Es cierto que existen las herramientas usuales, como el syllabus y el programa, pero carece profundidad y falla en su uso. Por esto se define el siguiente tratamiento:

Se envía al correo del estudiante un email informativo, sobre el estado actual del curso, fechas de próximas evaluaciones, temas actualmente revisados, y temas futuros a revisar, que sirva como una introducción al contenido de la próxima semana, y permita facilitar la reincorporación de alumnos que ya han faltado a clases

> Nota: **Si el experimento se evalúa a nivel de estudiante** es difícil diseñar buenos experimentos para tratar a nivel de estudiante, puesto que cualquier información que se hace llegar a una parte importante de los estudiantes inevitablemente llegará a la otra por otras vías de comunicación.

Como los requerimientos necesarios están todos implementados actualmente en la Universidad, el único costo, se asocia 1 o 2 horas extra semanales que el profesor de la asignatura dedicará a la redacción del correo y creacion del contenido asociado. Como cada grupo debe ser independiente se requiere una duracion mínima de 1 semestre por grupo, es decir el experimento tendría una duracion de almenos 1 año y medio.

1.2 Pregunta 2

Defina los grupos de tratamiento y control para implementar su experimento. Describa en detalle el mecanismo de asignacion aleatorio que permite la comparacion entre grupos.

Todo los grupos deben ser definidos desde cursos obligatorios en la respectiva carrera del alumno y debe contener almenos 2 examenes parciales tipo certamen. De esta forma se elimina la variabilidad

respecto a la metodología de evaluación que se está usando y los resultados pueden aplicarse a la mayoria de ramos en la Unversidad.

Para los grupos de control, se eligen alumnos aleatoriamente que pertenezcan al mismo curso/semestre.

Los grupos de tratamiento se eligen de la misma forma, pero deben pertenecer al mismo curso y a distinto semestre que los grupos de control, para minimizar las interacciones entre alumnos compartiendo información.

> Nota: En nuestra experiencia cuando las asistencias se miden de forma individual, es decir, con firma individual, la asistencia de alumnos tiende a ser mas alta, por esto quizas sea necesario otro grupo de control con conteo simple para medir el efecto de pasar lista sobre la asistencia.

1.3 Pregunta 3

Que metodo considera el mas apropiado para la estimacion del efecto promedio? (pre-test, pre-post test, Salomon 4 group). Justifique su respuesta en base a las ventajas y desventajas de cada metodo.

Se considera el método pre-post test, puesto que solo es necesario medir el efecto del tratamiento sobre los distintos grupos y no se requiere observar el comportamiento antes de incluir el tratamiento.

1.4 Pregunta 4

Ahora suponga que no es posible implementar un experimento a nivel de estudiante, sino a nivel de clase. Como ajustaria los elementos de su experimento para poder ser implementado a nivel de cluster? Sea especifico respecto tanto del tratamiento como del metodo de asignacion aleatorio y potencial comparacion entre grupos de tratamiento y control.

Ahora los grupos corresponderían a alumnos en grupos completos, solo comparables si son el mismo ramo, en distinto semestre, así se puede usar conteo total o por alumnos y no es necesario pasar lista individual.

Los cursos ramo/semestre son elegidos de forma aleatoria, los grupos de control y tratamiento deben permanecer al mismo periodo (semestre par o semestre impar).

1.5 Pregunta 5

Suponga que en vez de un experimento, se planifica que sea un programa implementado a nivel de toda la Universidad. Como ajustaria los elementos descritos anteriormente para poder comparar el efecto de la intervencion.

Si es un programa implementado a nivel de universidad los grupos se implementan por ramo en todos los ramos de la universidad, de esta forma se pueden comparar los resultados por año.

2 Parte 2 - Estimación de efectos promedio de tratamiento

2.1 Pregunta 6

A partir de sus respuestas en Parte 1, genere data para 40 grupos (considere cada grupo como una clase) con 50 estudiantes cada uno (asuma que los estudiantes son asignados aleatoriamente a cada clase). Cada estudiante debe tener data de asistencia en un periodo, generando una variable binaria aleatoria tal que la asistencia promedio a traves de todos los grupos es de 80%.

