

Tarea3_Auil_Cabezas

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. 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).
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.
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.
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.
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.

2 Desarrollo Parte 1 - Experimentos

1. Para implementar un experimento a nivel de estudiante, tenemos la idea de separar de manera aleatoria a todos los estudiantes que pertenezcan al mismo ramo en dos secciones distintas del mismo tamaño, donde cada una de estas tendrá un horario distinto dentro del mismo día. Esta desición nace de la necesidad que se tiene de separar la clase debido a los pocos metros cuadrados que posee cada estudiante dentro de la sala de clases al ser solo un grupo. Cabe destacar que los estudiantes no tendrán la opción de asistir a un horario de clases distinto al asignado en su sección y la asistencia no será obligatoria en ningún caso. El costo asociado a este experimento será el pago al profesor que realizará las clases de ambas secciones de manera separada(pudiendo hacerla en una sola), además de conseguir la misma sala para ambos horarios(pudiendo ser utilizada para realizar clases de cualquier otra asignatura). Todo esto

se evaluará a lo largo de un año y en nto a los materiales, consideraremos la infraestructura de la universidad(salas,baños,proyectores,etc) ademas de una cantidad de 80 estudiantes los cuales seran divididos en ambas secciones de igual cantidad.

2. El grupo de control será aquella sección que tenga clases a las 8 am, pues este es el horario en el que creemos que existe la menor tasa de asistencia a clases de parte de los estudiantes, mientras que el grupo de tratamiento será definido como aquella sección que tenga clases en un horario distinto al grupo de control, donde proponemos que este horario sea a las 11 am, puesto que creemos que es una hora donde los estudiantes tendrán la posibilidad de tener más horas de descanso, por lo que suponemos que podrian asistir de una manera más recurrente a las clases. La asignacion aleatoria de los estudiantes a cada sección será realizada mediante un código de python o excel que separe al total de estos en dos grupos de igual cantidad de personas.
3. El método que consideramos más apropiado es el pre-post-test, ya que, como el objetivo principal de este experimento radica en aumentar la tasa de asistencia a clases, se vuelve necesario medir esta antes de realizar el experimento para despues hacer una comparativa con los resultados obtenidos tras realizar esta intervención y si el experimento afecta realmente en la tasa de asistencia. En cambio post-test posee una desventaja la cual radica en que se obtendrán resultados pero no se sabrá si dichos números son producidos de manera directa o no por el experimento realizado, además entregará  resultado que pertenece a solo a los efectos promedios, por consiguiente pudiendo inducir  una gran perdida de la muestra. Con Solomon cuatro grupos, serviria para reducir la influencia de las variables de confusión y permitir que los investigadores prueben si la misma prueba previa tien un efecto sobre los sujetos. Sin embargo este tipo de diseños es mucho más complejo de configurar y de analizar, pues combate muchos de los problemas de validez interna que pueden afectar de manera directa a la investigación.
4. Para implementar un experimento a nivel de clase en vez de estudiante, planteamos la segmentacion por sector en donde residen los estudiantes dentro de su día a día en época universitaria, en donde se pueden encontrar sectores como Concepción centro, Pedro de Valdivia, sector Lomas (Concepción), San Pedro, Hualpén, entre otras. La forma de asignar la clase y sus alumnos se realiza de manera aleatoria, obteniendo estudiantes de los diversos sectores mencionados anteriormente por clase. Para dicho experimento, el segmento elegido por nosotros como control será el sector de Concepción centro,  que por la cercanía que este sector tiene con respecto a la Universidad de Concepción se  que aquellos estudiantes que residan en este tienen una mayor tasa de asistencia a clases.
5. El programa que pretendemos implementar a nivel Universidad de Concepción se planifica de la siguiente forma: Normalmente los programas de las asignaturas consideran dos horas de clases teóricas y otras dos destinadas a las clases prácticas, sin embargo se cree que uno de los factores que aumentaría la tasa de asistencia de los jóvenes universitarios a clases es la motivación que estos tienen para ir a sus aulas o laboratorios, por lo que se piensa que que los alumnos valoran de mejor forma los conocimientos llevados a la práctica que estudiar tanto la teoría. Es por esto que se plantea una reducción de una hora a las clases teóricas y un aumento de una hora a las clases prácticas, de manera de ver como se comporta el porcentaje de asistencias con esta nueva modalidad a nivel universidad. Para la realización de este experimento se plantea un horizonte de trabajo de dos años, donde posteriormente se compararán las tasas de asistencias de ambos periodos. El grupo de control en este caso será el método actual con 2 horas teoricas y 2 horas practicas  y el grupo tratamiento será el

método propuesto con 1 hora teorica y 3 horas prácticas.

3 Parte 2 - Estimacion de efectos promedio de tratamiento (data simulada)

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 talque la asistencia promedio a traves de todos los grupos es de 80%.
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.
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.
9. Estime el efecto promedio del tratamiento usando pre-post test con la data generada. Muestre que el efecto es equivalente usando ambos metodos.
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.

4 Desarrollo Parte 2 - Estimacion de efectos promedio de tratamiento (data simulada)

4.1 Carga de Bibliotecas

```
[1]: 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 sklearn
import scipy
from scipy.linalg import eig, cholesky
from scipy.stats import norm
import linearmodels.panel as lmp
from pylab import plot, show, axis, subplot, xlabel, ylabel, grid

%matplotlib inline
```

4.2 6.

```
[2]: # experiment parameters
np.random.seed(43) #set seed
nsize = 4000 #sample size

# we create simulated data starting from a given variance-covariance matrix

# variance-covariance matrix (simetric)
cov = np.array([
    [ 3.40, -2.75, -2.00],
    [-2.75,  5.50,  1.50],
    [-2.00,  1.50,  1.25]
])

X = norm.rvs(size=(3, nsize))
evals, evecs = eigh(cov)
c = np.dot(evecs, np.diag(np.sqrt(evals)))
Xa = np.dot(c, X)
Xa = Xa.transpose()
X = X.transpose()
X = pd.DataFrame(X)
Xa = pd.DataFrame(Xa)
Xc = pd.DataFrame(np.c_[X,Xa], columns=['X1','X2','X3','X4','X5','X6'])

#time periods and treatment asignment
Xc['p'] = 1
Xc.loc[0:1999,'p'] = 0
tr = np.random.binomial(1, 0.5, size=2000) #treatment status
Xc.loc[0:1999,'T'] = tr
Xc.loc[2000:3999,'T'] = tr

#Se generan los 40 grupos de 50 estudiantes tanto para el periodo pre y post_
↪experimento, por ende la muestra total será
# de 4000 observaciones(40*50*2)

Xc['c1']=1
Xc.loc[50:99,'c1']=2
Xc.loc[100:149,'c1']=3
Xc.loc[150:199,'c1']=4
Xc.loc[200:249,'c1']=5
Xc.loc[250:299,'c1']=6
Xc.loc[300:349,'c1']=7
Xc.loc[350:399,'c1']=8
Xc.loc[400:449,'c1']=9
Xc.loc[450:499,'c1']=10
Xc.loc[500:549,'c1']=11
```

```
Xc.loc[550:599, 'c1']=12
Xc.loc[600:649, 'c1']=13
Xc.loc[650:699, 'c1']=14
Xc.loc[700:749, 'c1']=15
Xc.loc[750:799, 'c1']=16
Xc.loc[800:849, 'c1']=17
Xc.loc[850:899, 'c1']=18
Xc.loc[900:949, 'c1']=19
Xc.loc[950:999, 'c1']=20
Xc.loc[1000:1049, 'c1']=21
Xc.loc[1050:1099, 'c1']=22
Xc.loc[1100:1149, 'c1']=23
Xc.loc[1150:1199, 'c1']=24
