Tarea2_Munoz

October 13, 2022

1 TAREA 2: Laboratorio de métodos aplicados avanzados

1.0.1 Autor: Sebastián Muñoz

1.0.2 Fecha: 05/10/22

1.1 Carga de datos y limpieza

[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
import linearmodels.panel as lmp
import seaborn as sbn
import numpy.linalg as la
from scipy import stats
import pingouin as pg

%matplotlib inline
```

[2]:
```
charls = pd.read_csv('../data/charls.csv')
#Se transforman los missing values para drinkly en valores NaN con el fin de␣
 ↪poder ser reconocidos por Pandas para poder descatarlos de forma correcta.
charls['drinkly']=charls['drinkly'].replace('.m:missing',np.nan)
charls.dropna(inplace=True)
charls.reset_index(drop=True, inplace=True)
charls.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34354 entries, 0 to 34353
Data columns (total 15 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   cesd      34354 non-null  int64
 1   child     34354 non-null  int64
 2   drinkly   34354 non-null  object
```

```
 3   female    34354 non-null  int64
 4   hrsusu    34354 non-null  float64
 5   hsize     34354 non-null  int64
 6   inid      34354 non-null  float64
 7   intmonth  34354 non-null  int64
 8   married   34354 non-null  int64
 9   retired   34354 non-null  int64
 10  schadj    34354 non-null  int64
 11  urban     34354 non-null  int64
 12  wave      34354 non-null  int64
 13  wealth    34354 non-null  float64
 14  age       34354 non-null  int64
dtypes: float64(3), int64(11), object(1)
memory usage: 3.9+ MB
```

Cargada la database, se identifica un total de 34371 observaciones, donde se distingue la presencia de variables ficticias, tales como: drinkly; female; married; retired y urban, las demás variables de la base de datos son cuantitativas.

A modo de limpiar la data, se removieron muestras que incluian valores NaN, que provenian principalmente desde la variable drinkly, descartando un total de 71 observaciones, dejando entonces un total de 34354 para análisis...

[3]:
```
sbn.pairplot(charls)
```

[3]:
```
<seaborn.axisgrid.PairGrid at 0x1b9205b0be0>
```

Del diagrama de dispersión, a modo general llama la atención lo siguiente: - De la correlacion entre todas las variables con la variable dependiente de resultado 'cesd', la que presenta un claro caso de outliers es wealth vs cesd. - La proporción entre hombres y mujeres es bastante cercana más no igual. - La proporción de personas con más observaciones en la variable 'hrsusu' correspondea 0 horas trabajo a la semana. - En 'retired', casi el 70% de las observaciones corresponden a personas que están jubiladas, lo cual podría estar correlacionado al punto anterior. - En 'married', casi el 90% de los encuestados son casados

[4]:
```
#corr = pg.pairwise_corr(charls, columns=['hrsusu','retired'],␣
 ↪method='spearman')
corr = pg.pairwise_corr(charls[['hrsusu','retired']], method='spearman')
corr
```

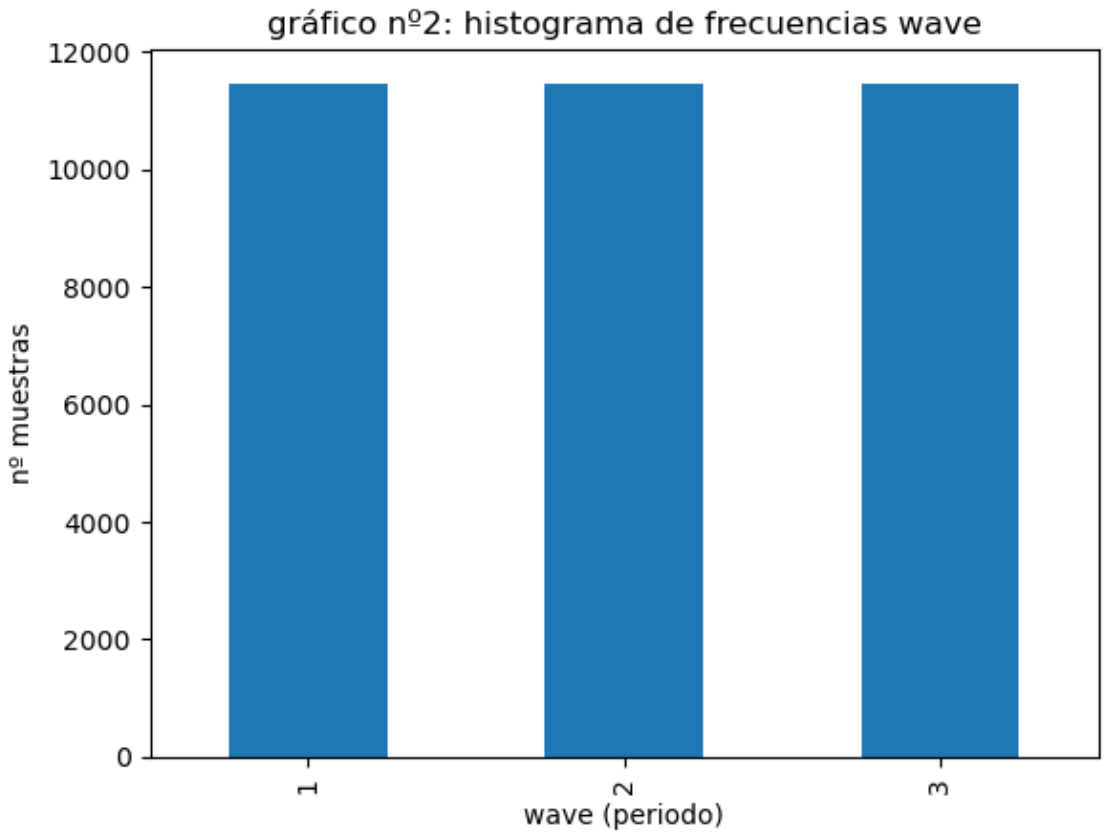
[4]:
```
         X        Y    method alternative      n         r          CI95%  \
0  hrsusu  retired  spearman   two-sided  34354 -0.707089  [-0.71, -0.7]