[120]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf
import scipy
from scipy.linalg import eigh, cholesky
from scipy.stats import norm,ttest_ind
import linearmodels.panel as lmp
from pylab import plot, show, axis, subplot, xlabel, ylabel, grid
%matplotlib inline
from numpy import random
import numpy as np

n_groups = 40
n_alumnos = 50
n_dias = 120 #Se preparan 120 días

nsize = n_groups*n_alumnos*n_dias


Xc = pd.DataFrame([i for i in range(nsize)],columns=["id"])

#Variable de periodo
Xc['p'] = 1
Xc.loc[0:nsize//2-1,'p'] = 0

#Variable de asistencia para el primer periodo

#Aigno los grupos
size = n_alumnos * n_dias//2
groups = [i for i in range(n_groups)]*2
for i,g in enumerate(groups):

    Xc.loc[(i)*(size):((i+1)*(size))-1,'cl'] = g

alumnos = n_alumnos*n_groups
size = 60
base = 2000 * 60
```

```
for i in range(alumnos):
    #Primer periodo
    Xc.loc[(i)*(size):((i+1)*(size))-1,"asistencia"] = np.random.binomial(1, 0.
 ↪8, size=size)
    Xc.loc[(i)*(size):((i+1)*(size))-1,"alumno"] = i
    Xc.loc[(i)*(size):((i+1)*(size))-1,"dia"] = [i for i in range(60)]

    #Segundo periodo
    Xc.loc[base + (i*size):base + (i+1)*(size)-1,"alumno"] = i
    Xc.loc[base + (i*size):base + (i+1)*(size)-1,"dia"] = [i for i in range(60)]

print(Xc)
```

```
            id  p    cl  asistencia  alumno   dia
0            0  0   0.0         1.0     0.0   0.0
1            1  0   0.0         0.0     0.0   1.0
2            2  0   0.0         1.0     0.0   2.0
3            3  0   0.0         1.0     0.0   3.0
4            4  0   0.0         1.0     0.0   4.0
…          … ..   …           …       …   …
239995  239995  1  39.0         NaN  1999.0  55.0
239996  239996  1  39.0         NaN  1999.0  56.0
239997  239997  1  39.0         NaN  1999.0  57.0
239998  239998  1  39.0         NaN  1999.0  58.0
239999  239999  1  39.0         NaN  1999.0  59.0

[240000 rows x 6 columns]
```

2.2 Pregunta 7

Genere un mecanismo de asignacion aleatorio a nivel de estudiante y muestre que en la data generada permite que ambos grupos (tratamiento y control) tienen una asistencia promedio comparable.

[132]:
```
from statsmodels.stats.power import TTestIndPower
from statsmodels.stats.contingency_tables import mcnemar

#Calcula
effect = 0.2
alpha = 0.005
power = 0.95
analysis = TTestIndPower()
result = analysis.solve_power(effect, power = power, nobs1= None, ratio = 1.0,␣
 ↪alpha = alpha)
print('Sample Size: %.3f' % round(result))

#Como se necesitan hay 2000 muestras en total y se pueden usar 993 para obtener␣
 ↪una significancia de 0.005 en el análisis
```

```
alumnos = n_alumnos*n_groups
size = 60
base = 2000 * 60
for i in range(alumnos):
    tratamiento = [random.randint(0,2)]*size
    Xc.loc[(i)*(size):((i+1)*(size))-1,"T"] = tratamiento
    Xc.loc[base + (i*size):base + (i+1)*(size)-1,"T"] = tratamiento


tratamiento = Xc[Xc["T"] == 1]
control =     Xc[Xc["T"] == 0]

control.describe()
tratamiento.describe()
```

Sample Size: 993.000

[132]:

	id	p	…	dia	T
count	116400.000000	116400.000000	…	116400.000000	116400.0
mean	120401.500000	0.500000	…	29.500000	1.0
std	69224.968535	0.500002	…	17.318177	0.0
min	0.000000	0.000000	…	0.000000	1.0
25%	60299.750000	0.000000	…	14.750000	1.0
50%	119999.500000	0.500000	…	29.500000	1.0
75%	180299.250000	1.000000	…	44.250000	1.0
max	239999.000000	1.000000	…	59.000000	1.0

[8 rows x 7 columns]

2.3 Pregunta 8

Genere un tratamiento que incrementa la participacion en el grupo de tratamiento en 10 puntos porcentuales. Ademas en la data posterior al experimento, asuma que la participacion promedio cayo a 75%. Estime el efecto promedio del tratamiento usando solo post-test.

[148]:
```
alumnos = n_alumnos*n_groups
size = 60