Xc.loc[1200:1249, 'c1']=25
Xc.loc[1250:1299, 'c1']=26
Xc.loc[1300:1349, 'c1']=27
Xc.loc[1350:1399, 'c1']=28
Xc.loc[1400:1449, 'c1']=29
Xc.loc[1450:1499, 'c1']=30
Xc.loc[1500:1549, 'c1']=31
Xc.loc[1550:1599, 'c1']=32
Xc.loc[1600:1649, 'c1']=33
Xc.loc[1650:1699, 'c1']=34
Xc.loc[1700:1749, 'c1']=35
Xc.loc[1750:1799, 'c1']=36
Xc.loc[1800:1849, 'c1']=37
Xc.loc[1850:1899, 'c1']=38
Xc.loc[1900:1949, 'c1']=39
Xc.loc[1950:1999, 'c1']=40
Xc.loc[2050:2099, 'c1']=2
Xc.loc[2100:2149, 'c1']=3
Xc.loc[2150:2199, 'c1']=4
Xc.loc[2200:2249, 'c1']=5
Xc.loc[2250:2299, 'c1']=6
Xc.loc[2300:2349, 'c1']=7
Xc.loc[2350:2399, 'c1']=8
Xc.loc[2400:2449, 'c1']=9
Xc.loc[2450:2499, 'c1']=10
Xc.loc[2500:2549, 'c1']=11
Xc.loc[2550:2599, 'c1']=12
Xc.loc[2600:2649, 'c1']=13
Xc.loc[2650:2699, 'c1']=14
Xc.loc[2700:2749, 'c1']=15
Xc.loc[2750:2799, 'c1']=16
Xc.loc[2800:2849, 'c1']=17
Xc.loc[2850:2899, 'c1']=18
Xc.loc[2900:2949, 'c1']=19
```

```

Xc.loc[2950:2999, 'c1']=20
Xc.loc[3000:3049, 'c1']=21
Xc.loc[3050:3099, 'c1']=22
Xc.loc[3100:3149, 'c1']=23
Xc.loc[3150:3199, 'c1']=24
Xc.loc[3200:3249, 'c1']=25
Xc.loc[3250:3299, 'c1']=26
Xc.loc[3300:3349, 'c1']=27
Xc.loc[3350:3399, 'c1']=28
Xc.loc[3400:3449, 'c1']=29
Xc.loc[3450:3499, 'c1']=30
Xc.loc[3500:3549, 'c1']=31
Xc.loc[3550:3599, 'c1']=32
Xc.loc[3600:3649, 'c1']=33
Xc.loc[3650:3699, 'c1']=34
Xc.loc[3700:3749, 'c1']=35
Xc.loc[3750:3799, 'c1']=36
Xc.loc[3800:3849, 'c1']=37
Xc.loc[3850:3899, 'c1']=38
Xc.loc[3900:3949, 'c1']=39
Xc.loc[3950:3999, 'c1']=40

#outcome variable
alpha=0.85
beta=0.61
coef=-0.17

#Estos valores de alpha, beta y coef nacen gracias a la resolución de Xc["y"]
↳ en cada combinacion posible de T y p(4 casos)
#Por ejemplo, en el 1° caso en que T=0(sin tratamiento) y p=0(periodo inicial)
↳ resulta que el valor alpha(simil al valor Z)
#que entrega el 0.8(80% de asistencia) en la distribucion normal es alpha = 0.84

#Para el 2° caso, T=0(sin tratamiento) y p=1(periodo post experimento) resulta
↳ que Xc["y"]=alpha + coef = 0.75(75% asistencia)
# el valor Z que entrega este valor es de 0.68 y despejando el coef dado el
↳ valor alpha conocido anteriormente, tenemos
#que coef = -0.17

#Y de manera similar para el 3° caso con T=1 y p=1 tenemos que alpha + beta +
↳ coef = 0.9(90% de asistencia), y dados los
#valores ya conocidos de alpha y coef, calculamos que beta = 0.61

Xc['y'] = alpha*(Xc['X1']) + beta*(Xc['X2']*Xc['T']*Xc['p']) +
↳ coef*(Xc['X3']*Xc['p'])