   p-unc  power
0    0.0    1.0
```

De la correlación entre las variables 'hrsusu' y 'retired' usando el metodo Spearman (ya que las variables no presentan distribución normal como para utilizar otro método como Person por ejemplo), se confirma que existe una alta correlación y negativa entre ambas variables, significativa con un valor-p practicamente 0, es por esa razón que para el modelo se eliminará una de ellas y en este caso se escogerá la variable binaria 'retired', ya que 'hrsusu' es continua y tambien por que al abarcar un amplio rango de valores puede entregar una mayor información al momento de explicar la variable de resultado.

[5]:
```
plt.title('gráfico nº2: histograma de frecuencias wave')
plt.xlabel('wave (periodo)')
plt.ylabel('nº muestras')
charls['wave'].value_counts().sort_index().plot(kind='bar')
charls['wave'].value_counts()
```

[5]:
```
1    11457
3    11452
2    11445
Name: wave, dtype: int64
```



Observando el gráfico nº2, se observa que no existe una atrición importante en el panel de datos (pero si la hay), una de las formas para corregir esto podría ser realizando imputación o bien por medio del uso de pesos relativos, sin embargo para efectos prácticos se asumirá que el panel está balanceado dado que el hecho de haber una atrición baja, no es impedimento para las librerias estadisticas para poder realizar una regresión no significativa.

A modo de controlar el efecto promedio que tiene el tiempo o periodo sobre la variable de resultado, se creará un total de 2 variables dicotomicas de 'wave', donde:

- wave_2 = {1: El periodo de la encuesta es el nº2; 0: en otro caso}
- wave_3 = {1: El periodo de la encuesta es el nº3; 0: en otro caso}

Categoria de referencia: El periodo de la encuesta es el nº1 (wave_2=0 y wave_3=0)

Además, como la información con respecto a la variable binaria 'drinkly' en la data estaba de forma de string y no directamente como valores en el conjunto {0,1}, se convertirán manualmente mediante la función map en dichos valores númericos con el fin de evitar eliminar esta variable por dicha problemática.

[6]:
```
charls['wave_2']=charls.wave.map({1:0, 2:1, 3:0})
charls['wave_3']=charls.wave.map({1:0, 2:0, 3:1})

charls['Drinkly']=charls.drinkly.map({'1.Yes':1, '0.None':0})
```

```
charls.drop(['drinkly'], axis=1, inplace=True)
```

Como la variable intmonth es categorica ya que representa el mes del año en el que la persona fue entrevistada, podría convertirse en una serie de variables dummy, sin embargo el hacer esto implica tener en el modelo 9 variables dicotomicas (pues ninguna persona fue entrevistada en abril ((intmonth=4) o mayo (intmonth=5)) generando un exceso de variables explicativas.

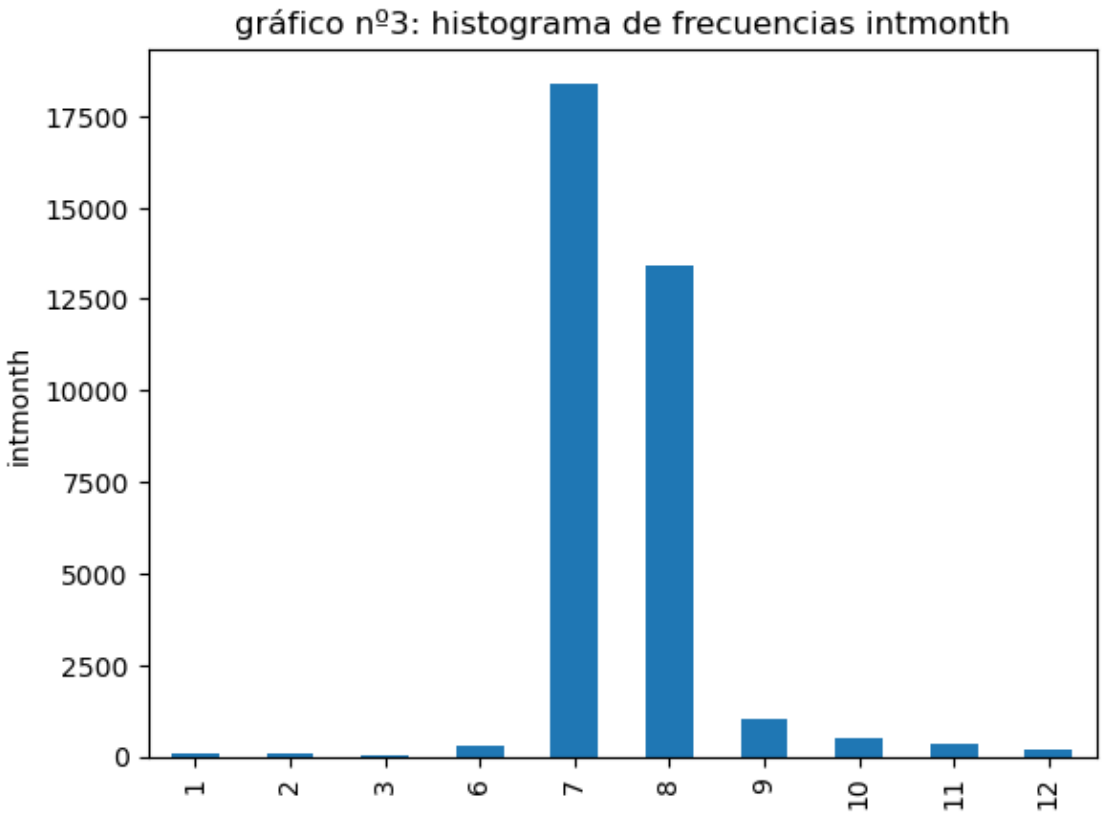
A pesar de eso, en base al gráfico nº3 y por simple inspección, resulta interesante pensar que es posible agrupar la variable intmonth en grupos equitativos, en este caso 3 cuatrimestres, donde en el primero se incluyan los meses desde el 1 hasta el 4, en el segundo desde el 5 hasta el 8 y en el último los meses 9 hasta el 12. Con esta idea en mente, en el gráfico se puede observar que el 1º cuatrimestre es en el que menos observaciones tiene registradas; el 2º cuatrimestre, el cual concentra por lejos la mayor parte de las observaciones en el mes 7 y 8; por último el 3º cuatrimestre presenta una concentracion de observaciones mucho menor que el 2º cuatrimestre pero mayor al 1º cuatrimestre. Al hacer esto se reduce enormemente la cantidad de variables a incluir en el modelo a final a solamente 2 variables dummy:

- intmonth_2 = {1: Si la persona fue encuestada en el 2º cuatrimestre; 0: en otro caso}
- intmonth_3 = {1: Si la persona fue encuestada en el 3ª cuatrimestre; 0: en otro caso}

Categoria de referencia: La persona no fue encuestada en el 1º cuatrimestre (intmonth_2=0 e intmonth_3=0)

[7]:
```
plt.title('gráfico nº3: histograma de frecuencias intmonth')
plt.ylabel('intmonth')
charls['intmonth'].value_counts().sort_index().plot(kind='bar')
charls['intmonth_2']=charls.intmonth.map({1:0, 2:0, 3:0, 4:0, 5:1, 6:1, 7:1, 8:
 ↪1, 9:0, 10:0, 11:0, 12:0})
charls['intmonth_3']=charls.intmonth.map({1:0, 2:0, 3:0, 4:0, 5:0, 6:0, 7:0, 8:
 ↪0, 9:1, 10:1, 11:1, 12:1})