#Genero la asistencia despues del tratamiento con un 90%, solo para aquellos␣
 ↪con tratamiento
for i in range(alumnos):
    if Xc.loc[base + (i*size),"T"] == 1:
        Xc.loc[base + (i*size):base + (i+1)*(size)-1,"asistencia"] = np.random.
 ↪binomial(1, 0.90, size=size)
    if Xc.loc[base + (i*size),"T"] == 0:
```

```
        Xc.loc[base + (i*size):base + (i+1)*(size)-1,"asistencia"] = np.random.
 ↪binomial(1, 0.65, size=size)

#Como la mitad de la muestra subió a una media de 90%, esto implica que la otra␣
 ↪mitad tiene ahora una nueva distribucion binomial con un 65% de asistencia␣
 ↪promedio, por lo que
#Usando solo post-test el efecto promedio es de un 25%
print(Xc)
print(f"Post-test Efecto promedio: 25%")
```

```
            id  p    cl  asistencia  alumno   dia    T
0            0  0   0.0         1.0     0.0   0.0  1.0
1            1  0   0.0         0.0     0.0   1.0  1.0
2            2  0   0.0         1.0     0.0   2.0  1.0
3            3  0   0.0         1.0     0.0   3.0  1.0
4            4  0   0.0         1.0     0.0   4.0  1.0
…          … ..   …           …       …   …   …
239995  239995  1  39.0         1.0  1999.0  55.0  1.0
239996  239996  1  39.0         0.0  1999.0  56.0  1.0
239997  239997  1  39.0         1.0  1999.0  57.0  1.0
239998  239998  1  39.0         1.0  1999.0  58.0  1.0
239999  239999  1  39.0         1.0  1999.0  59.0  1.0

[240000 rows x 7 columns]
Post-test Efecto promedio: 25%
```

2.4 Pregunta 9

Estime el efecto promedio del tratamiento usando pre-post test con la data generada. Muestre que el efecto es equivalente usando ambos metodos.

[169]:
```
#Calculo entonces la media de los
#Comparo el grupo de tratamiento en el periodo 1 vs periodo 2

tratamiento_pre =  Xc.query("T == 1 and p == 0")
tratamiento_post = Xc.query("T == 1 and p == 1")
control_pre =      Xc.query("T == 0 and p == 0")
control_post =     Xc.query("T == 0 and p == 1")

mean_pre = np.average( tratamiento_pre["asistencia"])
mean_post = np.average(tratamiento_post["asistencia"])
mean_control_pre = np.average( control_pre["asistencia"])
mean_control_post = np.average(control_post["asistencia"])

print(f"Efecto promedio pre-post: {mean_post-mean_control_pre}")

#Comparamos los controles
```

```
print(f"Diferencia promedio pre-post: {mean_pre-mean_control_post}")

#Diferencia total
print(f"Diferencia Total:␣
 ↪{mean_post-mean_control_pre+mean_pre-mean_control_post}")
```

```
Efecto promedio pre-post: 0.0985927334601141
Diferencia promedio pre-post: 0.15280585860607887
Diferencia Total: 0.25139859206619297
```

2.5 Pregunta 10

Estime el efecto ajustando los errores estandar por cluster (la variable grupo representa cada clase). Cual es la diferencia entre ambas estimaciones? Explique porque es esperable (o no) encontrar diferencias entre ambos metodos.

[176]:
```
y=Xc['asistencia']
Xc['dd']= Xc['p']*Xc['T']
X=Xc[['p','T','dd']]
X = sm.add_constant(X)
model = sm.OLS(y, X)
results2 = model.fit()
results3 = model.fit(cov_type="cluster", cov_kwds={'groups': Xc['cl']})
print(results3.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             asistencia   R-squared:                       0.048
Model:                            OLS   Adj. R-squared:                  0.048
Method:                 Least Squares   F-statistic:                     5804.
Date:                Mon, 28 Nov 2022   Prob (F-statistic):           1.04e-51
Time:                        18:31:11   Log-Likelihood:            -1.2100e+05
No. Observations:              240000   AIC:                         2.420e+05
Df Residuals:                  239996   BIC:                         2.420e+05
Df Model:                           3
Covariance Type:              cluster
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.8006      0.002    443.751      0.000       0.797       0.804
p             -0.1524      0.002    -68.446      0.000      -0.157      -0.148
T              0.0004      0.002      0.150      0.880      -0.004       0.005
dd             0.2507      0.003     96.237      0.000       0.246       0.256
==============================================================================
Omnibus:                    40914.589   Durbin-Watson:                   1.995
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            66585.923
Skew:                          -1.290   Prob(JB):                         0.00
Kurtosis:                       2.957   Cond. No.                         6.77
```