```

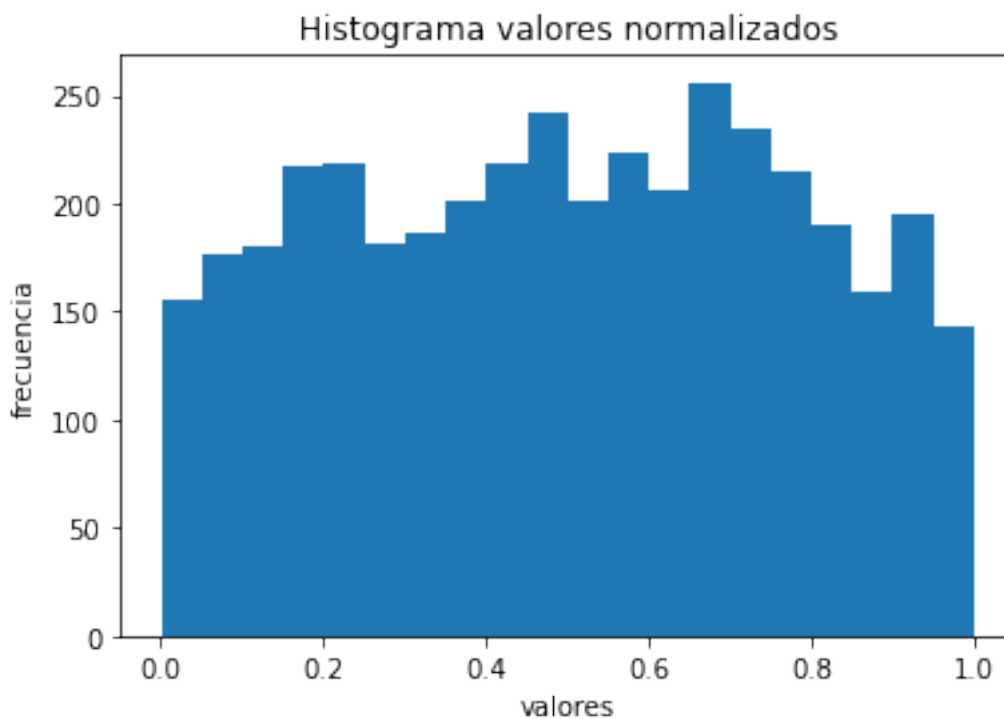
```
Xc.describe()
```

```
[2]:
```

	X1	X2	X3	X4	X5 \
count	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000
mean	0.009930	0.010328	-0.003044	-0.012652	-0.002598
std	1.002325	0.998561	1.010796	1.855616	2.370292
min	-3.309156	-3.678544	-3.031178	-5.810348	-7.701771
25%	-0.673139	-0.642355	-0.680084	-1.271982	-1.570262
50%	0.019748	0.029617	-0.025739	0.020108	0.004478
75%	0.706789	0.675787	0.688850	1.201377	1.614102
max	3.098634	3.497461	3.818642	7.121031	7.781027

	X6	p	T	cl	y
count	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000
mean	0.010753	0.500000	0.48150	20.50000	0.006967
std	1.123150	0.500063	0.49972	11.54484	0.910272
min	-4.383292	0.000000	0.00000	1.00000	-2.893001
25%	-0.722168	0.000000	0.00000	10.75000	-0.621658
50%	0.004757	0.500000	0.00000	20.50000	0.021362
75%	0.756007	1.000000	1.00000	30.25000	0.612902
max	3.519005	1.000000	1.00000	40.00000	3.701233

```
[3]: Xc["y_10"]=norm.cdf(Xc["y"])
plt.hist(Xc["y_10"], 20)
plt.ylabel('frecuencia')
plt.xlabel('valores')
plt.title('Histograma valores normalizados')
plt.show()
```



```
[4]: Xc["y_10"]
```

```
[4]: 0      0.586593
      1      0.219995
      2      0.373830
      3      0.324670
      4      0.767110
      ...
      3995    0.836497
      3996    0.795264
      3997    0.179705
      3998    0.742747
      3999    0.683556
      Name: y_10, Length: 4000, dtype: float64
```

4.3 7.

```
[5]: alpha=0.765
      beta=0.922
      teta=0.732

      Xc.loc[(Xc["y_10"]>=alpha) & (Xc["p"]==0) , "asistencia"]=0
      Xc.loc[(Xc["y_10"]<alpha) & (Xc["p"]==0) , "asistencia"]=1
```



```
Xc.loc[(Xc["y_10"]>=beta) & (Xc["p"]==1) & (Xc["T"]==1), "asistencia"]=0
Xc.loc[(Xc["y_10"]<beta) & (Xc["p"]==1) & (Xc["T"]==1), "asistencia"]=1
Xc.loc[(Xc["y_10"]>=teta) & (Xc["p"]==1) & (Xc["T"]==0), "asistencia"]=0
Xc.loc[(Xc["y_10"]<teta) & (Xc["p"]==1) & (Xc["T"]==0), "asistencia"]=1

Xc.groupby(by=["p", "T"]).mean()
```

```
[5]:
```

		X1	X2	X3	X4	X5	X6	c1	\
p	T								
0	0.0	0.009246	0.049406	0.014577	-0.017696	-0.075874	0.017018	20.143684	
	1.0	0.032544	0.016248	0.046399	0.065372	-0.114841	-0.027711	20.883697	
1	0.0	0.013097	0.004007	-0.061116	-0.101291	0.128912	0.061898	20.143684	
	1.0	-0.015357	-0.030868	-0.008926	0.010207	0.046933	-0.012606	20.883697	

		y	y_10	asistencia
p	T			
0	0.0	0.007859	0.502426	0.808100
	1.0	0.027663	0.510064	0.794393
1	0.0	0.021523	0.509610	0.750241
	1.0	-0.030365	0.489838	0.900312

4.4 8.

4.5 Post-test

```
[6]: y = Xc.loc[2000:3999, "asistencia"] #se toma solo la mitad de los datos
      ↪ totales(Xc["p"]==1)
X = Xc.loc[2000:3999, "T"]
X = sm.add_constant(X)
model = sm.OLS(y, X)
results = model.fit()
print(results.summary())
```

```

OLS Regression Results
=====
Dep. Variable:      asistencia      R-squared:      0.039
Model:              OLS             Adj. R-squared: 0.038
Method:             Least Squares   F-statistic:    80.03
Date:               Mon, 28 Nov 2022 Prob (F-statistic): 8.21e-19
Time:               18:45:47         Log-Likelihood: -874.41
No. Observations:   2000             AIC:            1753.
Df Residuals:       1998             BIC:            1764.
Df Model:           1
Covariance Type:    nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	0.7502	0.012	64.452	0.000	0.727	0.773

T	0.1501	0.017	8.946	0.000	0.117	0.183
---	--------	-------	-------	-------	-------	-------

```
=====
```

Omnibus:	477.856	Durbin-Watson:	2.038
Prob(Omnibus):	0.000	Jarque-Bera (JB):	883.065
Skew:	-1.582	Prob(JB):	1.76e-192
Kurtosis:	3.762	Cond. No.	2.58

```
=====
```

Notes:

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

4.6 9.

4.7 Pre-post test

```
[7]: 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()
print(results2.summary())
```

OLS Regression Results

```
=====
```

Dep. Variable:	asistencia	R-squared:	0.019
Model:	OLS	Adj. R-squared:	0.019
Method:	Least Squares	F-statistic:	26.20
Date:	Mon, 28 Nov 2022	Prob (F-statistic):	8.86e-17
Time:	18:45:47	Log-Likelihood:	-1877.7
No. Observations:	4000	AIC:	3763.
Df Residuals:	3996	BIC:	3788.
Df Model:	3		
Covariance Type:	nonrobust		

```
=====
```

	coef	std err	t	P> t	[0.025	0.975]
--	------	---------	---	------	--------	--------

```
-----
```

const	0.8081	0.012	67.222	0.000	0.785	0.832
p	-0.0579	0.017	-3.403	0.001	-0.091	-0.025
T	-0.0137	0.017	-0.791	0.429	-0.048	0.020
dd	0.1638	0.025	6.685	0.000	0.116	0.212

```
=====
```

Omnibus:	898.162	Durbin-Watson:	2.010
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1635.706
Skew:	-1.546	Prob(JB):	0.00
Kurtosis:	3.507	Cond. No.	6.75

```
=====
```

Notes:

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

4.8 10.

4.9 Clusters

```
[8]: results3 = model.fit(cov_type="cluster", cov_kwds={'groups': Xc['cl']})  
print(results3.summary())
```

```

                        OLS Regression Results
=====
Dep. Variable:          asistencia    R-squared:                0.019
Model:                  OLS          Adj. R-squared:            0.019
Method:                 Least Squares  F-statistic:              30.57
Date:                  Mon, 28 Nov 2022  Prob (F-statistic):      2.46e-10
Time:                  18:45:47       Log-Likelihood:           -1877.7
No. Observations:      4000          AIC:                     3763.
Df Residuals:          3996          BIC:                     3788.
Df Model:               3
Covariance Type:       cluster
=====

```

	coef	std err	z	P> z	[0.025	0.975]
const	0.8081	0.011	74.533	0.000	0.787	0.829
p	-0.0579	0.013	-4.322	0.000	-0.084	-0.032
T	-0.0137	0.014	-1.005	0.315	-0.040	0.013
dd	0.1638	0.021	7.957	0.000	0.123	0.204

```
=====
Omnibus:                898.162    Durbin-Watson:           2.010
Prob(Omnibus):           0.000    Jarque-Bera (JB):        1635.706
Skew:                   -1.546    Prob(JB):                0.00
Kurtosis:               3.507    Cond. No.                6.75
=====
```

Notes:

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

5 Parte 3 - Experimentos naturales

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

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.

12. Estime el efecto del tratamiento usando diferencias en diferencias, comparando entre los periodos 2 y 3.
13. Compare el efecto del tratamiento generando grupos pseudo-equivalentes, en particular entre individuos solo con 3 hijos (tratamiento) y 2 hijos (control).
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?
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 predecir 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.

```
[9]: charls = pd.read_csv('../TAREA 4-5/charls.csv')
charls.dropna(inplace=True)
charls.reset_index(drop=True, inplace=True)

charls.describe()
```

```
[9]:
```

	age	bnrps	cesd	child	dnrps \
count	21045.000000	21045.000000	21045.000000	21045.000000	21045.000000
mean	59.386553	59.610683	8.656878	2.825232	0.740889
std	9.016106	51.905928	6.307677	1.372179	0.438157
min	20.000000	0.000000	0.000000	0.000000	0.000000
25%	52.000000	0.000000	4.000000	2.000000	0.000000
50%	59.000000	60.000000	7.000000	3.000000	1.000000
75%	65.000000	74.875404	12.000000	4.000000	1.000000
max	95.000000	300.000000	30.000000	10.000000	1.000000

	female	hrsusu	hsize	intmonth	married \
count	21045.000000	21045.000000	21045.000000	21045.000000	21045.000000
mean	0.521026	2.548166	3.585222	7.511143	0.907674
std	0.499570	1.757182	1.720136	0.865851	0.289492
min	0.000000	0.000000	1.000000	1.000000	0.000000
25%	0.000000	0.000000	2.000000	7.000000	1.000000
50%	1.000000	3.401197	3.000000	7.000000	1.000000
75%	1.000000	4.025352	5.000000	8.000000	1.000000
max	1.000000	5.123964	16.000000	12.000000	1.000000

	nrps	retage	retired	schadj	urban \
count	21045.000000	21045.000000	21045.000000	21045.000000	21045.000000
mean	0.519078	1.280969	0.204942	4.162414	0.206652
std	0.499648	3.830963	0.403669	3.540039	0.404914

min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000
50%	1.000000	0.000000	0.000000	4.000000	0.000000
75%	1.000000	0.000000	0.000000	8.000000	0.000000
max	1.000000	51.000000	1.000000	16.000000	1.000000

	wave	wealth	inid
count	21045.000000	2.104500e+04	21045.000000
mean	1.909385	6.783959e+03	12747.082870
std	0.817975	5.453065e+04	7769.025809
min	1.000000	-1.648450e+06	1.000000
25%	1.000000	1.000000e+02	5176.000000
50%	2.000000	1.000000e+03	13314.000000
75%	3.000000	6.800000e+03	19650.000000
max	3.000000	1.040000e+06	25403.000000

[]: