Tarea2_Almonacid_Gonzalez

October 14, 2022

1 Tarea 2

Autores

- José González Cortés
- Felipe Almonacid Contreras

2 Preparacion de librerías y funciones

[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 numpy.linalg as la
from scipy import stats

def hausman(fe, re):
    try:
        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
    except Exception as e:
        return None,None,None
%matplotlib inline
```

3 Pregunta 1: Análisis de Datos

Cargar la base de datos *charls.csv* en el ambiente. Identifique los tipos de datos que se encuentran en la base, realice estadisticas descriptivas sobre las variables importantes

(Hint: Revisar la distribuciones, datos faltantes, outliers, etc.) y limpie las variables cuando sea necesario.

En esta sección se hizo un análisis de las variables que están presentes en la base de datos: #### Las variables tienen la siguiente descripcion:

- INID: identificador unico
- wave: periodo de la encuesta (1-3)
- cesd: puntaje en la escala de salud mental (0-30)
- child: numero de hijos
- drinkly: bebio alcohol en el ultimo mes (binario)
- hrsusu: horas promedio trabajo semanal
- hsize: tamano del hogar
- intmonth: mes en que fue encuestado/a (1-12)
- married: si esta casado/a (binario)
- retired: si esta pensionado/a (binario)
- schadj: años de escolaridad
- urban: zona urbana (binario)
- wealth: riqueza neta (miles RMB)
- age: edad al entrar a la encuesta (no varia entre periodos)

Algunas de estás variables presentaron valores que no correspondian para hacer la ejecución de los modelos por lo que hicieron correcciones de estos datos, estas correcciones fueron: - Se pudo observar que la base de datos presentaba identificadores repetidos una gran cantidad de veces, especificamente desde el dato 10,058, para la resolusión de esta problematica se eliminaron todos los "inid" mayores al identificador del dato 10,058. - El caso de que en la variable drinkly se almacenaban string (0.Nones y 1.Yes), además tenia un caracter no correspondiente que era ".m:missing", para arreglar esto se tuvo que eliminar los individuos que presentaban el caracter ".m:missing" y posteriormente a eso se transformó "0.Nones" a int(0) y "1.Yes" a int(1).

[2]:
```
#Se cargan los datos desde los archivos csv
charls = pd.read_csv('../../data/charls.csv')
charls.dropna(inplace=True)

#Se limpian los valores no válidos en la base de datos
indexNames = charls[charls['inid'] > 94004308001].index #Los al datos que␣
 ↪tienen mal asignado el inid
charls.drop(indexNames , inplace=True)
indexNames = charls[charls['drinkly'] == ".m:missing"].index #Existía almenos␣
 ↪un valor que contenia este dato
charls.drop(indexNames , inplace=True)
value = charls.iloc[indexNames]["inid"]
for v in value.array:
    indexNames = charls[charls['inid'] == v].index
    charls.drop(indexNames , inplace=True)


#Transformamos la variable Drinkly
def funct(s):
```

```
    return int(s.split(".")[0])
charls["drinkly"] = charls["drinkly"].map(funct)

charls.reset_index(drop=True, inplace=True)
```

3.1 Sanity Check

Se comprueban todos los valores unicos presentes en cada columna, para confirmar que no existan valores invalidos.

[3]:
```
for c in charls:
    columna = charls[c].unique()
    columna.sort()
    print(f"{c}: {columna}\n")
```

```
cesd: [ 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
 24 25 26 27 28 29 30]

child: [ 0  1  2  3  4  5  6  7  8  9 10]

drinkly: [0 1]

female: [0 1]

hrsusu: [  0.    1.    2.    3.    4.    4.5   5.    6.    7.    8.    9.   10.
  12.   14.   15.   16.   17.5  18.   20.   21.   22.   22.5  24.   24.5
  25.   25.5  26.   27.   27.5  28.   30.   31.5  32.   32.5  33.   35.
  36.   38.5  40.   42.   42.5  44.   45.   48.   49.   50.   52.5  54.
  55.   56.   59.5  60.   62.5  63.   65.   66.   66.5  70.   72.   73.5
  77.   78.   80.5  84.   90.   91.   96.   98.  105.  112.  119.  120.
 126.  133.  140.  144.  154.  161.  168. ]

hsize: [ 1  2  3  4  5  6  7  8  9 10 11 12 13]

inid: [1.0104101e+10 1.0104101e+10 1.0104102e+10 … 9.4004303e+10 9.4004303e+10
 9.4004308e+10]

intmonth: [ 1  2  6  7  8  9 10 11 12]

married: [0 1]

retired: [0 1]

schadj: [ 0  4  8 12 14 16]

urban: [0 1]

wave: [1 2 3]
```

```
wealth: [-975000. -596000. -499500. … 1040000. 1479000. 8001500.]