```
==============================================================================

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
```

3 Parte 3: Experimentos naturales.

Usando la data **charls.csv**, responda las siguientes preguntas relativas a experimentos naturales.

3.1 Pregunta 11

Simule un experimento natural (e.g. intervencion de politica publica) tal que se reduce la proporcion de individuos con 3 hijos o mas que declaran beber alcohol en el tercer periodo a la mitad. Para ello, genere una variable de tratamiento (todos los individuos con mas de 2 hijos son parte de la intervencion), y una nueva variable llamada *sdrinlky*, talque es identica a *drinkly* en los periodos 1 y 2 , pero sustituya los valores aleatoriamente en el periodo 3 para generar el efecto esperado.

[240]:
```
charls = pd.read_csv('../../data/charls.csv')
charls.dropna(inplace=True)
charls.reset_index(drop=True, inplace=True)
charls.drop(indexes)
#Creo la variable de tratamiento
for i in range(len(charls)):
    if charls.loc[i,"child"] > 2:
        charls.loc[i,"treatement"] = 1
    else:
        charls.loc[i,"treatement"] = 0

drinkly = charls.query("child >= 3 and wave == 3")["drinkly"]
drinkly = [int(i) for i in drinkly if i in ["0","1"]]
mean = np.average(drinkly)

#Creo la variable drinkly
for i in range(len(charls)):
    if charls.loc[i,"wave"] == 3 and charls.loc[i,"treatement"] == 1:
        #Genero la variable binaria, con una distribucion binomial con la mitad␣
 ↪de la media, que drinkly original de la weave 3
        charls.loc[i,"sdrinkly"] = np.random.binomial(1, mean/2, size=1)
    else:
        try:
            charls.loc[i,"sdrinkly"] = int(charls.loc[i,"drinkly"])
        except:
            charls.loc[i,"sdrinkly"] = 0

means = np.average(charls["sdrinkly"].dropna(inplace=False))
means2 = np.average(charls.query("wave == 3 and treatement == 1")["sdrinkly"].
 ↪dropna(inplace=False))
```

```
means3 = np.average(charls.query("wave == 3 and treatement == 0")["sdrinkly"])
print(f"Media total wave 3: {means}")
print(f"Media total wave 3 con tratamiento: {means2}")
print(f"Media total wave 3 sin tratamiento: {means3}")
```

```
Media total wave 3: 0.31081016868614875
Media total wave 3 con tratamiento: 0.16306361951822113
Media total wave 3 sin tratamiento: 0.3700170357751278
```

3.2 Pregunta 12

Estime el efecto del tratamiento usando diferencias en diferencias, comparando entre los periodos 2 y 3.

[241]:
```
charls = charls[charls['wave'] > 1] #Filtro para el periodo 2 y 3
y = charls["sdrinkly"]
charls["dd"] = charls["treatement"]*charls["wave"]
X=charls[['wave','treatement','dd']]
X = sm.add_constant(X)
model = sm.OLS(y,X)
result = model.fit()
print(result.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                sdrinkly   R-squared:                       0.033
Model:                            OLS   Adj. R-squared:                  0.033
Method:                 Least Squares   F-statistic:                     146.3
Date:                Mon, 28 Nov 2022   Prob (F-statistic):           2.94e-93
Time:                        19:39:54   Log-Likelihood:                -8080.3
No. Observations:               12965   AIC:                         1.617e+04
Df Residuals:                   12961   BIC:                         1.620e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.3577      0.029     12.215      0.000       0.300       0.415
wave           0.0041      0.012      0.355      0.722      -0.019       0.027
treatement     0.2550      0.040      6.348      0.000       0.176       0.334
dd            -0.1540      0.016     -9.684      0.000      -0.185      -0.123
==============================================================================
Omnibus:                    12640.788   Durbin-Watson:                   1.797
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2196.154
Skew:                           0.809   Prob(JB):                         0.00
Kurtosis:                       1.798   Cond. No.                         39.6
==============================================================================
```

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

```
/tmp/ipykernel_56237/4229658461.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  charls["dd"] = charls["treatement"]*charls["wave"]
```

El efecto del tratamiento se estima en un 15%

3.3 Pregunta 13

Compare el efecto del tratamiento generando grupos pseudo-equivalentes, en particular entre individuos solo con 3 hijos (tratamiento) y 2 hijos (control).

[244]:
```
hijos2 =  charls.query("wave == 3 and child == 2")
hijos3 =  charls.query("wave == 3 and child > 2")


media_control = np.average(hijos2["sdrinkly"])
media_tratamiento = np.average(hijos3["sdrinkly"])

print(media_tratamiento-media_control)
```

```
-0.20633621142007136
```

Comparando ambos grupos en la wave 3, el tratamiento disminuye la probabilidad de "drinkly" en cerca de un 20%

3.4 Pregunta 14

Estime el efecto anterior usando la variable *married* como instrumento para determinar el efecto del tratamiento en la pregunta 12. Como se interpreta el efecto en este caso?

[247]:
```
Xa=charls[['married','female','age','hsize','wealth','schadj','retired','nrps','child']]
ya=charls['sdrinkly']
Xa = sm.add_constant(Xa)

model = sm.OLS(ya, Xa)
results = model.fit(cov_type="HC1")
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                sdrinkly   R-squared:                       0.146
Model:                            OLS   Adj. R-squared:                  0.145
```

Method:	Least Squares	F-statistic:	242.0
Date:	Mon, 28 Nov 2022	Prob (F-statistic):	0.00
Time:	19:59:23	Log-Likelihood:	-7273.2
No. Observations:	12965	AIC:	1.457e+04
Df Residuals:	12955	BIC:	1.464e+04
Df Model:	9		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.7249	0.041	17.843	0.000	0.645	0.805
married	-0.0469	0.014	-3.291	0.001	-0.075	-0.019
female	-0.3190	0.008	-38.081	0.000	-0.335	-0.303
age	-0.0026	0.001	-4.644	0.000	-0.004	-0.002
hsize	-0.0005	0.002	-0.228	0.820	-0.005	0.004
wealth	2.522e-08	6.41e-08	0.394	0.694	-1e-07	1.51e-07
schadj	0.0024	0.001	1.997	0.046	4.53e-05	0.005
retired	-0.0735	0.009	-8.081	0.000	-0.091	-0.056
nrps	0.0120	0.008	1.471	0.141	-0.004	0.028
child	-0.0215	0.003	-6.634	0.000	-0.028	-0.015

Omnibus:	2103.824	Durbin-Watson:	1.824
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1329.422
Skew:	0.663	Prob(JB):	2.09e-289
Kurtosis:	2.163	Cond. No.	6.75e+05

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 6.75e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Se usa married como instrumento de estimacion para sdrinkly

3.5 Pregunta 15

Finalmente, asuma que la intervencion se implementa en todos los individuos. Genere una nueva variable de tratamiento un nueva variable llamada *tdrinkly* donde el efecto es una reduccion de 50% en la prevalencia de consumo de alcohol en toda la poblacion en el tercer periodo (identica a *drinkly* en los periodos 1 y 2). Genere una variable *cdrinkly* que es identica a *drinkly* en los periodos 1 y 2 y use la informacion de ambos periodos para predicir el valor esperado de *drinkly* en el tercer periodo, estos seran los valores de *cdrinkly* en el periodo 3 (contrafactual). Finalmente, estime el efecto de la intervencion en toda la poblacion comparando entre *tdrinkly* (datos reales) versus *cdrinkly* contrafactual.