charls.drop(['intmonth'], axis=1, inplace=True)
```



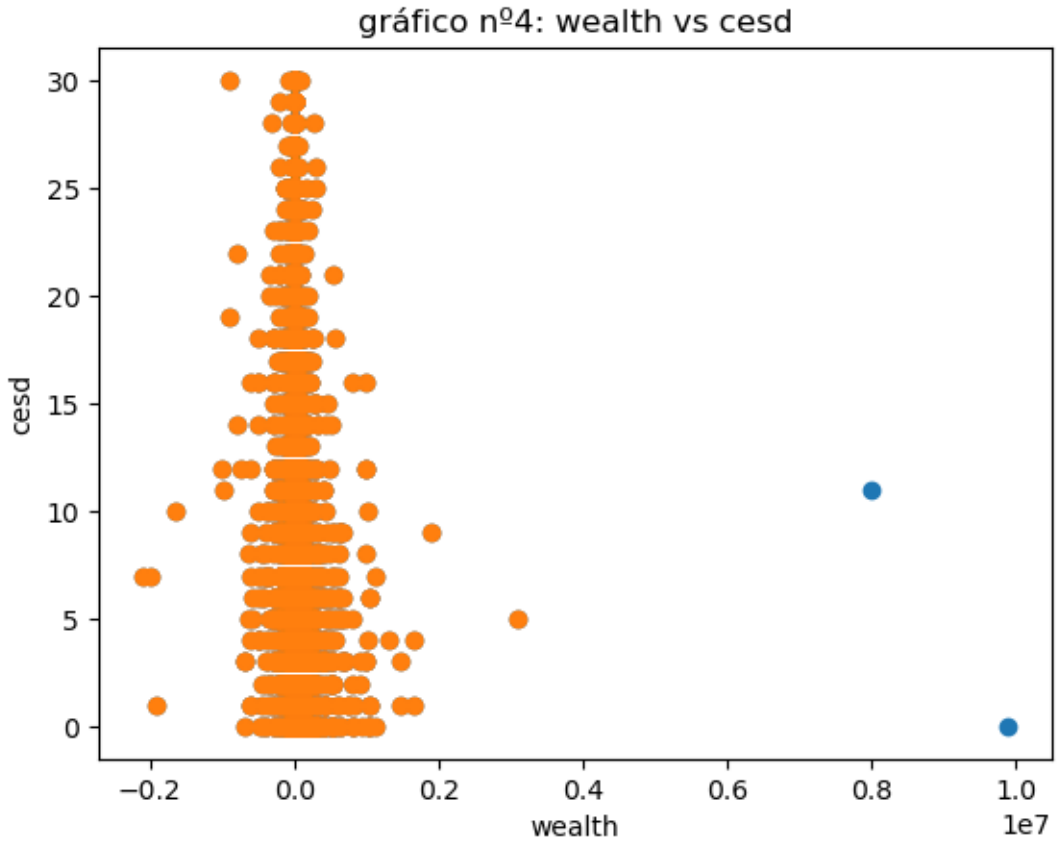
En el gráfico nº4 se detectaron dos outliers y para eliminarlos se ha hecho un subset_0 descartando las observaciones anómalas observadas en dicha gráfica, en el segundo gráfico a partir del susbet_0 se detectaron nuevos outliers y de la misma forma se han ido eliminando; luego de descartar los outliers el total de muestras se redujo a 34352 observaciones.

[8]:
```
plt.title('gráfico nº4: wealth vs cesd')
plt.xlabel('wealth')
plt.ylabel('cesd')
plt.scatter(charls['wealth'],charls['cesd'])
subset_0=charls.loc[charls['wealth']<0.6*(10**7)]
plt.scatter(subset_0['wealth'],subset_0['cesd'])

subset_0.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 34352 entries, 0 to 34353
Data columns (total 18 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   cesd        34352 non-null  int64
 1   child       34352 non-null  int64
 2   female      34352 non-null  int64
```

```
 3   hrsusu      34352 non-null  float64
 4   hsize       34352 non-null  int64
 5   inid        34352 non-null  float64
 6   married     34352 non-null  int64
 7   retired     34352 non-null  int64
 8   schadj      34352 non-null  int64
 9   urban       34352 non-null  int64
 10  wave        34352 non-null  int64
 11  wealth      34352 non-null  float64
 12  age         34352 non-null  int64
 13  wave_2      34352 non-null  int64
 14  wave_3      34352 non-null  int64
 15  Drinkly     34352 non-null  int64
 16  intmonth_2  34352 non-null  int64
 17  intmonth_3  34352 non-null  int64
dtypes: float64(3), int64(15)
memory usage: 5.0 MB
```



[9]:
```
X=subset_0[['child','Drinkly','hrsusu','hsize','female','intmonth_2','intmonth_3','married','r
Xm=(X.groupby(subset_0['inid']).transform('mean'))
```

```
X_ID=subset_0[['inid','wave','cesd','child','Drinkly','hrsusu','hsize','female','married','ret
Xc=pd.DataFrame(np.c_[X_ID, Xm],␣
 ↪columns=['inid','wave','cesd','child','Drinkly','hrsusu','hsize','female','married','retire

#se establece el index del panel
Xc = Xc.set_index(["inid","wave"])
Xc.describe()
```

[9]:

	cesd	child	Drinkly	hrsusu	hsize	\
count	34352.000000	34352.000000	34352.000000	34352.000000	34352.000000	
mean	8.173643	2.708197	0.331596	27.970722	3.441168	
std	6.183829	1.399927	0.470794	28.273270	1.702504	
min	0.000000	0.000000	0.000000	0.000000	1.000000	
25%	3.000000	2.000000	0.000000	0.000000	2.000000	
50%	7.000000	2.000000	0.000000	21.000000	3.000000	
75%	12.000000	3.000000	1.000000	54.000000	4.000000	
max	30.000000	11.000000	1.000000	168.000000	16.000000	

	female	married	retired	schadj	urban	\
count	34352.000000	34352.000000	34352.000000	34352.000000	34352.000000	
mean	0.537873	0.874592	0.283914	4.677632	0.356195	
std	0.498571	0.331185	0.450902	3.857608	0.478881	
min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	1.000000	0.000000	0.000000	0.000000	
50%	1.000000	1.000000	0.000000	4.000000	0.000000	
75%	1.000000	1.000000	1.000000	8.000000	1.000000	
max	1.000000	1.000000	1.000000	18.000000	1.000000	

	…	mintmonth_2	mintmonth_3	mmarried	mretired	\
count	…	34352.000000	34352.000000	34352.000000	34352.000000	
mean	…	0.935026	0.059676	0.874592	0.283914	
std	…	0.113486	0.108508	0.184905	0.229963	
min	…	0.333333	0.000000	0.000000	0.000000	
25%	…	0.940367	0.000000	0.867284	0.157658	
50%	…	0.992424	0.006873	0.901709	0.250000	
75%	…	1.000000	0.055556	0.964912	0.333333	
max	…	1.000000	0.666667	1.000000	1.000000	

	mschadj	murban	mwealth	mage	mwave_2	\
count	34352.000000	34352.000000	34352.000000	34352.000000	34352.000000	
mean	4.677632	0.356195	9370.798561	58.223539	0.333168	
std	2.347706	0.381082	26626.492141	5.243503	0.005736	
min	0.000000	0.000000	-325000.000000	16.000000	0.000000	
25%	3.764706	0.000000	400.000000	56.418803	0.333333	
50%	4.363636	0.250000	5016.666667	58.000000	0.333333	

```
75%        5.500000      0.617284   12873.398374     60.406977      0.333333
max       16.000000      1.000000  663010.000000     89.000000      0.500000

            mwave_3
count  34352.000000
mean       0.333314
std        0.005720
min        0.000000
25%        0.333333
50%        0.333333
75%        0.333333
max        0.500000

[8 rows x 31 columns]
```

1.2 Modelo Pooled OLS

La finalidad del analisis será comparar modelos con la misma cantidad de variables explicativas, a modo de evitar concluir o interpretar de forma incorrecta los resultados; entonces haciendo un adelanto al modelo de efectos fijos (FE), al ser las variables 'female' y 'age' constantes por individuo en los periodos de la encuesta, no se pueden incluir en dicho modelo; por lo que la carencia de estas variables en FE y la presencia de las mismas en Efectos aleatorios (RE), se presentaría un problema al realizar el test de Hausman: 1) por la diferencia en la dimensión del vector de parametros y covarianzas entre ambos modelos y 2) en el cálculo de los grados de libertad para dicho test; es por eso que se decidió de antemano erradicar esas 2 variables tanto en Pooled OLS, FE y RE y poder comparar modelos similares en el número de variables explicativas.

Por otro lado, se incluirán todas las demás variables de la base de datos, esto debido a que al menos desde el punto de vista lógico, el efecto que cada una de ellas pueda incidir en la variable explicativa 'cesd', parece razonable y hace sentido.

[10]:
```
y=Xc['cesd']
#X=Xc[['child','Drinkly','hrsusu','hsize','female','married','schadj','urban','wealth','age',
X=Xc[['child','Drinkly','hrsusu','hsize','married','schadj','urban','wealth','intmonth_2','int
X=sm.add_constant(X)

model=lmp.PooledOLS(y,X)
OLS=model.fit(cov_type="robust")
print(OLS)
```

```
                          PooledOLS Estimation Summary
================================================================================
Dep. Variable:                    cesd   R-squared:                        0.0635
Estimator:                   PooledOLS   R-squared (Between):              0.0845
No. Observations:                34352   R-squared (Within):               0.0386
Date:                Wed, Oct 05 2022   R-squared (Overall):              0.0635
Time:                        13:41:58   Log-likelihood                -1.102e+05
Cov. Estimator:                Robust
```

		F-statistic:	194.07
Entities:	3459	P-value	0.0000
Avg Obs:	9.9312	Distribution:	F(12,34339)
Min Obs:	2.0000		
Max Obs:	468.00	F-statistic (robust):	197.23
		P-value	0.0000
Time periods:	3	Distribution:	F(12,34339)
Avg Obs:	1.145e+04		
Min Obs:	1.144e+04		
Max Obs:	1.146e+04		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	10.096	0.3974	25.405	0.0000	9.3175	10.875
child	0.1116	0.0253	4.4157	0.0000	0.0620	0.1611
Drinkly	-0.8509	0.0683	-12.450	0.0000	-0.9848	-0.7169
hrsusu	-0.0059	0.0012	-4.8609	0.0000	-0.0082	-0.0035
hsize	-0.0272	0.0197	-1.3839	0.1664	-0.0658	0.0113
married	-1.4521	0.1119	-12.974	0.0000	-1.6715	-1.2327
schadj	-0.2279	0.0089	-25.729	0.0000	-0.2453	-0.2106
urban	-1.1106	0.0698	-15.904	0.0000	-1.2475	-0.9738
wealth	-4.87e-06	5.652e-07	-8.6158	0.0000	-5.978e-06	-3.762e-06
intmonth_2	1.3278	0.3739	3.5511	0.0004	0.5949	2.0607
intmonth_3	1.1758	0.3940	2.9848	0.0028	0.4037	1.9480
wave_2	-0.4635	0.0776	-5.9716	0.0000	-0.6157	-0.3114
wave_3	-0.1952	0.0836	-2.3359	0.0195	-0.3590	-0.0314

En base a estos resultados, se puede decir a modo general que en realidad todas las variables explicativas que se ingresaron al modelo son significativas, dado que el valor p de cada una de ellas no sobrepasa el valor de 0.1, por lo que se rechaza la hipotesis nula de que estos coeficientes sean iguales a 0.

Analizando el valor de los coeficientes, para este modelo el coef. que es de mayor importancia o el que más aporta a la explicación de la variable dependiente es el de married; ya que indica que el hecho de que una persona esté casada, disminuye 1.4 el puntaje en la escala de salud mental de la persona que contestó la encuesta.

Además, es interesante ver que la conversión de la variable intmonth que representaba 12 meses a 2 variables dicotomicas que representaban 3 cuatrimestres resultaron ser significativas para este modelo, donde la segunda variable más robusta para el modelo fue intmonth_2, que indica que el hecho de que se encueste a la persona en el segundo cuatrimestre del año aumenta 1.3 el puntaje en la escala de salud mental de esa persona.

En principio se añadieron 2 variables dicotomicas para medir el impacto de cada uno de los 3 periodos en los que se realizó la encuesta, las conclusiones que se pueden hacer al respecto dado que estas variables resultaron ser significativas es que, el efecto que tiene el periodo 2 (wave_2=1) sobre la variable explicativa (cesd) con respecto al periodo inicial (periodo 1), hace que se disminuyan

casi 0.5 puntos en la escala de salud mental.

El efecto que tiene la variable 'wealth' sobre la variable resultado (puntaje en la escala de salud mental) es prácticamente nulo.

1.3 Efectos fijos

[11]:
```
#X=Xc[['child','Drinkly','hrsusu','hsize','married','schadj','urban','wealth','intmonth_2','in
X=Xc[['child','Drinkly','hrsusu','hsize','married','schadj','urban','wealth','intmonth_2','int
X=sm.add_constant(X)
model=lmp.PanelOLS(y,X, entity_effects=True)
fe=model.fit(cov_type="robust")
print(fe)
```

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:                    cesd   R-squared:                        0.0416
Estimator:                   PanelOLS   R-squared (Between):              0.0681
No. Observations:               34352   R-squared (Within):               0.0416
Date:                Wed, Oct 05 2022   R-squared (Overall):              0.0588
Time:                        13:41:58   Log-likelihood                -1.055e+05
Cov. Estimator:                Robust
                                        F-statistic:                      111.78
Entities:                        3459   P-value                           0.0000
Avg Obs:                       9.9312   Distribution:                F(12,30881)
Min Obs:                       2.0000
Max Obs:                       468.00   F-statistic (robust):             94.615
                                        P-value                           0.0000
Time periods:                       3   Distribution:                F(12,30881)
Avg Obs:                    1.145e+04
Min Obs:                    1.144e+04
Max Obs:                    1.146e+04

                             Parameter Estimates
================================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
--------------------------------------------------------------------------------
const          10.468     0.4151     25.217     0.0000      9.6541      11.281
child          0.0052     0.0321     0.1628     0.8707     -0.0576      0.0681
Drinkly       -0.8575     0.0767    -11.183     0.0000     -1.0078     -0.7072
hrsusu        -0.0102     0.0013    -7.7333     0.0000     -0.0128     -0.0076
hsize         -0.0905     0.0229    -3.9523     0.0001     -0.1354     -0.0456
married       -1.3545     0.1354    -10.003     0.0000     -1.6199     -1.0891
schadj        -0.2090     0.0112    -18.645     0.0000     -0.2310     -0.1870
urban         -0.6131     0.1156    -5.3047     0.0000     -0.8396     -0.3866
wealth     -2.616e-06   4.99e-07    -5.2426     0.0000  -3.594e-06  -1.638e-06
intmonth_2     1.2413     0.3727     3.3303     0.0009      0.5107      1.9719
intmonth_3     1.0090     0.3927     2.5692     0.0102      0.2392      1.7787
```

```
wave_2         -0.4671     0.0713     -6.5472     0.0000     -0.6070     -0.3273
wave_3         -0.2638     0.0776     -3.3988     0.0007     -0.4159     -0.1117
==============================================================================

F-test for Poolability: 2.8428
P-value: 0.0000
Distribution: F(3458,30881)

Included effects: Entity
```

En este modelo (FE), a comparación con el modelo anterior: - La variable 'child' cambió el signo y también aumentó su efecto considerablemente pero a costa de que se perdiera significancia. - La variable 'Drinkly' aumentó su efecto. - La variable 'urban' disminuyó su efecto considerablemente. - La variable 'wealth' "disminuyó" su efecto, pero el valor del coef. sigue siendo prácticamente 0 o nulo.

El P-Value al final de los resultados concluye que el modelo de efectos fijos es significativo y explica de mejor forma la variable dependiente en comparación al modelo Pooled OLS.

1.4 Efectos aleatorios

[12]:
```
#X=Xc[['child','Drinkly','hrsusu','hsize','female','married','retired','schadj','urban','weal
X=Xc[['child','Drinkly','hrsusu','hsize','married','schadj','urban','wealth','intmonth_2','int
X=sm.add_constant(X)
model=lmp.RandomEffects(y,X)
re=model.fit(cov_type="robust")
print(re)
re.variance_decomposition
```

```
                        RandomEffects Estimation Summary
================================================================================
Dep. Variable:                    cesd   R-squared:                        0.1831
Estimator:                RandomEffects   R-squared (Between):              0.0861
No. Observations:                 34352   R-squared (Within):               0.0412
Date:                Wed, Oct 05 2022   R-squared (Overall):              0.0553
Time:                        13:41:58   Log-likelihood                -1.073e+05
Cov. Estimator:                Robust
                                        F-statistic:                      641.31
Entities:                        3459   P-value                           0.0000
Avg Obs:                       9.9312   Distribution:                F(12,34339)
Min Obs:                       2.0000
Max Obs:                       468.00   F-statistic (robust):             134.74
                                        P-value                           0.0000
Time periods:                       3   Distribution:                F(12,34339)
Avg Obs:                    1.145e+04
Min Obs:                    1.144e+04
Max Obs:                    1.146e+04

                             Parameter Estimates
```

```
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
const          10.976     0.3911     28.066     0.0000      10.209      11.742
child          0.0116     0.0274     0.4251     0.6708     -0.0420      0.0653
Drinkly        -0.8610     0.0689    -12.489     0.0000     -0.9961     -0.7259
hrsusu         -0.0090     0.0012    -7.4455     0.0000     -0.0114     -0.0066
hsize          -0.0802     0.0205    -3.9184     0.0001     -0.1204     -0.0401
married        -1.3647     0.1161    -11.751     0.0000     -1.5923     -1.1371
schadj         -0.2141     0.0096    -22.393     0.0000     -0.2329     -0.1954
urban          -0.8986     0.0912    -9.8554     0.0000     -1.0774     -0.7199
wealth     -3.255e-06  4.899e-07    -6.6447     0.0000  -4.216e-06  -2.295e-06
intmonth_2     1.2604     0.3571     3.5297     0.0004      0.5605      1.9602
intmonth_3     1.1075     0.3761     2.9448     0.0032      0.3704      1.8446
wave_2         -0.4653     0.0712    -6.5401     0.0000     -0.6048     -0.3259
wave_3         -0.2396     0.0773    -3.0981     0.0019     -0.3912     -0.0880
==============================================================================
```

[12]:
```
Effects                  13.222446
Residual                 30.216269
Percent due to Effects    0.304393
Name: Variance Decomposition, dtype: float64
```

En este modelo (RE), a comparación con el modelo anterior: - La variable 'child' permanece como no significativa. - La variable 'urban' aumentó el efecto que habia disminuido en el modelo de fijos. - el efecto de 'wealth' sigue siendo prácticamente despreciable.

Al asumir con este modelo que $Cov(x_{itk}, \mu_i) = 0, \forall t \in \{1,2,3\}, \forall k \in \{1,..,12\}$, entonces la informacion al final de los resultados indica que la proporción del error compuesto: $\mu_i + U_i$ correspondiente al error no observable fijo μ_i es de 0.3 y el restante 0.7 es perteneciente al error no observable que varia en el tiempo o idiosincrático U_i

1.5 Comparación entre modelos

[13]:
```
print(lmp.compare({"Pooled": OLS, "FE": fe, "RE":re}, precision="pvalues"))
```

```
                            Model Comparison
================================================================================
                                Pooled             FE                 RE
--------------------------------------------------------------------------------
Dep. Variable                     cesd           cesd               cesd
Estimator                   PooledOLS        PanelOLS     RandomEffects
No. Observations                34352          34352              34352
Cov. Est.                      Robust         Robust             Robust
R-squared                      0.0635         0.0416             0.1831
R-Squared (Within)             0.0386         0.0416             0.0412
R-Squared (Between)            0.0845         0.0681             0.0861
R-Squared (Overall)            0.0635         0.0588             0.0553
```

F-statistic	194.07	111.78	641.31
P-value (F-stat)	0.0000	0.0000	0.0000
==========	==========	==========	==========
const	10.096	10.468	10.976
	(0.0000)	(0.0000)	(0.0000)
child	0.1116	0.0052	0.0116
	(1.01e-05)	(0.8707)	(0.6708)
Drinkly	-0.8509	-0.8575	-0.8610
	(0.0000)	(0.0000)	(0.0000)
hrsusu	-0.0059	-0.0102	-0.0090
	(1.174e-06)	(1.066e-14)	(9.881e-14)
hsize	-0.0272	-0.0905	-0.0802
	(0.1664)	(7.756e-05)	(8.931e-05)
married	-1.4521	-1.3545	-1.3647
	(0.0000)	(0.0000)	(0.0000)
schadj	-0.2279	-0.2090	-0.2141
	(0.0000)	(0.0000)	(0.0000)
urban	-1.1106	-0.6131	-0.8986
	(0.0000)	(1.136e-07)	(0.0000)
wealth	-4.87e-06	-2.616e-06	-3.255e-06
	(0.0000)	(1.594e-07)	(3.084e-11)
intmonth_2	1.3278	1.2413	1.2604
	(0.0004)	(0.0009)	(0.0004)
intmonth_3	1.1758	1.0090	1.1075
	(0.0028)	(0.0102)	(0.0032)
wave_2	-0.4635	-0.4671	-0.4653
	(2.372e-09)	(5.953e-11)	(6.232e-11)
wave_3	-0.1952	-0.2638	-0.2396
	(0.0195)	(0.0007)	(0.0019)
==========	==========	==========	==========
Effects		Entity	

P-values reported in parentheses

P-values reportados en los parentesis:

- Para la variabkle 'child', a partir de los modelos FE y RE; a comparación con el modelo Pooled OLS, pasó de ser significativo a no serlo.
- El efecto de 'Drinkly' no varía demasiado y para los 3 modelos resultó ser significativo.
- 'hsize' para Pooled OLS no era significativo y tanto en FE como en RE resultó sí serlo.
- La variable 'married' es la que mayor efecto tiene en la variable resultado, en los tres modelos.
- 'urban' permaneció siendo significativo al 99% de confianza en los tres modelos, sin embargo entre ellos el valor de su efecto es muy variable.
- En los tres modelos, 'wealth' fue significativo pero su efecto es practicamente nulo en la variable resultado.
- Tanto 'intmonth_2' como 'intmonth_3' resultaron ser significativas en los tres modelos.
- Las variables 'wave_2' y 'wave_3' son significativas en todos los modelos, e indican que el efecto que tiene tanto estar en el periodo 2 como en el periodo 3 hace disminuir el valor de la

variable resultado, esto esto, el puntaje de la escala de salud mental.

[14]:
```
def hausman(fe, re):
 diff = fe.params-re.params
 psi = fe.cov - re.cov
 dof = diff.size -1
 W = diff.dot(la.inv(psi)).dot(diff)
 pval = stats.chi2.sf(W, dof)
 return W, dof, pval


htest = hausman(fe, re)
print("Hausman Test: chi-2 = {0}, df = {1}, p-value = {2}".format(htest[0],␣
 ↪htest[1], htest[2]))
```

```
Hausman Test: chi-2 = 25.560123100482496, df = 12, p-value =
0.012380618444528552
```

El rechazo del test de Hausman al 95% de confianza refleja que el supuesto que se utiliza en el modelo RE, es decir, $Cov(x_{itk}, \mu_i) = 0, \forall t \in \{1, 2, 3\}, \forall k \in \{1, .., 12\}$ es incorrecto, favoreciendo el modelo de FE por sobre el RE.

Por lo tanto, dado el resultado del test de Hausman y el valor p entre el modelo FE y Pooled OLS, se concluye que el modelo más significativo es el Modelo de efectos fijos para este set de variables explicativas.

1.6 Modelo de efectos aleatorios correlacionados

[15]:
```
X=Xc[['child','Drinkly','hrsusu','hsize','married','schadj','urban','wealth','intmonth_2','int
X=sm.add_constant(X)
model=lmp.RandomEffects(y,X)
cre=model.fit(cov_type="robust")
print(cre)
```

```
                        RandomEffects Estimation Summary
================================================================================
Dep. Variable:                    cesd   R-squared:                        0.1848
Estimator:                RandomEffects   R-squared (Between):              0.1050
No. Observations:                 34352   R-squared (Within):               0.0416
Date:                Wed, Oct 05 2022   R-squared (Overall):              0.0636
Time:                        13:41:59   Log-likelihood                -1.072e+05
Cov. Estimator:                Robust
                                        F-statistic:                      324.27
Entities:                        3459   P-value                           0.0000
Avg Obs:                       9.9312   Distribution:                F(24,34327)
Min Obs:                       2.0000
Max Obs:                       468.00   F-statistic (robust):             76.176
                                        P-value                           0.0000
Time periods:                       3   Distribution:                F(24,34327)
```

```
Avg Obs:                    1.145e+04
Min Obs:                    1.144e+04
Max Obs:                    1.146e+04
```

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	15.143	7.4586	2.0303	0.0423	0.5239	29.762
child	0.0052	0.0307	0.1698	0.8652	-0.0550	0.0655
Drinkly	-0.8575	0.0740	-11.583	0.0000	-1.0026	-0.7124
hrsusu	-0.0102	0.0013	-7.9578	0.0000	-0.0127	-0.0077
hsize	-0.0905	0.0223	-4.0582	0.0000	-0.1342	-0.0468
married	-1.3545	0.1294	-10.471	0.0000	-1.6081	-1.1010
schadj	-0.2090	0.0106	-19.646	0.0000	-0.2299	-0.1882
urban	-0.6131	0.1097	-5.5902	0.0000	-0.8281	-0.3981
wealth	-2.616e-06	4.833e-07	-5.4133	0.0000	-3.563e-06	-1.669e-06
intmonth_2	1.2413	0.3821	3.2484	0.0012	0.4923	1.9903
intmonth_3	1.0090	0.4015	2.5132	0.0120	0.2221	1.7959
wave_2	-0.4671	0.0711	-6.5715	0.0000	-0.6064	-0.3278
wave_3	-0.2638	0.0776	-3.4013	0.0007	-0.4158	-0.1118
mchild	0.0066	0.0659	0.1007	0.9198	-0.1226	0.1359
mDrinkly	-0.1482	0.1974	-0.7508	0.4528	-0.5352	0.2387
mhrsusu	0.0125	0.0039	3.1650	0.0016	0.0047	0.0202
mhsize	0.0630	0.0570	1.1050	0.2692	-0.0487	0.1747
mintmonth_2	1.2277	1.3667	0.8983	0.3690	-1.4511	3.9065
mintmonth_3	2.3206	1.4256	1.6279	0.1036	-0.4735	5.1147
mmarried	-0.1262	0.2853	-0.4423	0.6583	-0.6853	0.4329
mschadj	-0.0094	0.0237	-0.3973	0.6912	-0.0559	0.0371
murban	-1.0331	0.1931	-5.3509	0.0000	-1.4115	-0.6546
mwealth	-1.106e-05	1.594e-06	-6.9343	0.0000	-1.418e-05	-7.931e-06
mwave_2	-11.964	9.4935	-1.2602	0.2076	-30.571	6.6438
mwave_3	-4.3575	13.419	-0.3247	0.7454	-30.660	21.945

[16]:
```
print(lmp.compare({"FE": fe, "RE": re, "CRE": cre}, precision="pvalues"))
```

Model Comparison

	FE	RE	CRE
Dep. Variable	cesd	cesd	cesd
Estimator	PanelOLS	RandomEffects	RandomEffects
No. Observations	34352	34352	34352
Cov. Est.	Robust	Robust	Robust
R-squared	0.0416	0.1831	0.1848
R-Squared (Within)	0.0416	0.0412	0.0416
R-Squared (Between)	0.0681	0.0861	0.1050

R-Squared (Overall)	0.0588	0.0553	0.0636
F-statistic	111.78	641.31	324.27
P-value (F-stat)	0.0000	0.0000	0.0000
==========	==========	==========	==========
const	10.468	10.976	15.143
	(0.0000)	(0.0000)	(0.0423)
child	0.0052	0.0116	0.0052
	(0.8707)	(0.6708)	(0.8652)
Drinkly	-0.8575	-0.8610	-0.8575
	(0.0000)	(0.0000)	(0.0000)
hrsusu	-0.0102	-0.0090	-0.0102
	(1.066e-14)	(9.881e-14)	(1.776e-15)
hsize	-0.0905	-0.0802	-0.0905
	(7.756e-05)	(8.931e-05)	(4.956e-05)
married	-1.3545	-1.3647	-1.3545
	(0.0000)	(0.0000)	(0.0000)
schadj	-0.2090	-0.2141	-0.2090
	(0.0000)	(0.0000)	(0.0000)
urban	-0.6131	-0.8986	-0.6131
	(1.136e-07)	(0.0000)	(2.285e-08)
wealth	-2.616e-06	-3.255e-06	-2.616e-06
	(1.594e-07)	(3.084e-11)	(6.229e-08)
intmonth_2	1.2413	1.2604	1.2413
	(0.0009)	(0.0004)	(0.0012)
intmonth_3	1.0090	1.1075	1.0090
	(0.0102)	(0.0032)	(0.0120)
wave_2	-0.4671	-0.4653	-0.4671
	(5.953e-11)	(6.232e-11)	(5.051e-11)
wave_3	-0.2638	-0.2396	-0.2638
	(0.0007)	(0.0019)	(0.0007)
mchild			0.0066
			(0.9198)
mDrinkly			-0.1482
			(0.4528)
mhrsusu			0.0125
			(0.0016)
mhsize			0.0630
			(0.2692)
mintmonth_2			1.2277
			(0.3690)
mintmonth_3			2.3206
			(0.1036)
mmarried			-0.1262
			(0.6583)
mschadj			-0.0094
			(0.6912)
murban			-1.0331
			(8.808e-08)

```
mwealth                                                          -1.106e-05
                                                                (4.154e-12)
mwave_2                                                             -11.964
                                                                   (0.2076)
mwave_3                                                             -4.3575
                                                                   (0.7454)
====================== ============== ================ =================
Effects                          Entity
-------------------------------------------------------------------------

P-values reported in parentheses
```

Lo realmente interesante es que si se comparan todas las variables explicativas entre los modelos FE y CRE, son casi idénticos; por lo más probable es que esto se produjo debido a que no se usaron las variables explicativas que se usaron en los modelos en FE, RE y CRE no son constantes en el tiempo.

A continuación se incluirán las variables 'age' y 'female' en RE y CRE:

[17]:
```
X=Xc[['child','Drinkly','hrsusu','hsize','married','schadj','urban','wealth','intmonth_2','int
X=sm.add_constant(X)
model=lmp.RandomEffects(y,X)
re=model.fit(cov_type="robust")

X=Xc[['child','Drinkly','hrsusu','hsize','married','schadj','urban','wealth','intmonth_2','int
X=sm.add_constant(X)
model=lmp.RandomEffects(y,X)
cre=model.fit(cov_type="robust")

print(lmp.compare({"FE": fe, "RE": re, "CRE": cre}, precision="pvalues"))
```

```
                                Model Comparison
=============================================================================
                                    FE                RE               CRE
-----------------------------------------------------------------------------
Dep. Variable                     cesd              cesd              cesd
Estimator                     PanelOLS     RandomEffects     RandomEffects
No. Observations                 34352             34352             34352
Cov. Est.                       Robust            Robust            Robust
R-squared                       0.0416            0.1923            0.1941
R-Squared (Within)              0.0416            0.0503            0.0507
R-Squared (Between)             0.0681            0.1020            0.1221
R-Squared (Overall)             0.0588            0.0640            0.0730
F-statistic                     111.78            583.96            295.30
P-value (F-stat)                0.0000            0.0000            0.0000
====================     =============   ===============   ===============
const                           10.468            9.9755            14.910
                               (0.0000)          (0.0000)          (0.0400)
child                           0.0052            0.0491            0.0345
```

	(0.8707)	(0.1123)	(0.3230)
Drinkly	-0.8575	-0.2680	-0.2889
	(0.0000)	(0.0005)	(0.0004)
hrsusu	-0.0102	-0.0081	-0.0091
	(1.066e-14)	(9.515e-11)	(4.579e-12)
hsize	-0.0905	-0.0934	-0.1026
	(7.756e-05)	(5.548e-06)	(4.279e-06)
married	-1.3545	-1.3102	-1.2903
	(0.0000)	(0.0000)	(0.0000)
schadj	-0.2090	-0.1668	-0.1597
	(0.0000)	(0.0000)	(0.0000)
urban	-0.6131	-0.9845	-0.6971
	(1.136e-07)	(0.0000)	(1.797e-10)
wealth	-2.616e-06	-3.302e-06	-2.639e-06
	(1.594e-07)	(3.274e-11)	(6.818e-08)
intmonth_2	1.2413	1.2344	1.2418
	(0.0009)	(0.0005)	(0.0011)
intmonth_3	1.0090	1.1049	1.0269
	(0.0102)	(0.0032)	(0.0103)
wave_2	-0.4671	-0.4721	-0.4728
	(5.953e-11)	(2.718e-11)	(2.447e-11)
wave_3	-0.2638	-0.2490	-0.2705
	(0.0007)	(0.0013)	(0.0005)
age		-0.0039	-0.0011
		(0.4307)	(0.8389)
female		1.4134	1.4119
		(0.0000)	(0.0000)
mchild			0.0101
			(0.8914)
mDrinkly			0.1328
			(0.5487)
mhrsusu			0.0128
			(0.0019)
mhsize			0.0687
			(0.2363)
mintmonth_2			0.8168
			(0.5502)
mintmonth_3			2.0177
			(0.1559)
mmarried			-0.0885
			(0.7605)
mschadj			-0.0083
			(0.7376)
murban			-1.0366
			(7.098e-08)
mwealth			-1.154e-05
			(6.286e-13)
mwave_2			-13.205

```
                                                                (0.1378)
mwave_3                                                         -5.7318
                                                                (0.6641)
mage                                                             0.0011
                                                                (0.9285)
mfemale                                                          0.1790
                                                                (0.3730)
====================== ============== ================ =================
Effects                           Entity
-------------------------------------------------------------------------
```

P-values reported in parentheses

Comparando nuevamente los coeficientes entre FE y CRE: - Tanto los modelos CRE como en RE consideran que la variable 'age' no es significativa, en cambio, 'female' si lo es, y además es la que más aporta de todas las variables en el efecto de la variable resultado. - Los coeficientes entre el modelo FE y CRE no son iguales como en la comparación anterior, sin embargo, los valores de los coef., el signo y el valor de los efectos se aproximan bastante, y como el modelo CRE incluye variables explicativas que son constantes en el tiempo, se concluye finalmente que para esta base de datos y de entre estos 4 modelos (Pooled OLS, FE, RE y CRE) el que mejor ajusta los datos de panel es el modelo de Efectos Aleatorios Correlacionados (CRE).