age: [16 17 26 27 29 31 32 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74
 75 76 77 78 79 80 81 82 83 84 86 87 89]
```

Al notar que todos los valores están correctos en cada variables, se prosiguió con la contrucción del panel de datos, considerando que para las medias de algunas variables, se eliminaron las que parecen ser constantes en el tiempo.

[4]:
```
# Se construyen las variables para el analisis de panel
X=charls[["cesd","child","drinkly","hrsusu",
          "hsize","intmonth","married",
          "retired","schadj","urban","wealth",
          "age","female"]]
#Se agrupan todos los valores por el identificador
Xm=(X.groupby(charls['inid']).transform('mean'))
del Xm["age"]
del Xm["female"]
del Xm["schadj"]
#Se eliminaron algunas variables constantes para no ser incluidas en las medias
Xinid=charls[['inid','wave',"cesd","child",
              "drinkly","hrsusu","hsize",
              "intmonth","married","retired",
              "schadj","urban","wealth","age",
              "female"]]
Xc = pd.DataFrame(
    np.c_[Xinid, Xm],
    columns=['inid','wave',"cesd","child",
             "drinkly","hrsusu","hsize",
             "intmonth","married","retired",
             "schadj","urban","wealth","age",
             "female","mcesd","mchild",
             "mdrinkly","mhrsusu","mhsize",
             "mintmonth","mmarried","mretired",
             "murban","mwealth"])
#Estructura del panel
Xc = Xc.set_index(["inid","wave"])
Xc.describe()
```

[4]:

	cesd	child	drinkly	hrsusu	hsize	\
count	10028.000000	10028.000000	10028.000000	10028.000000	10028.000000	
mean	8.860690	2.768049	0.324491	27.936528	3.653171	
std	6.288328	1.435647	0.468208	27.247127	1.784314	
min	0.000000	0.000000	0.000000	0.000000	1.000000	
25%	4.000000	2.000000	0.000000	0.000000	2.000000	

```
50%        8.000000      2.000000      0.000000     24.000000      3.000000
75%       13.000000      3.000000      1.000000     49.000000      5.000000
max       30.000000     10.000000      1.000000    168.000000     13.000000

           intmonth       married       retired        schadj         urban  \
count  10028.000000  10028.000000  10028.000000  10028.000000  10028.000000
mean       7.591943      0.858097      0.269047      4.097128      0.315716
std        1.100524      0.348968      0.443486      3.605418      0.464824
min        1.000000      0.000000      0.000000      0.000000      0.000000
25%        7.000000      1.000000      0.000000      0.000000      0.000000
50%        7.000000      1.000000      0.000000      4.000000      0.000000
75%        8.000000      1.000000      1.000000      4.000000      1.000000
max       12.000000      1.000000      1.000000     16.000000      1.000000

       …         mcesd        mchild      mdrinkly       mhrsusu  \
count  …  10028.000000  10028.000000  10028.000000  10028.000000
mean   …      8.860690      2.768049      0.324491     27.936528
std    …      5.125862      1.360613      0.406834     21.278312
min    …      0.000000      0.000000      0.000000      0.000000
25%    …      5.000000      2.000000      0.000000      8.333333
50%    …      8.000000      2.333333      0.000000     28.000000
75%    …     12.000000      3.333333      0.666667     43.333333
max    …     28.666667      9.666667      1.000000    119.000000

            mhsize     mintmonth      mmarried      mretired        murban  \
count  10028.000000  10028.000000  10028.000000  10028.000000  10028.000000
mean       3.653171      7.591943      0.858097      0.269047      0.315716
std        1.459551      0.630634      0.332878      0.365991      0.464824
min        1.000000      5.000000      0.000000      0.000000      0.000000
25%        2.333333      7.333333      1.000000      0.000000      0.000000
50%        3.666667      7.666667      1.000000      0.000000      0.000000
75%        4.666667      8.000000      1.000000      0.333333      1.000000
max       10.000000     10.000000      1.000000      1.000000      1.000000

            mwealth
count  1.002800e+04
mean   1.017937e+04
std    6.285634e+04
min   -3.250000e+05
25%    8.333333e+01
50%    1.070000e+03
75%    8.616667e+03
max    2.672550e+06

[8 rows x 23 columns]
```

Para la implementación de los modelos, primero se descutió de las variables que se deben considerar

para el desarrollo. Por ende las variables que consideramos que deben estar y las que no se presentan a continuación:

3.2 Variables que consideramos que si deberían ir:

- child:
- hrsusu
- married
- retired
- schadj
- urban
- wealth
- age
- female

3.3 Variables que no consideramos que deberían ir:

- drinkly; saber si un encuestado ingirió alcohol el mes pasado no es una variable acorde a lo que se esta prediciendo, en caso de que la variable fuese cuanto alcohol ingiere frecuentemente en un mes sería una varible acorde para determinar el nivel de salud mental de las personas.
- hsize; esta variable no debe tener mayor peso dentro de la variable a estimar, ya que, el tamaño de un hogar puede ser delimitado, además esta muy relacionado con que la persona viva en zona urbana o no.
- intmonth; en caso de los meses se considera que no es importante para el modelo debido a que se expresa como una variable cualitativa, ya que, los encuestados si pueden variar su salud mental pero no es un factor que al aumentar o disminuir los meses afecte la salud mental.

4 Pregunta 2: Modelo de Minimo Cuadrado Agrupado

Ejecute un modelo Pooled OLS para explicar el puntaje en la escala de salud mental (CESD). Seleccione las variables dependientes a incluir en el modelo final e interprete su significad

[5]:
```
y = Xc["cesd"]
X = Xc[["child","hrsusu","married","retired",
        "schadj","urban","wealth","age",
        "female"]]
X = sm.add_constant(X)
model = lmp.PooledOLS(y,X)
OLS = model.fit(cov_type="robust")
final_model = OLS
print(OLS)
```

```
                          PooledOLS Estimation Summary
================================================================================
Dep. Variable:                    cesd   R-squared:                        0.0716
Estimator:                  PooledOLS   R-squared (Between):              0.1078
No. Observations:               10028   R-squared (Within):               0.0003
```

Date:	Wed, Oct 05 2022	R-squared (Overall):	0.0716
Time:	13:41:50	Log-likelihood	-3.229e+04
Cov. Estimator:	Robust		
		F-statistic:	85.815
Entities:	3345	P-value	0.0000
Avg Obs:	2.9979	Distribution:	F(9,10018)
Min Obs:	2.0000		
Max Obs:	3.0000	F-statistic (robust):	83.767
		P-value	0.0000
Time periods:	3	Distribution:	F(9,10018)
Avg Obs:	3342.7		
Min Obs:	3341.0		
Max Obs:	3345.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	10.754	0.5808	18.515	0.0000	9.6157	11.893
child	0.0765	0.0494	1.5491	0.1214	-0.0203	0.1733
hrsusu	0.0060	0.0028	2.1450	0.0320	0.0005	0.0115
married	-1.4550	0.2004	-7.2615	0.0000	-1.8478	-1.0623
retired	0.5049	0.1860	2.7141	0.0067	0.1402	0.8695
schadj	-0.1953	0.0182	-10.721	0.0000	-0.2310	-0.1596
urban	-1.7967	0.1323	-13.578	0.0000	-2.0560	-1.5373
wealth	-2.343e-06	1.968e-06	-1.1905	0.2339	-6.201e-06	1.515e-06
age	-0.0103	0.0084	-1.2159	0.2241	-0.0268	0.0063
female	1.5284	0.1315	11.626	0.0000	1.2707	1.7861

5 Análisis Pooled OLS

Al realizar un modelo Pooled OLS para estimar el puntaje en escala de salud mental, se logra determinar que el modelo posee un valor-p de 0.000 y un valor F-statistic de 84.050, co

Las variables que son significativas para el modelo son las siguientes: - child: El valor-p (0.0999) de este parametro esta muy al limite de ser significativa o no para el modelo,además de que sus posibles valores del parametro pueden pasar por el cero, pero se estima que tener hijos es un factor importante para determinar la salud mental de los encuestados. Por otro lado,este parametro presenta un valor de 0,0812, lo que quiere decir que al variar en 1 la cantidad de hijos afecta en 0,0812 el puntaje de salud mental. Es importante mencionar que al aumentar child, se empeora la salud mental de las personas, ya que, aumenta el cesd subiendo su puntaje. - hrsusu: Las horas laborales son significativas para el modelo con un valor-p de 0.0338. Por otro lado, el valor de este parametro es de 0.0060 lo que significa que el cambio marginal es muy bajo. - married: Esta varible originalmente es binaria y el parametro estimado presenta un valor de -1.4475, lo que quiere decir que si el individuo esta casado (1) el puntaje de cesd disminuye y si el individuo tiene un estado civil de soltero (0) implica que este parametro desaparece. - retired: Se puede concluir que estar retirado es un factor que si es significativo pero no fundamental (valor-p: 0.0089) para determinar

el puntaje de salud mental, en este caso el parametro presentó un valor de 0.4868 lo que quiere decir que si el individuo se encuentra retirado su cesd aumenta en 0.4868 pero en caso contrario el parametro desaparece. - schadj: En el caso de los años de escolaridad, se puede decir que es fundamental para estimar el cesd. Este parametro presentó un valor de -0.1951, es decir, que al variar en 1 este parametro va a afectar en -0.1951 el cesd, además se logra concluir que a mayor grado de escolaridad menor es cesd por lo que es mejor salud mental. - urban: Esta variable es significativa al modelo y cuyo valor que posee el parametro es -1.798, es decir, que si el individuo vive en zona urbana el valor de cesd disminuye lo que implica una mejor salud mental, pero en caso de que no viva en zona urbana este parametro desaparece. - female: Ser mujer es significativo para el modelo otorgandoles un parametro de 1.5347, es decir, que si el encuestado es mujer (1) el resultado de cesd aumenta en 1.5347 pero en caso contrario (0) desaparece el parametro.

Las variables que no presentaron significancia para el modelo son el siguiente: - wealth (valor-p: 0.2340): A pesar de que la riqueza neta es un factor que influye en el día a día de las personas es un factor que no influyente para el modelo, esto puede ser debido a la gran cartera de valores que esta variable puede adquirir, además que los valores de cesd presentes en la base de datos que se otrogó no siguen un aumento o disminución que dependa de la riqueza neta. - age (valor-p: 0.2469): La edad es una variable que tiene poca relación con la salud mental y el modelo avala esa noción, ya que presento un parametro sumamente bajo para el modelo y su valor-p es el más alto de todos por lo que no es significativo.

6 Pregunta 3: Modelo de Efectos Fijos

Ejecute un modelo de efectos fijos para explicar el puntaje en la escala de salud mental (CESD). Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

[6]:
```
y = Xc["cesd"]
X = Xc[["child","hrsusu","married","retired","wealth"]]
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:                        0.0026
Estimator:                   PanelOLS   R-squared (Between):              0.0178
No. Observations:               10028   R-squared (Within):               0.0026
Date:                Wed, Oct 05 2022   R-squared (Overall):              0.0127
Time:                        13:41:50   Log-likelihood                -2.718e+04
Cov. Estimator:              Robust
                                        F-statistic:                      3.4878
Entities:                        3345   P-value                           0.0038
Avg Obs:                       2.9979   Distribution:                  F(5,6678)
Min Obs:                       2.0000
Max Obs:                       3.0000   F-statistic (robust):             2.8075
                                        P-value                           0.0154
```

```
Time periods:                       3   Distribution:                  F(5,6678)
Avg Obs:                       3342.7
Min Obs:                       3341.0
Max Obs:                       3345.0

                            Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
const          9.4565     0.5267     17.954     0.0000      8.4240      10.489
child          0.1446     0.0958     1.5087     0.1314     -0.0433      0.3324
hrsusu         0.0001     0.0029     0.0354     0.9717     -0.0056      0.0058
married       -1.2790     0.5045    -2.5349     0.0113     -2.2681     -0.2899
retired        0.3851     0.2029     1.8982     0.0577     -0.0126      0.7828
wealth      -4.84e-07  8.656e-07    -0.5591     0.5761  -2.181e-06   1.213e-06
==============================================================================

F-test for Poolability: 3.8655
P-value: 0.0000
Distribution: F(3344,6678)

Included effects: Entity
```

7 Análisis FE

Se realizó un modelo de Efectos Fijos para estimar el puntaje en escala de salud mental (cesd). El modelo posee un valor-p de 0.0026 y un valor F-statistic de 3.4547 concluyendo que el modelo explica muy levemente la variable cesd con respecto a las varibles independientes, pero es peor modelo que el Pooled OLS por las pocas variables que se utilizan.

Las variables que son significativas para este modelo son las siguientes: - married: Ser casado es una variables que posee un valor-p de 0.0112, siendo un valor significativo pero que esta al margen para el modelo, por eso se asigna un parametro de -1.2789, es decir, que ser casado (1) baja en un -1.2789 el valor de cesd y en caso contrario (0) desaparece.

- retired: Estar retirado es una variables que posee un valor-p de 0.0748, siendo un valor significativo pero que esta al margen para el modelo, por eso se asigna un parametro de 0.3615, es decir, que estar retirado (1) baja en un 0.3615 el valor de cesd y en caso contrario (0) desaparece. Cabe mencionar, que el parametro puede tener un valor de 0 en algún momento por lo que hace menos significativa la variable.

Las varibles que no presentan siginificancia para el modelo son las siguientes: - child: La cantidad de hijos no es influyente para este modelo, ya que presenta un valor-p de 0.1158 y su parametro puede ser 0. Por otra parte, se puede decir que a mayor cantidad de hijso implica un mayor valor de cesd por lo que peor la salud mental. - hrsusu: Para este modelo, las horas semanales promedio de trabajo es una varible que no posee nada de significancia, esto es por su valor-p excesivamente alto (0.9927) por lo que su parametros es bajo. - wealth: Por último, la variable de riqueza neta posee un valor-p de 0.5952 y un parametro demasiado bajo por lo que no es para nada significativo para el modelo

Cabe destacar que para realizar este modelo se necesita eliminar las variables schadj, urban, age y female, ya que presentaban un error de colinealidad, esto es debido a que son variables que se mantienen constante por cada individuos en los tres periodos que fueron entrevistados. Además desde un punto de vista teórico, el modelo no considera suficientes variables para ser útil.

8 Pregunta 4: Model de Efectos Aleatorios

Ejecute un modelo de efectos aleatorios para explicar el puntaje en la escala de salud mental (CESD). Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

[7]:
```
X = Xc[["child","hrsusu","married",
        "retired","schadj","urban",
        "wealth","age","female"]]
X = sm.add_constant(X)
model=lmp.RandomEffects(y,X)
re=model.fit(cov_type="robust")
print(re)
```

```
                        RandomEffects Estimation Summary
================================================================================
Dep. Variable:                     cesd   R-squared:                        0.0395
Estimator:                RandomEffects   R-squared (Between):              0.1061
No. Observations:                 10028   R-squared (Within):               0.0022
Date:                Wed, Oct 05 2022     R-squared (Overall):              0.0711
Time:                        13:41:50     Log-likelihood                -2.922e+04
Cov. Estimator:                Robust
                                          F-statistic:                      45.813
Entities:                        3345     P-value                           0.0000
Avg Obs:                       2.9979     Distribution:                 F(9,10018)
Min Obs:                       2.0000
Max Obs:                       3.0000     F-statistic (robust):             46.133
                                          P-value                           0.0000
Time periods:                       3     Distribution:                 F(9,10018)
Avg Obs:                       3342.7
Min Obs:                       3341.0
Max Obs:                       3345.0


                             Parameter Estimates
================================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
--------------------------------------------------------------------------------
const          10.979     0.7556     14.530     0.0000      9.4979      12.460
child          0.0942     0.0590     1.5973     0.1102     -0.0214      0.2098
hrsusu         0.0027     0.0026     1.0447     0.2962     -0.0024      0.0077
married       -1.4255     0.2486    -5.7333     0.0000     -1.9129     -0.9381
retired        0.4213     0.1749     2.4092     0.0160      0.0785      0.7640
schadj        -0.1982     0.0245    -8.0771     0.0000     -0.2463     -0.1501
```

```
urban          -1.8047     0.1820    -9.9168     0.0000     -2.1614     -1.4479
wealth     -1.211e-06  1.064e-06    -1.1383     0.2550  -3.297e-06   8.746e-07
age            -0.0132     0.0110    -1.1970     0.2313     -0.0347      0.0084
female          1.5074     0.1803     8.3605     0.0000      1.1539      1.8608
==============================================================================
```

9 Análisis RE

Se logra determinar que el modelo posee un valor-p de 0.0394 y un valor F-statistic de 45.739 concluyendo que el modelo explica levemente la variable cesd con respecto a las varibles independientes seleccionadas.

Las variables que son significativas para el modelo son las siguientes: - child: El valor-p (0.0942) de este parametro esta muy al limite de ser significativa o no para el modelo,además de que sus posibles valores del parametro pueden pasar por el cero, pero se estima que tener hijos es un factor importante para determinar la salud mental de los encuestados. Por otro lado,este parametro presenta un valor de 0,0993, lo que quiere decir que al variar en 1 la cantidad de hijos afecta en 0,0993 el puntaje de salud mental. Es importante mencionar que al aumentar número de hijos, se empeora la salud mental de las personas, ya que, aumenta el cesd subiendo su puntaje. - married: Esta varible originalmente es binaria y el parametro estimado presenta un valor de -1.4255, lo que quiere decir que si el individuo esta casado (1) el puntaje de cesd disminuye y si el individuo tiene un estado civil de soltero (0) implica que este parametro desaparece. - retired: Se puede concluir que estar retirado es un factor que si es significativo pero no fundamental (valor-p: 0.0160) para determinar el puntaje de salud mental, en este caso el parametro presentó un valor de 0.4213 lo que quiere decir que si el individuo se encuentra retirado su salud mental aumenta en 0.4868 pero en caso contrario el parametro desaparece. - schadj: En el caso de los años de escolaridad, se puede decir que es fundamental para estimar el cesd. Este parametro presentó un valor de -0.1978, es decir, que al variar en 1 este parametro va a afectar en -0.1978 el cesd, además se logra concluir que a mayor grado de escolaridad menor es cesd por lo que es mejor salud mental. - urban: Esta variable es significativa al modelo y cuyo valor que posee el parametro es -1.8060, es decir, que si el individuo vive en zona urbana el valor de cesd disminuye lo que implica una mejor salud mental, pero en caso de que no viva en zona urbana este parametro desaparece. - female: Ser mujer es significativo para el modelo otorgandoles un parametro de 1.5078, es decir, que si el encuestado es mujer (1) el resultado de cesd aumenta en 1.5078 pero en caso contrario (0) desaparece el parametro.

Las variables que no presentaron significancia para el modelo son el siguiente: - hrsusu (valor-p: 0.3446): En este modelo las horas laborales no son significativas por el valor-p que este parametro presenta. Por otro lado, el valor de este parametro es de 0.0027 por lo que afecta muy poco o casi nada a la variable predictiva cuando hrsusu varie en 1. - wealth (valor-p: 0.2571): A pesar de que la riqueza neta es un factor que influye en el día a día de las personas es un factor que no influyente para el modelo, esto puede ser debido a la gran cartera de valores que esta variable puede adquirir, además que los valores de cesd presentes en la base de datos que se otrogó no siguen un aumento o disminución que dependa de la riqueza neta. - age (valor-p: 0.2469): La edad es una variable que tiene poca relación con la salud mental y el modelo avala esa noción, ya que presento un parametro sumamente bajo para el modelo y su valor-p es alto de por lo que no es significativo.

Cabe destacar que los valores de los parametros de este modelo son muy similares con los del Pooled OLS, aunque varian en la cantidad de variables significativas, además el valor R-squared y F-estatistic del modelo RE es casi la mitad de los respectivos valores del modelo Pooled OLS. Todo

lo anterior es debido a que los dos modelos evaluan las mismas variables.

10 Pregunta 5: Comparativa

5. Comente los resultados obtenidos en 2, 3 y 4. ¿Cuáles y por qué existen las diferencias entre los resultados?. En su opinión, ¿Cuál sería el más adecuado para responder la pregunta de investgación y por qué? ¿Qué variables resultaron ser robustas a la especificación?

10.1 Comparativa de modelos

Para la realizar un análisis profundo de los modelos, se hicieron tres tablas comparativas. La primera tabla comparativa se hizo con respectos a los 3 modelos donde los modelos Pooled OLS y RE presentan las misma cantidad de variables (9) mientras que el modelo FE cuenta con menos variables (5), la segunda tabla compara los 3 modelos pero con la misma cantidad de variables en todos (estas variables son las del modelo FE) y por último se hace una comparativa del modelo Pooled OLS con la cantidad de variables original (9) y la reducida (5). ### Primera comparativa

[8]:
```
print(lmp.compare({"FE": fe, "RE": re, "Pooled": OLS}).summary)
```

```
                          Model Comparison
=========================================================================
                                   FE                RE         Pooled
-------------------------------------------------------------------------
Dep. Variable                    cesd              cesd           cesd
Estimator                   PanelOLS     RandomEffects      PooledOLS
No. Observations               10028             10028          10028
Cov. Est.                     Robust            Robust         Robust
R-squared                     0.0026            0.0395         0.0716
R-Squared (Within)            0.0026            0.0022         0.0003
R-Squared (Between)           0.0178            0.1061         0.1078
R-Squared (Overall)           0.0127            0.0711         0.0716
F-statistic                   3.4878            45.813         85.815
P-value (F-stat)              0.0038            0.0000         0.0000
====================     ===========   ===============   ============
const                         9.4565            10.979         10.754
                             (17.954)          (14.530)       (18.515)
child                         0.1446            0.0942         0.0765
                             (1.5087)          (1.5973)       (1.5491)
hrsusu                        0.0001            0.0027         0.0060
                             (0.0354)          (1.0447)       (2.1450)
married                       -1.2790           -1.4255        -1.4550
                            (-2.5349)         (-5.7333)      (-7.2615)
retired                       0.3851            0.4213         0.5049
                             (1.8982)          (2.4092)       (2.7141)
wealth                     -4.84e-07        -1.211e-06     -2.343e-06
                            (-0.5591)         (-1.1383)      (-1.1905)
schadj                                         -0.1982        -0.1953
                                              (-8.0771)      (-10.721)
```

```
urban                                          -1.8047        -1.7967
                                              (-9.9168)      (-13.578)
age                                            -0.0132        -0.0103
                                              (-1.1970)      (-1.2159)
female                                          1.5074         1.5284
                                               (8.3605)       (11.626)
====================== ============ ================= ==============
Effects                        Entity
-------------------------------------------------------------------------

T-stats reported in parentheses
```

10.1.1 Segunda comparativa

[9]:
```
y = Xc["cesd"]
X = Xc[["child","hrsusu","married",
        "retired","wealth"]]
X = sm.add_constant(X)
model = lmp.PanelOLS(y,X, entity_effects=True)
rfe = model.fit(cov_type="robust")
model=lmp.RandomEffects(y,X)
rre=model.fit(cov_type="robust")
model = lmp.PooledOLS(y,X)
ROLS = model.fit(cov_type="robust")
print(lmp.compare({"FE": rfe, "RE": rre, "Pooled": ROLS}))
```

```
                            Model Comparison
=========================================================================
                                   FE                RE         Pooled
-------------------------------------------------------------------------
Dep. Variable                    cesd              cesd           cesd
Estimator                   PanelOLS     RandomEffects      PooledOLS
No. Observations               10028             10028          10028
Cov. Est.                     Robust            Robust         Robust
R-squared                     0.0026            0.0091         0.0175
R-Squared (Within)            0.0026            0.0015        -0.0029
R-Squared (Between)           0.0178            0.0237         0.0278
R-Squared (Overall)           0.0127            0.0163         0.0175
F-statistic                   3.4878            18.470         35.748
P-value (F-stat)              0.0038            0.0000         0.0000
=====================     ===========   ===============   ============
const                         9.4565            9.6232         9.6960
                             (17.954)          (31.712)       (38.364)
child                         0.1446            0.2228         0.2523
                             (1.5087)          (4.0732)       (5.6311)
hrsusu                        0.0001            0.0016         0.0030
                             (0.0354)          (0.6068)       (1.0447)
married                       -1.2790           -1.6858        -1.8242
```

```
                            (-2.5349)         (-6.8349)      (-9.2437)
retired                       0.3851            0.1476        -0.0690
                             (1.8982)          (0.8550)      (-0.3740)
wealth                      -4.84e-07        -1.541e-06     -3.293e-06
                            (-0.5591)         (-1.3834)      (-1.4137)
======================= ============= ================= ==============
Effects                        Entity
-------------------------------------------------------------------------

T-stats reported in parentheses
```

[10]:
```
### Hussman Test
values = [rre,rfe,ROLS]
hmatrix = [[ hausman(i, j)[2] for i in values] for j in values]
#print(f"Hausman Test: chi-2 = {htest[0]}, df = {htest[1]}, p-value =␣
 ↪{htest[2]}")
print("Haussman p-value Matrix")
print(pd.DataFrame(hmatrix))
```

```
Haussman p-value Matrix
              0        1    2
0           NaN  0.12084  1.0
1  1.000000e+00      NaN  1.0
2  1.582774e-47  0.00002  NaN
```

Realizando el test de haussman contra todos los modelos, se puede observar que, los coeficientes de POLS son significativamente distintos a los de FE y RE.

Si revisamos FE Y RE con el test de haussman al compararlos RE y FE tienen un p-value de 0.12084, por lo que no es lo suficientemente pequeño para negar la hipotesis de que los estimadores son distintos, en este caso, elegimos el modelo que parece ser mas consistente, que en este caso es Random Effects, puesto que sus coeficientes son consistentes con los entregados por PooledOLS, además que no sufre problemas de colinealidad con variables que se consideran importantes en el modelo.

10.1.2 Tercera comparativa

[11]:
```
print(lmp.compare({"Pooled":OLS, "RPooled": ROLS}))
```

```
                Model Comparison
================================================
                             Pooled      RPooled
------------------------------------------------
Dep. Variable                  cesd         cesd
Estimator                PooledOLS    PooledOLS
No. Observations             10028        10028
Cov. Est.                   Robust       Robust
R-squared                   0.0716       0.0175
R-Squared (Within)          0.0003      -0.0029
```

```
R-Squared (Between)          0.1078       0.0278
R-Squared (Overall)          0.0716       0.0175
F-statistic                  85.815       35.748
P-value (F-stat)             0.0000       0.0000
==================== ============ ============
const                        10.754       9.6960
                           (18.515)     (38.364)
child                        0.0765       0.2523
                           (1.5491)     (5.6311)
hrsusu                       0.0060       0.0030
                           (2.1450)     (1.0447)
married                     -1.4550      -1.8242
                          (-7.2615)    (-9.2437)
retired                      0.5049      -0.0690
                           (2.7141)    (-0.3740)
schadj                      -0.1953
                          (-10.721)
urban                       -1.7967
                          (-13.578)
wealth                   -2.343e-06   -3.293e-06
                          (-1.1905)    (-1.4137)
age                         -0.0103
                          (-1.2159)
female                       1.5284
                           (11.626)
-----------------------------------------------

T-stats reported in parentheses
```

En todos los casos anteriores PooledOLS presenta una mejor descripcion de la variable cesd, con una mayor significancia y un estadistico F mayor, así como valores mas grandes de R-squared.

11 Pregunta 6: Modelo Efectos Aleatorios Correlacionados

Ejecute un modelo de efectos aleatorios correlacionados (CRE) para explicar el puntaje en la escala de salud mental (CESD). Seleccione las variables dependientes a incluir en el modelo final e interprete su significado. Es este modelo adecuado, dada la data disponible, para modelar el componente no observado?

[12]:
```
X = Xc[["child","hrsusu","married","retired",
        "schadj","murban","wealth",
        "mchild","mhrsusu","mmarried",
        "mretired","mwealth"]]
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:	0.0328
Estimator:	RandomEffects	R-squared (Between):	0.0879
No. Observations:	10028	R-squared (Within):	0.0026
Date:	Wed, Oct 05 2022	R-squared (Overall):	0.0592
Time:	13:41:50	Log-likelihood	-2.921e+04
Cov. Estimator:	Robust		
		F-statistic:	28.291
Entities:	3345	P-value	0.0000
Avg Obs:	2.9979	Distribution:	F(12,10015)
Min Obs:	2.0000		
Max Obs:	3.0000	F-statistic (robust):	28.783
		P-value	0.0000
Time periods:	3	Distribution:	F(12,10015)
Avg Obs:	3342.7		
Min Obs:	3341.0		
Max Obs:	3345.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	11.237	0.4151	27.070	0.0000	10.424	12.051
child	0.1446	0.1006	1.4369	0.1508	-0.0526	0.3417
hrsusu	0.0001	0.0029	0.0353	0.9718	-0.0056	0.0058
married	-1.2790	0.4854	-2.6350	0.0084	-2.2305	-0.3275
retired	0.3851	0.2023	1.9039	0.0569	-0.0114	0.7816
schadj	-0.2533	0.0243	-10.437	0.0000	-0.3009	-0.2057
murban	-1.6858	0.1880	-8.9661	0.0000	-2.0543	-1.3172
wealth	-4.84e-07	1.394e-06	-0.3472	0.7284	-3.216e-06	2.248e-06
mchild	-0.1290	0.1217	-1.0598	0.2893	-0.3676	0.1096
mhrsusu	0.0093	0.0064	1.4543	0.1459	-0.0032	0.0218
mmarried	-0.2090	0.5662	-0.3691	0.7121	-1.3188	0.9009
mretired	0.4231	0.4019	1.0528	0.2925	-0.3647	1.2109
mwealth	-4.683e-06	3.615e-06	-1.2953	0.1952	-1.177e-05	2.403e-06

12 Análisis CRE

El modelo CRE presenta una precision muy similar al RE, pero existen varios efectos notables en respecto a ciertas variables.

Variables notables: - hrsusu: Las horas trabajadas semanales promedio, no muestran ninguna significancia por lo que su coeficiente es mas probablente 0, pero si mhrsusu, que parece ser mas significativo para la regresion, además posee un coeficiente mas grande y significativamente distinto de 0.

- married: La variable married es significativa, pero no así la mmarried que presenta un p-value de 0.7807 y un coeficiente muy bajo, que es probablemente cero

Variables no signficativas: - wealth: Wealth y mwalth no presentan significancia ni un coeficiente lo suficientemente alto para importar en el modelo - child: No es significativa ni tampoco mchild

13 Pregunta 7: Predicción de Distribución Mediante Modelo CRE

7. Usando el modelo CRE, prediga la distribucion del componente no observado. Que puede inferir respecto de la heterogeneidad fija en el tiempo y su impacto en el puntaje CESD?

El modelo CRE presenta un p-value: 0.0000, si se comparan los coeficientes de la matriz de medias, con los componentes que varian en el tiempo, estas tienen coeficientes mas pequeños, en general y p-values muy altos, por lo que existe una alta probabilidad de que el componente que varia en el tiempo del error sea muy pequeño.

Esto implica que el error asociado a cada individuo (Heterogeneidad), es muy alto, es decir la distribución, del error tiene una alta varianza.

El impacto en el puntaje CESD, significaría que, cada individuo posee un puntaje base de CESD distinto, que afecta su variabilidad en el tiempo.

8. Usando sus respuestas anteriores, que modelo prefiere? que se puede inferir en general respecto del efecto de las variables explicativas sobre el puntaje CESD?

El modelo preferido para explicar el puntaje de CESD es PooledOLS, por que posee mejor precision así como mayor significancia para las variables. Y comparativamente presenta mejores regresiones.

[13]:
```
print(final_model.summary)
```

```
                          PooledOLS Estimation Summary
================================================================================
Dep. Variable:                    cesd   R-squared:                        0.0716
Estimator:                   PooledOLS   R-squared (Between):              0.1078
No. Observations:                10028   R-squared (Within):               0.0003
Date:                Wed, Oct 05 2022   R-squared (Overall):              0.0716
Time:                        13:41:50   Log-likelihood                -3.229e+04
Cov. Estimator:                Robust
                                        F-statistic:                      85.815
Entities:                        3345   P-value                           0.0000
Avg Obs:                       2.9979   Distribution:                 F(9,10018)
Min Obs:                       2.0000
Max Obs:                       3.0000   F-statistic (robust):             83.767
                                        P-value                           0.0000
Time periods:                       3   Distribution:                 F(9,10018)
Avg Obs:                       3342.7
Min Obs:                       3341.0
Max Obs:                       3345.0

                            Parameter Estimates
================================================================================
```

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	10.754	0.5808	18.515	0.0000	9.6157	11.893
child	0.0765	0.0494	1.5491	0.1214	-0.0203	0.1733
hrsusu	0.0060	0.0028	2.1450	0.0320	0.0005	0.0115
married	-1.4550	0.2004	-7.2615	0.0000	-1.8478	-1.0623
retired	0.5049	0.1860	2.7141	0.0067	0.1402	0.8695
schadj	-0.1953	0.0182	-10.721	0.0000	-0.2310	-0.1596
urban	-1.7967	0.1323	-13.578	0.0000	-2.0560	-1.5373
wealth	-2.343e-06	1.968e-06	-1.1905	0.2339	-6.201e-06	1.515e-06
age	-0.0103	0.0084	-1.2159	0.2241	-0.0268	0.0063
female	1.5284	0.1315	11.626	0.0000	1.2707	1.7861

En base a las variables explicativas, se puede decir que las variables dicotomicas tienden a tener el mayor efecto sobre el cesd, justificado por la presencia de un componente no observado probablemente asociado a estas variables, además las variables discretas con mayor impacto son child y schadj. Y se pueden hacer ciertas inferencias respecto a los cambios marginales, donde: - Por cada hijo extra, el cesd aumnenta en 0.0812 puntos. - Por cada año de escolaridad, el puntaje de cae 0.1951 puntos. - Se requieren 2322000 wealth (Sea cual sea la unidad monetaria) para disminuir el puntaje cesd en 1.

[ ]: