Tarea_2_Conejeros_Gonzalez

October 14, 2022

0.1 Tarea 2: Felipe Conejeros y Mabel González

0.2 Librerías a utilizar

[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 sns

%matplotlib inline
```

0.3 Pregunta 1: Limpieza y análisis descriptivo de los datos

0.3.1 Limpieza de datos

[2]:
```
charls = pd.read_csv('../data/charls.csv')
charls.head()
```

[2]:

	cesd	child	drinkly	female	hrsusu	hsize	inid	intmonth	married	retired	schadj	urban	wave	wealth	age
0	6	2	0.None	1	0.0	4	1.010410e+10	7	1	0	0	0	1	-5800.0	46
1	7	2	0.None	1	49.0	4	1.010410e+10	7	1	0	0	0	2	100.0	46
2	5	2	0.None	1	56.0	7	1.010410e+10	8	1	0	0	0	3	-59970.0	46
3	0	2	1.Yes	0	63.0	4	1.010410e+10	7	1	0	4	0	1	-5800.0	48
4	5	2	1.Yes	0	49.0	4	1.010410e+10	7	1	0	4	0	2	100.0	48

Dado que los valores de la variable inid mantienen un error de la base de datos fuente, se procederá a eliminar las observaciones desde la fila 10.057.

[3]:
```
charls = charls.drop(range(10057,34371),axis=0)
charls.head(-1)
```

[3]:

	cesd	child	drinkly	female	hrsusu	hsize	inid	intmonth	\
0	6	2	0.None	1	0.0	4	1.010410e+10	7	
1	7	2	0.None	1	49.0	4	1.010410e+10	7	
2	5	2	0.None	1	56.0	7	1.010410e+10	8	
3	0	2	1.Yes	0	63.0	4	1.010410e+10	7	
4	5	2	1.Yes	0	49.0	4	1.010410e+10	7	
…	…	…	…	…	…	…	…	…	
10051	5	1	0.None	1	0.0	5	9.400430e+10	7	
10052	3	1	1.Yes	1	0.0	4	9.400430e+10	8	
10053	4	2	0.None	1	0.0	2	9.400431e+10	8	
10054	5	2	0.None	1	0.0	2	9.400431e+10	7	
10055	5	2	0.None	1	0.0	4	9.400431e+10	8	

	married	retired	schadj	urban	wave	wealth	age
0	1	0	0	0	1	-5800.0	46
1	1	0	0	0	2	100.0	46
2	1	0	0	0	3	-59970.0	46
3	1	0	4	0	1	-5800.0	48
4	1	0	4	0	2	100.0	48
…	…	…	…	…	…	…	…
10051	1	1	8	1	2	505000.0	69
10052	1	1	8	1	3	1479000.0	69
10053	0	1	8	1	1	3000.0	61
10054	0	1	8	1	2	32500.0	61
10055	0	1	8	1	3	0.0	61

[10056 rows x 15 columns]

De la columna “drinkly”, se tienen valores tipo string, por lo que reemplazará de la siguiente forma:
* 0.None: 0 * 1.Yes: 1

[4]:
```
for i in range(len(charls)):
    if charls["drinkly"][i] == "0.None":
        charls["drinkly"][i] = 0

for i in range(len(charls)):
    if charls["drinkly"][i] == "1.Yes":
        charls["drinkly"][i] = 1
```

```
<ipython-input-4-cfdd89e65fa3>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  charls["drinkly"][i] = 0
```

```
<ipython-input-4-cfdd89e65fa3>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  charls["drinkly"][i] = 1
```

Por otro lado, se encontraron valores ".m:missing" en la misma columna, por lo que las filas que lo contengan serán eliminadas.

[5]:
```
k = 0

id_missing = []
wave_missing = []

for i in range(len(charls)):
    if charls["drinkly"][i] == '.m:missing':
        id_missing.append(i)
        wave_missing.append(charls["wave"][i])
        k += 1

print("Se encontraron", k, "filas con string '.m:missing' de la columna␣
 ↪'drinkly', las cuales serán eliminadas")
print("Se eliminará el inid completo, considerando sus distintos 'wave'")

for i in id_missing:
    if charls["wave"][i] == 2:
        print("Se elimina la fila", i, "con 'inid'", charls["inid"][i], "y␣
 ↪'wave'", charls["wave"][i])
        print("Se elimina la fila", i - 1, "con 'inid'", charls["inid"][i - 1],␣
 ↪"y 'wave'",charls["wave"][i - 1])
        print("Se elimina la fila", i + 1, "con 'inid'", charls["inid"][i + 1],␣
 ↪"y 'wave'",charls["wave"][i + 1])
        print("-------------------------------------------------------------")

    if charls["wave"][i] == 3:
        print("Se elimina la fila", i, "con 'inid'", charls["inid"][i], "y␣
 ↪'wave'", charls["wave"][i])
        print("Se elimina la fila", i - 1, "con 'inid'", charls["inid"][i - 1],␣
 ↪"y 'wave'",charls["wave"][i - 1])
        print("Se elimina la fila", i - 2, "con 'inid'", charls["inid"][i - 2],␣
 ↪"y 'wave'",charls["wave"][i - 2])
        print("-------------------------------------------------------------")

for i in id_missing:
    if charls["wave"][i] == 2:
        charls = charls.drop([i])
```

```
        charls = charls.drop([i - 1])
        charls = charls.drop([i + 1])
    elif charls["wave"][i] == 3:
        charls = charls.drop([i])
        charls = charls.drop([i - 1])
        charls = charls.drop([i - 2])
```

```
Se encontraron 7 filas con string '.m:missing' de la columna 'drinkly', las
cuales serán eliminadas
Se eliminará el inid completo, considerando sus distintos 'wave'
Se elimina la fila 4712 con 'inid' 56059207001.0 y 'wave' 3
Se elimina la fila 4711 con 'inid' 56059207001.0 y 'wave' 2
Se elimina la fila 4710 con 'inid' 56059207001.0 y 'wave' 1
------------------------------------------------------------
Se elimina la fila 4813 con 'inid' 56059314002.0 y 'wave' 2
Se elimina la fila 4812 con 'inid' 56059314002.0 y 'wave' 1
Se elimina la fila 4814 con 'inid' 56059314002.0 y 'wave' 3
------------------------------------------------------------
Se elimina la fila 5878 con 'inid' 57457309001.0 y 'wave' 2
Se elimina la fila 5877 con 'inid' 57457309001.0 y 'wave' 1
Se elimina la fila 5879 con 'inid' 57457309001.0 y 'wave' 3
------------------------------------------------------------
Se elimina la fila 6326 con 'inid' 58202302001.0 y 'wave' 3
Se elimina la fila 6325 con 'inid' 58202302001.0 y 'wave' 2
Se elimina la fila 6324 con 'inid' 58202302001.0 y 'wave' 1
------------------------------------------------------------
Se elimina la fila 6394 con 'inid' 58202320002.0 y 'wave' 2
Se elimina la fila 6393 con 'inid' 58202320002.0 y 'wave' 1
Se elimina la fila 6395 con 'inid' 58202320002.0 y 'wave' 3
------------------------------------------------------------
Se elimina la fila 9142 con 'inid' 74981324002.0 y 'wave' 2
Se elimina la fila 9141 con 'inid' 74981324002.0 y 'wave' 1
Se elimina la fila 9143 con 'inid' 74981324002.0 y 'wave' 3
------------------------------------------------------------
Se elimina la fila 9227 con 'inid' 75376118001.0 y 'wave' 3
Se elimina la fila 9226 con 'inid' 75376118001.0 y 'wave' 2
Se elimina la fila 9225 con 'inid' 75376118001.0 y 'wave' 1
------------------------------------------------------------
```

[6]:
```
charls.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10036 entries, 0 to 10056
Data columns (total 15 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   cesd      10036 non-null  int64
 1   child     10036 non-null  int64
```

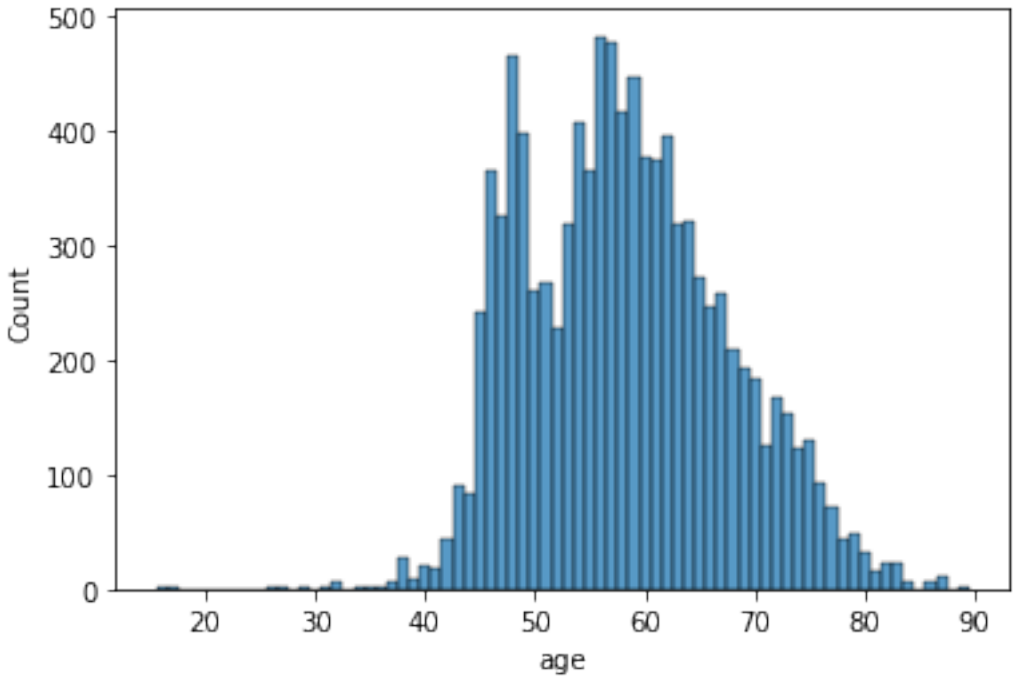
```
 2   drinkly    10036 non-null  object
 3   female     10036 non-null  int64
 4   hrsusu     10036 non-null  float64
 5   hsize      10036 non-null  int64
 6   inid       10036 non-null  float64
 7   intmonth   10036 non-null  int64
 8   married    10036 non-null  int64
 9   retired    10036 non-null  int64
 10  schadj     10036 non-null  int64
 11  urban      10036 non-null  int64
 12  wave       10036 non-null  int64
 13  wealth     10036 non-null  float64
 14  age        10036 non-null  int64
dtypes: float64(3), int64(11), object(1)
memory usage: 1.2+ MB
```

De la salida anterior, se observa que "drinkly" es una variable de tipo objeto, por lo que se cambiará a float con el fin de usar los modelos.

[7]:
```
charls = charls.astype({"drinkly":"float"})
```

[8]:
```
sns.histplot(charls, x = "age", discrete = True)
charls.value_counts("age")
```

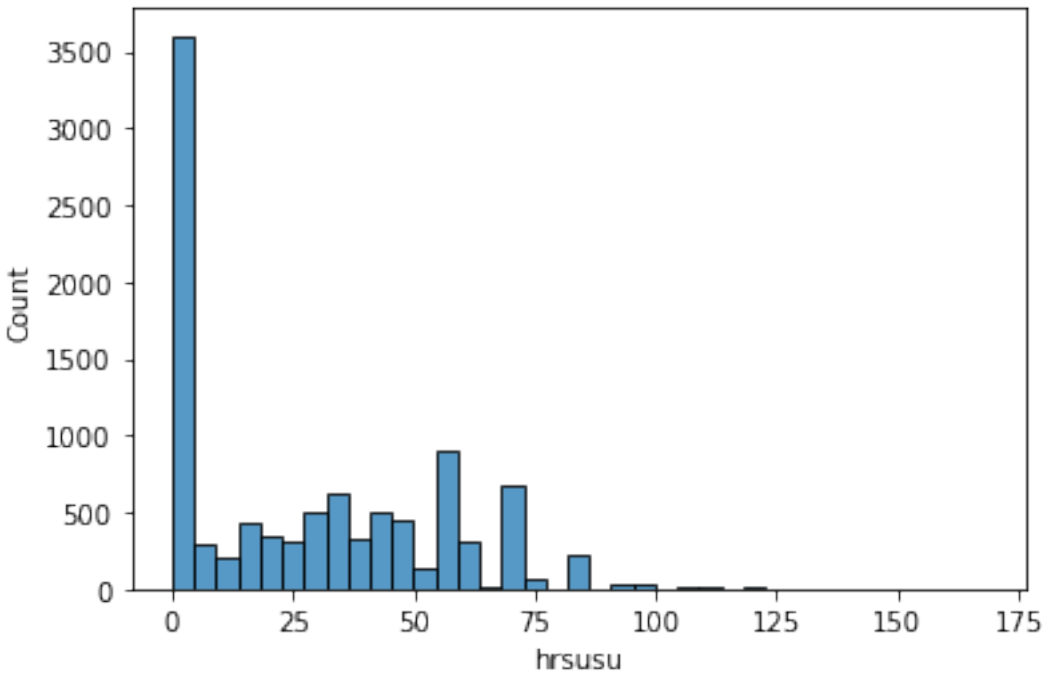
[8]:
```
age
56    483
57    477
48    465
59    447
58    417
      …
31      3
29      3
27      3
26      3
89      3
Length: 61, dtype: int64
```



Del gráfico anterior, correspondiente a la distribución de “age”, no se observan outliers.

[9]:
```
sns.histplot(charls, x = "hrsusu")
charls.value_counts("hrsusu")
```

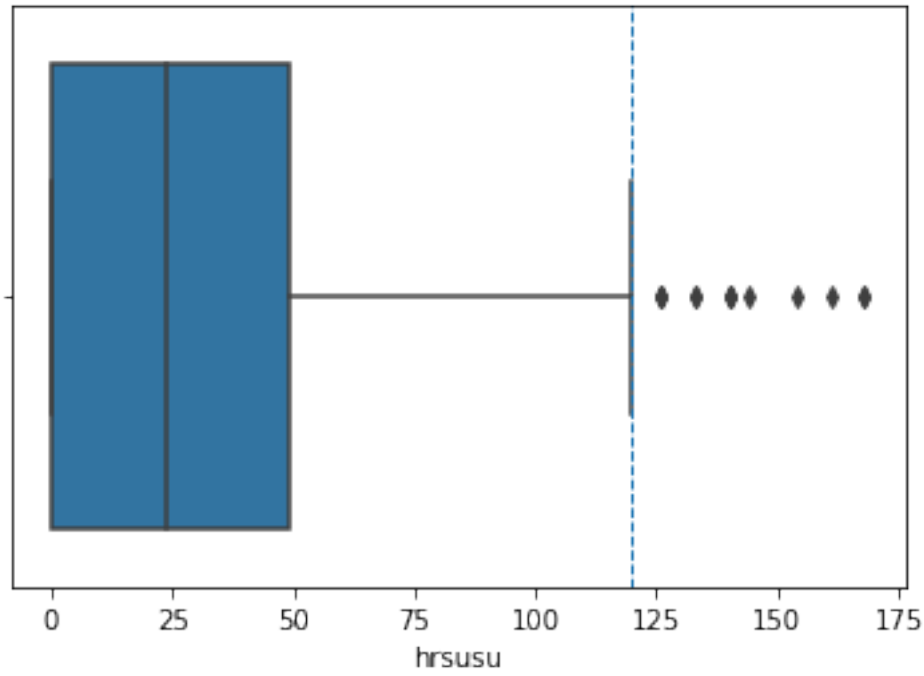
[9]:
```
hrsusu
0.0     3441
56.0     901
70.0     650
42.0     457
35.0     405
        …
32.5       1
66.5       1
38.5       1
62.5       1
90.0       1
Length: 79, dtype: int64
```



De la figura anterior, se identifican posibles outliers para la variable “hrsusu”.

[10]:
```
sns.boxplot(x = charls["hrsusu"])
plt.axvline(120, 0,20, ls='--', lw=1)
```

[10]:
```
<matplotlib.lines.Line2D at 0x21c0a808a60>
```



Del boxplot, se verifica la existencia de outliers, que corresponden a un "hrsusu" mayor que 120. Estas observaciones serán eliminadas.

[11]:
```
borrar = charls[charls["hrsusu"] > 120].index
charls = charls.drop(borrar)
```

[12]:
```
# charls data
charls.dropna(inplace=True)
charls.reset_index(drop=True, inplace=True)


#variable construction
X=charls[['cesd','child','drinkly','female','hrsusu','hsize','intmonth','married','retired','s
Xm=(X.groupby(charls['inid']).transform('mean'))

Xid=charls[['inid','wave','cesd','child','drinkly','female','hrsusu','hsize','intmonth','marri
Xc=pd.DataFrame(np.c_[Xid, Xm],␣
  ↪columns=['inid','wave','cesd','child','drinkly','female','hrsusu','hsize','intmonth','marri
Xc = Xc.set_index(['inid','wave'])
```

0.3.2 Análisis descriptivo

[13]:
```
charls.describe()
```

[13]:

	cesd	child	drinkly	female	hrsusu
count	10023.000000	10023.000000	10023.000000	10023.000000	10023.000000
mean	8.867106	2.768433	0.323855	0.542851	27.809189
std	6.291789	1.435903	0.467969	0.498185	26.942820
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	4.000000	2.000000	0.000000	0.000000	0.000000
50%	8.000000	2.000000	0.000000	1.000000	24.000000
75%	13.000000	3.000000	1.000000	1.000000	49.000000
max	30.000000	10.000000	1.000000	1.000000	120.000000

	hsize	inid	intmonth	married	retired
count	10023.000000	1.002300e+04	10023.000000	10023.000000	10023.000000
mean	3.654195	4.886154e+10	7.593236	0.857927	0.269081
std	1.785813	2.297920e+10	1.101360	0.349143	0.443504
min	1.000000	1.010410e+10	1.000000	0.000000	0.000000
25%	2.000000	3.110611e+10	7.000000	1.000000	0.000000
50%	3.000000	5.630230e+10	7.000000	1.000000	0.000000
75%	5.000000	6.403312e+10	8.000000	1.000000	1.000000
max	13.000000	1.017910e+11	12.000000	1.000000	1.000000

	schadj	urban	wave	wealth	age
count	10023.000000	10023.000000	10023.000000	1.002300e+04	10023.000000
mean	4.092787	0.315574	1.999401	1.019523e+04	58.229372
std	3.604440	0.464767	0.816415	9.958253e+04	9.233720
min	0.000000	0.000000	1.000000	-9.750000e+05	16.000000
25%	0.000000	0.000000	1.000000	0.000000e+00	51.000000
50%	4.000000	0.000000	2.000000	3.000000e+02	58.000000
75%	4.000000	1.000000	3.000000	4.075000e+03	64.000000
max	16.000000	1.000000	3.000000	8.001500e+06	89.000000

La tabla anterior corresponde a la base de datos final que se analizará, la cual cuenta con un total de 10.023 observaciones. * Para la variable “cesd”, correspondiente al puntaje en la escala de la salud mental, se observa un promedio de 8,86, con un mínimo de 0 y un máximo de 30. * “child”, perteneciente al número de hijos, tiene una media de 2,76, con un valor mínimo de 0 y máximo de 10. * La variables binarias “drinkly” y “female” mantienen un promedio de 0,32 y 0,54, respectivamente. * “hrsusu” corresponde a las horas promedio de trabajo semanal, la cual mantiene un promedio de 27,8, un máximo de 120 (aproximadamente 17 horas por día) y un mínimo de 0. * “hsize” pertenece al tamaño del hogar, con un promedio de 3,65, un máximo de 13 y un mínimo de 1. * La variable “intmonth” corresponde al mes en que el individuo fue encuestado. * “married” y “retired” son variables binarias correspondientes a que si la persona está o no casada y pensionada, la cual mantiene una media de 0,85 y de 0,26, respectivamente. * La variable “schadj” corresponde a los años de escolaridad, con un promedio de 4,1, un máximo de 16 y un mínimo de 0. * “urban” es una variable binaria para describir si la persona vive en la zona urbana o no, con un promedio de 0,31. * “wealth” corresponde a la riqueza neta, con un promedio de 1.0e+04, un máximo de 8.0e+06 y un mínimo de -9.7e+05. * Finalmente, la variable “age” corresponde a la edad al entrar a la encuesta.

0.4 Pregunta 2: Modelo Pooled OLS

[14]:
```
y=Xc['cesd']
X=Xc[['child','drinkly','female','hrsusu','hsize','married','retired','schadj','urban','wealth
X=sm.add_constant(X)

model = sm.OLS(y, X.astype(float))
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   cesd   R-squared:                       0.072
Model:                           OLS   Adj. R-squared:                  0.071
Method:                Least Squares   F-statistic:                     70.79
Date:               Wed, 05 Oct 2022   Prob (F-statistic):          9.09e-154
Time:                       20:14:17   Log-Likelihood:                -32281.
No. Observations:              10023   AIC:                         6.459e+04
Df Residuals:                  10011   BIC:                         6.467e+04
Df Model:                         11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         11.1256      0.600     18.544      0.000       9.950      12.302
child          0.0934      0.049      1.916      0.055      -0.002       0.189
drinkly       -0.0064      0.144     -0.044      0.965      -0.289       0.276
female         1.5231      0.143     10.633      0.000       1.242       1.804
hrsusu         0.0059      0.003      2.027      0.043       0.000       0.012
hsize         -0.0767      0.035     -2.172      0.030      -0.146      -0.007
married       -1.3845      0.187     -7.419      0.000      -1.750      -1.019
retired        0.4886      0.183      2.668      0.008       0.130       0.848
schadj        -0.1987      0.019    -10.543      0.000      -0.236      -0.162
urban         -1.8002      0.137    -13.112      0.000      -2.069      -1.531
wealth     -2.327e-06   6.12e-07     -3.804      0.000   -3.53e-06   -1.13e-06
age           -0.0131      0.008     -1.573      0.116      -0.030       0.003
==============================================================================
Omnibus:                      665.045   Durbin-Watson:                   1.332
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              805.282
Skew:                           0.693   Prob(JB):                    1.37e-175
Kurtosis:                       3.093   Cond. No.                     1.01e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 1.01e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
C:\Users\felip\anaconda3\lib\site-packages\statsmodels\tsa\tsatools.py:142:
FutureWarning: In a future version of pandas all arguments of concat except for
the argument 'objs' will be keyword-only
  x = pd.concat(x[::order], 1)
```

0.4.1 Interpretación:

El modelo Pooled OLS equivale a estimar el efecto promedio, asumiendo que el tiempo afecta a todas las unidades de la misma forma. El método asume que todas la personas encuestadas son idénticas, lo que puede ocasionar que se sobredimensione la verdadera relación entre las variables.

- El valor de 0,072 del R-cuadrado indica que sólo un 7,2% de la variación en la variable cesd es explicada por las variables consideradas.
- La prueba F entrega un valor-p muy pequeño, lo que implica que todas las variables en conjunto explican estadísticamente la variación del puntaje de salud mental.
- Los coeficientes de las variables: female, married, retired, schadj, urban y wealth son significativamente distintos de cero en un 99%, y los de las variables: hrsusu y hsize son significativos en un 95%. Por lo que estas variables explican estadísticamente el puntaje de salud mental.
- Las variables que resultaron ser no significativas son: child, drinkly y age. Drinkly es la variable que uno esperaría aportara a explicar la salud mental de una persona, sin embargo como se mide si una persona ha bebido alguna vez en el último mes tiene sentido que no sea determinante.
- En cuanto al genero, si el individuo es mujer, esto tiende a aumentar el puntaje cesd en 1,5231 puntos en promedio, disminuyendo su salud mental en comparación con los hombres.
- Frente a un aumento de 10 horas de trabajo semanal del individuo, el aumento promedio del puntaje cesd es de 0,059 unidades, lo que indica que en promedio la salud mental de las personas no se ve muy afectada por más hora de trabajo.
- Al aumentar el tamaño del hogar de una persona, en promedio el puntaje cesd disminuye en 0,0767 unidades.
- Las personas casadas tienden a disminuir el puntaje de su salud mental en al menos 1 punto, lo que indica que en promedio las personas casadas gozan de mejor salud mental.
- Las personas retiradas suman en promedio 0,4886 puntos en la escala cesd. Esta disminución en su salud mental se podría explicar porque las personas dejan de estar ocupadas con actividades diarias.
- Mientras más años de escolaridad tiene una persona, en promedio su salud mental tiende a mejorar en 0,1987 puntos.
- Vivir en una zona urbana puede afectar positivamente la salud mental de una persona, haciendo que su puntaje disminuya casi 2 puntos en promedio. Esto podría ser explicado porque las personas tienen mayor acceso a centros de salud, espacios recreativos, oportunidades laborales, etc.
- Las personas con más riqueza experimetan en promedio una mejora en su salud mental, pero en una magnitud casi imperceptible.

[15]:
```
charls.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10023 entries, 0 to 10022
Data columns (total 15 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   cesd       10023 non-null  int64  
 1   child      10023 non-null  int64  
 2   drinkly    10023 non-null  float64
 3   female     10023 non-null  int64  
 4   hrsusu     10023 non-null  float64
 5   hsize      10023 non-null  int64  
 6   inid       10023 non-null  float64
 7   intmonth   10023 non-null  int64  
 8   married    10023 non-null  int64  
 9   retired    10023 non-null  int64  
 10  schadj     10023 non-null  int64  
 11  urban      10023 non-null  int64  
 12  wave       10023 non-null  int64  
 13  wealth     10023 non-null  float64
 14  age        10023 non-null  int64  
dtypes: float64(4), int64(11)
memory usage: 1.1 MB
```

0.4.2 Estimadores robustos

[16]:
```
model=lmp.PooledOLS(y,X)
OLS=model.fit(cov_type="robust")
print(OLS)
```

```
                          PooledOLS Estimation Summary                          
================================================================================
Dep. Variable:                   cesd   R-squared:                        0.0722
Estimator:                  PooledOLS   R-squared (Between):              0.1079
No. Observations:               10023   R-squared (Within):               0.0011
Date:                Wed, Oct 05 2022   R-squared (Overall):              0.0722
Time:                        20:14:18   Log-likelihood                -3.228e+04
Cov. Estimator:                Robust                                           
                                        F-statistic:                      70.791
Entities:                        3346   P-value                           0.0000
Avg Obs:                       2.9955   Distribution:                F(11,10011)
Min Obs:                       1.0000                                           
Max Obs:                       3.0000   F-statistic (robust):             68.988
                                        P-value                           0.0000
Time periods:                       3   Distribution:                F(11,10011)
Avg Obs:                       3341.0                                           
Min Obs:                       3337.0                                           
```

```
Max Obs:                        3343.0

                            Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
const          11.126     0.6175     18.017     0.0000      9.9152      12.336
child          0.0934     0.0499     1.8708     0.0614     -0.0045      0.1912
drinkly       -0.0064     0.1426    -0.0448     0.9643     -0.2859      0.2732
female         1.5231     0.1437     10.600     0.0000      1.2415      1.8048
hrsusu         0.0059     0.0029     2.0574     0.0397      0.0003      0.0116
hsize         -0.0767     0.0346    -2.2176     0.0266     -0.1444     -0.0089
married       -1.3845     0.2023    -6.8452     0.0000     -1.7810     -0.9880
retired        0.4886     0.1868     2.6155     0.0089      0.1224      0.8547
schadj        -0.1987     0.0183    -10.855     0.0000     -0.2346     -0.1628
urban         -1.8002     0.1325    -13.585     0.0000     -2.0600     -1.5405
wealth     -2.327e-06  1.954e-06    -1.1908     0.2338  -6.158e-06   1.504e-06
age           -0.0131     0.0086    -1.5333     0.1252     -0.0299      0.0037
==============================================================================
```

0.4.3 Interpretación

Se arregla el problema de que los estimadores sean robutos. La etimación se hace más flexible ya que la matriz de covarianzas asume heterocedasticidad.

- Los resultados no varian mucho.
- Ahora la variable wealth ya no es significativa, por lo que en promedio la riqueza de las personas no ayuda a explicar su salud mental. Esta mejora tiene sentido ya que anteriormente el aporte de la variable era muy pequeño.

0.5 Pregunta 3: Modelo de Efectos Fijos

Para utilizar este modelo, se excluyeron las variables con efectos fijos, las cuales fueron: female, schadj, urban y age, ya que no varían en el tiempo y presentan problemas de correlación serial.

[17]:
```
X=Xc[['child','drinkly','hrsusu','hsize','married','retired','wealth']]
X=sm.add_constant(X)
model=lmp.PanelOLS(y,X)
fe=model.fit(cov_type="robust")
print(fe)
```

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:                    cesd   R-squared:                        0.0219
Estimator:                   PanelOLS   R-squared (Between):              0.0369
No. Observations:               10023   R-squared (Within):              -0.0079
Date:                Wed, Oct 05 2022   R-squared (Overall):              0.0219
Time:                        20:14:19   Log-likelihood                -3.255e+04
Cov. Estimator:                Robust
```

```
                                  F-statistic:                     32.052
Entities:                          3346   P-value                   0.0000
Avg Obs:                         2.9955   Distribution:          F(7,10015)
Min Obs:                         1.0000
Max Obs:                         3.0000   F-statistic (robust):        26.092
                                          P-value                   0.0000
Time periods:                       3     Distribution:          F(7,10015)
Avg Obs:                         3341.0
Min Obs:                         3337.0
Max Obs:                         3343.0

                            Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
const          10.034     0.2747     36.530     0.0000      9.4953      10.572
child          0.2492     0.0450     5.5341     0.0000      0.1610      0.3375
drinkly       -0.8938     0.1322    -6.7595     0.0000     -1.1530     -0.6346
hrsusu         0.0047     0.0029     1.6075     0.1080     -0.0010      0.0105
hsize         -0.0322     0.0348    -0.9243     0.3553     -0.1005      0.0361
married       -1.7562     0.2007    -8.7490     0.0000     -2.1497     -1.3627
retired       -0.1540     0.1857    -0.8294     0.4069     -0.5181      0.2100
wealth     -3.204e-06  2.316e-06    -1.3834     0.1666  -7.744e-06   1.336e-06
==============================================================================



C:\Users\felip\anaconda3\lib\site-packages\statsmodels\tsa\tsatools.py:142:
FutureWarning: In a future version of pandas all arguments of concat except for
the argument 'objs' will be keyword-only
  x = pd.concat(x[::order], 1)
```

0.5.1 Interpretación

Este modelo resta el promedio de las variables a cada observación, por lo que permite comparar a los individuos con ellos mismos en el tiempo en vez de con otros individuos.

- El valor de 0,0219 del R-cuadrado indica que sólo un 2,19% de la variación en la variable cesd es explicada por las variables consideradas.
- La prueba F entrega un valor-p igual a cero, lo que implica que todas las variables en conjunto explican estadísticamente la variación del puntaje de salud mental.
- Ahora las variables child y drinkly son significativas, por lo que el numero de hijos y la ingesta de alcohol ayudan a explicar la variación en la salud mental de los individuos en el tiempo, en conjunto con la variable married.
- Las variables hrsusu, hsize y retired se convirtieron en no significativas, mientras que wealth sigue siendo no significativa.
- El aumento en el numero de hijos de una persona aumenta el puntaje cesd en 0,2492 puntos,

disminuyendo su salud mental.

- Si una persona comienza a beber cuando antes no lo hacia, esto tiende a disminuir su puntaje cesd en 0,8938 puntos, mejorando su salud mental. Esto no tiene una explicación lógica inmediata, lo que puede deberse a que sólo se está considerando si una persona ha bebido alcohol en el último mes y no la frecuencia o cantidad en que lo ha hecho.
- Luego de que una persona se casa, esta tiende a mejorar su salud mental en al menos 1 punto.
- EL modelo no agrupado (fixed) es mejor que el agrupado (valor p=0)

0.6 Pregunta 4: Modelo de Efectos Aleatorios

[18]:
```
X=Xc[['child','drinkly','female','hrsusu','hsize','married','retired','schadj','urban','wealth
model=lmp.RandomEffects(y,X)
re=model.fit(cov_type="robust")
print(re)
```

```
                          RandomEffects Estimation Summary
================================================================================
Dep. Variable:                    cesd   R-squared:                        0.5197
Estimator:              RandomEffects   R-squared (Between):              0.7632
No. Observations:               10023   R-squared (Within):              -0.0025
Date:                Wed, Oct 05 2022   R-squared (Overall):              0.6774
Time:                        20:14:20   Log-likelihood                -2.924e+04
Cov. Estimator:                Robust
                                        F-statistic:                      985.00
Entities:                        3346   P-value                           0.0000
Avg Obs:                       2.9955   Distribution:                F(11,10012)
Min Obs:                       1.0000
Max Obs:                       3.0000   F-statistic (robust):             960.48
                                        P-value                           0.0000
Time periods:                       3   Distribution:                F(11,10012)
Avg Obs:                       3341.0
Min Obs:                       3337.0
Max Obs:                       3343.0

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
child          0.0256     0.0606     0.4218     0.6732     -0.0932      0.1443
drinkly        0.3537     0.1504     2.3521     0.0187      0.0589      0.6485
female         2.6603     0.1793     14.836     0.0000      2.3088      3.0118
hrsusu         0.0093     0.0026     3.5957     0.0003      0.0042      0.0144
hsize          0.0264     0.0345     0.7656     0.4439     -0.0412      0.0941
married        0.0671     0.2322     0.2892     0.7725     -0.3880      0.5222
retired        0.3614     0.1777     2.0336     0.0420      0.0130      0.7098
schadj        -0.0738     0.0239    -3.0889     0.0020     -0.1206     -0.0270
```

```
urban          -1.7715     0.1867    -9.4879     0.0000     -2.1375     -1.4055
wealth     -1.282e-06  1.085e-06    -1.1822     0.2372  -3.408e-06   8.438e-07
age            0.1281     0.0053     23.967     0.0000      0.1176      0.1386
==============================================================================
```

0.6.1 Interpretación

El modelo de efectos aleatorios permite variables explicativas que son constantes en el tiempo, siendo una ventaja frente al modelo de efectos fijos. Esto es posible debido a que supone que el efecto inobservable no está correlacionado con ninguna de las variables explicativas, ya sea que las variables explicativas estén fijas en el tiempo o no. Si bien este modelo logra entregar más información con respecto a la heterogeneidad, si se segmentaran los individuos se lograrían mejores resultados.

- El valor del R-cuadrado aumentó con respecto al modelo anterior, siendo ahora de 0,5197, y el valor p de la prueba F se mantiene igual a cero, por lo que todas las variables en conjunto explican estadísticamente la variación del puntaje de salud mental.
- Los coeficientes de las variables: female, hrsusu, schadj, urban y age son significativamente distintos de cero en un 99%, y los de las variables: drinkly y retired son significativos en un 95%. Por lo que estas variables explican estadísticamente el puntaje de salud mental.
- Las variables que resultaron ser no significativas son: child, hsize, married y wealth, donde child y married difieren de los modelos anteriores.
- En una población, las personas que beben al menos una vez al mes aumentan su puntaje cesd en 0,3537 puntos, por lo que su salud mental disminuye. Se evidencia que el efecto del alcohol es contrario al modelo anterior.
- En cuanto al genero, las mujeres de una población tienden a aumentar el puntaje cesd en 2,6603 puntos, disminuyendo su salud mental en comparación con los hombres.
- Frente a un aumento de 10 horas de trabajo semanal de las personas, su puntaje cesd aumenta en 0,093 unidades, por lo que la salud mental de la población no se ve muy afectada por más hora de trabajo.
- Mientras más años de escolaridad tienen las personas, su salud mental tiende a mejorar en 0,0738 puntos por cada año.
- Vivir en una zona urbana puede afectar positivamente la salud mental de las personas, haciendo que su puntaje disminuya 1,7715 puntos.
- Mientras más años cumplan las personas de una población, su puntaje de salud tiende a aumentar en 0,1281 puntos, de donde se extrae que las personas van empeorando su salud mental progresivamente.

A continuación, se extraerán algunas variables para poder comparar random effects y efectos fijos a través del test de Hausman.

[19]:
```
X=Xc[['child','drinkly','hrsusu','hsize','married','retired','wealth']]
model=lmp.RandomEffects(y,X)
re=model.fit(cov_type="robust")
print(re)
```

```
                          RandomEffects Estimation Summary
================================================================================
Dep. Variable:                    cesd   R-squared:                        0.4354
Estimator:              RandomEffects   R-squared (Between):              0.7078
No. Observations:               10023   R-squared (Within):              -0.0545
Date:                Wed, Oct 05 2022   R-squared (Overall):              0.6224
Time:                        20:14:20   Log-likelihood                -2.947e+04
Cov. Estimator:                Robust
                                        F-statistic:                     1103.3
Entities:                        3346   P-value                          0.0000
Avg Obs:                       2.9955   Distribution:                 F(7,10016)
Min Obs:                       1.0000
Max Obs:                       3.0000   F-statistic (robust):            1140.1
                                        P-value                          0.0000
Time periods:                       3   Distribution:                 F(7,10016)
Avg Obs:                       3341.0
Min Obs:                       3337.0
Max Obs:                       3343.0

                             Parameter Estimates
================================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
--------------------------------------------------------------------------------
child          1.1537     0.0524     22.019     0.0000      1.0510      1.2564
drinkly        0.2823     0.1499     1.8830     0.0597     -0.0116      0.5762
hrsusu         0.0256     0.0027     9.6378     0.0000      0.0204      0.0308
hsize          0.2711     0.0355     7.6333     0.0000      0.2015      0.3407
married        3.0070     0.2091     14.378     0.0000      2.5970      3.4170
retired        1.7569     0.1734     10.133     0.0000      1.4170      2.0967
wealth     -1.175e-06  9.241e-07    -1.2720     0.2034  -2.987e-06    6.36e-07
================================================================================
```

0.6.2 Interpretación

- En este nuevo modelo el valor del R-cuadrado disminuyó con respecto al modelo anterior, siendo ahora de 0,01, y el valor p de la prueba F se mantiene igual a cero.
- Las variables significativas y no significativas cambiaron, donde ahora solo child, drinkly y married explican el puntaje de salud mental de los individuos.

[20]:
```
re.variance_decomposition
```

[20]:
```
Effects                 22.521591
Residual                19.850840
Percent due to Effects   0.531515
Name: Variance Decomposition, dtype: float64
```

0.6.3 Interpretación

Se obtiene la descomposición de varianzas que indica que proporción es el mu y U. Si asumimos que la correlación es 0 las proporciones serán cerca de 50% y 50%. Esto indica que 50% de los efectos vienen de la heteregeonidad entre los individuos encuestados y la otra mitad vienen de los errores residuales de shocks que varían en el tiempo y entre individuo.

0.7 Model comparison

[21]:
```
print(lmp.compare({"FE": fe, "RE": re, "Pooled": OLS}))
```

```
                          Model Comparison
=================================================================
                                FE              RE       Pooled
-----------------------------------------------------------------
Dep. Variable                 cesd            cesd         cesd
Estimator                PanelOLS   RandomEffects    PooledOLS
No. Observations            10023           10023        10023
Cov. Est.                  Robust          Robust       Robust
R-squared                  0.0219          0.4354       0.0722
R-Squared (Within)        -0.0079         -0.0545       0.0011
R-Squared (Between)        0.0369          0.7078       0.1079
R-Squared (Overall)        0.0219          0.6224       0.0722
F-statistic                32.052          1103.3       70.791
P-value (F-stat)           0.0000          0.0000       0.0000
==================== ============ =============== ============
const                      10.034                       11.126
                          (36.530)                     (18.017)
child                      0.2492          1.1537       0.0934
                          (5.5341)        (22.019)     (1.8708)
drinkly                   -0.8938          0.2823      -0.0064
                          (-6.7595)        (1.8830)    (-0.0448)
hrsusu                     0.0047          0.0256       0.0059
                          (1.6075)        (9.6378)     (2.0574)
hsize                     -0.0322          0.2711      -0.0767
                          (-0.9243)        (7.6333)    (-2.2176)
married                   -1.7562          3.0070      -1.3845
                          (-8.7490)        (14.378)    (-6.8452)
retired                   -0.1540          1.7569       0.4886
                          (-0.8294)        (10.133)     (2.6155)
wealth                  -3.204e-06      -1.175e-06   -2.327e-06
                          (-1.3834)       (-1.2720)    (-1.1908)
female                                                    1.5231
                                                         (10.600)
schadj                                                   -0.1987
                                                        (-10.855)
urban                                                    -1.8002
                                                        (-13.585)
```

```
age                                                          -0.0131
                                                            (-1.5333)
--------------------------------------------------------------------

T-stats reported in parentheses
```

[22]:
```
import numpy.linalg as la
from scipy import stats

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
```

0.8 Pregunta 5: Modelos Pooled OLS, Efectos Fijos y Efectos Aleatorios

[23]:
```
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 = nan, df = 7, p-value = nan
```

Se tiene que los errores estándar de Pooled OLS ignoran la correlación serial positiva, por lo que no son las mismas variables que se comparan entre modelos con respecto a efectos fijos y aleatorios.

Por otro lado, luego de realizar el test de Hausman, se obtuvo un p-value < 0.05 que rechaza la hipótesis nula de igualdad al 95% de confianza, lo que significa que la estimación de EF es mejor por sobre el modelo de efectos aleatorios. Lo anterior considerando solo cambios dentro de los mismos individuos en el tiempo, por lo que no se puede obtener mucha información de cambios entre personas.

Las variables robustas entre efectos fijos y efectos aleatorios son: child, drinkly y married. Esto debido a que son significativas para ambos modelos y no presentaron grandes variaciones en los coeficientes obtenidos.

0.9 Pregunta 6: Modelo de Efectos Aleatorios Correlacionados

[24]:
```
X=Xc[['child','drinkly','female','hrsusu','hsize','married','retired','schadj','urban','wealth
X=sm.add_constant(X)
model=lmp.RandomEffects(y,X, check_rank=False)
cre=model.fit(cov_type="robust")
print(cre)
```

```
                        RandomEffects Estimation Summary
================================================================================
Dep. Variable:                   cesd   R-squared:                        0.0415
```

Estimator:	RandomEffects	R-squared (Between):	0.1097
No. Observations:	10023	R-squared (Within):	0.0039
Date:	Wed, Oct 05 2022	R-squared (Overall):	0.0743
Time:	20:14:23	Log-likelihood	-2.919e+04
Cov. Estimator:	Robust		
		F-statistic:	19.704
Entities:	3346	P-value	0.0000
Avg Obs:	2.9955	Distribution:	F(22,10000)
Min Obs:	1.0000		
Max Obs:	3.0000	F-statistic (robust):	2.0964
		P-value	0.0019
Time periods:	3	Distribution:	F(22,10000)
Avg Obs:	3341.0		
Min Obs:	3337.0		
Max Obs:	3343.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	10.505	0.9259	11.347	0.0000	8.6906	12.320
child	0.1519	0.1007	1.5085	0.1315	-0.0455	0.3493
drinkly	0.2146	0.1938	1.1076	0.2681	-0.1652	0.5944
female	0.7596					
hrsusu	-0.0009	0.0030	-0.2930	0.7695	-0.0067	0.0049
hsize	-0.1195	0.0428	-2.7889	0.0053	-0.2035	-0.0355
married	-1.1849	0.4840	-2.4480	0.0144	-2.1338	-0.2361
retired	0.3354	0.2026	1.6557	0.0978	-0.0617	0.7325
schadj	-0.0954					
urban	-0.8932	1.8697	-0.4777	0.6329	-4.5583	2.7718
wealth	-4.952e-07	1.426e-06	-0.3472	0.7285	-3.291e-06	2.301e-06
age	-0.0038					
mchild	-0.0818	0.1260	-0.6497	0.5159	-0.3287	0.1651
mdrinkly	-0.3436	0.3039	-1.1305	0.2583	-0.9393	0.2521
mfemale	0.7596	1.0262	0.7402	0.4592	-1.2519	2.7711
mhrsusu	0.0144	0.0065	2.2109	0.0271	0.0016	0.0272
mhsize	0.0742	0.0768	0.9665	0.3338	-0.0763	0.2246
mmarried	-0.2094	0.5695	-0.3676	0.7132	-1.3258	0.9070
mretired	0.4217	0.4068	1.0368	0.2998	-0.3756	1.2191
mschadj	-0.0954					
murban	-0.8932	1.9729	-0.4528	0.6507	-4.7605	2.9741
mwealth	-4.734e-06	3.651e-06	-1.2967	0.1948	-1.189e-05	2.422e-06
mage	-0.0038	0.0223	-0.1686	0.8662	-0.0475	0.0400

```
C:\Users\felip\anaconda3\lib\site-packages\linearmodels\panel\results.py:87:
RuntimeWarning: invalid value encountered in sqrt
  return Series(np.sqrt(np.diag(self.cov)), self._var_names, name="std_error")
```

0.9.1 Interpretación

Utilizar el modelo de efectos aleatorios correlacionados permite considerar la heterogeneidad de cada persona encuestada, por lo que da más información considerando la caracterización de cada individuo con variables como age, urban, female. En este caso se observa que solo 2 variables resultaron significativas: hsize y married, por lo que solo estas variables consiguen explicar la variación en el puntaje de salud mental de los individuos. Se tiene que:

- Cuando el tamaño del hogar aumenta en 1, el puntaje cesd disminuye en 0,1195 unidades, indicando que a mayor tamaño del hogar las personas tienden a tener una mejor salud mental.
- Las personas casadas tienden a disminuir su puntaje de salud mental en casi 2 unidades, por lo que su salud mental es mejor considerando esta variable.

Cabe destacar que al considerar todos los promedios en el modelo no se logra identificar existosamente la heterogeneidad no observada, por lo que se deberían considerar solo algunas medias para el análisis.

[25]:
```
print(lmp.compare({"FE": fe, "RE": re, "CRE": cre}))
```

```
                          Model Comparison
=========================================================================
                                  FE              RE             CRE
-------------------------------------------------------------------------
Dep. Variable                   cesd            cesd            cesd
Estimator                   PanelOLS   RandomEffects   RandomEffects
No. Observations               10023           10023           10023
Cov. Est.                     Robust          Robust          Robust
R-squared                     0.0219          0.4354          0.0415
R-Squared (Within)           -0.0079         -0.0545          0.0039
R-Squared (Between)           0.0369          0.7078          0.1097
R-Squared (Overall)           0.0219          0.6224          0.0743
F-statistic                   32.052          1103.3          19.704
P-value (F-stat)              0.0000          0.0000          0.0000
==================== ============ =============== ===============
const                         10.034                          10.505
                             (36.530)                        (11.347)
child                         0.2492          1.1537          0.1519
                             (5.5341)        (22.019)        (1.5085)
drinkly                      -0.8938          0.2823          0.2146
                            (-6.7595)        (1.8830)        (1.1076)
hrsusu                        0.0047          0.0256         -0.0009
                             (1.6075)        (9.6378)       (-0.2930)
hsize                        -0.0322          0.2711         -0.1195
                            (-0.9243)        (7.6333)       (-2.7889)
married                      -1.7562          3.0070         -1.1849
                            (-8.7490)        (14.378)       (-2.4480)
retired                      -0.1540          1.7569          0.3354
                            (-0.8294)        (10.133)        (1.6557)
wealth                   -3.204e-06      -1.175e-06      -4.952e-07
```

```
                        (-1.3834)        (-1.2720)        (-0.3472)
female                                                     0.7596

schadj                                                    -0.0954

urban                                                     -0.8932
                                                         (-0.4777)
age                                                       -0.0038

mchild                                                    -0.0818
                                                         (-0.6497)
mdrinkly                                                  -0.3436
                                                         (-1.1305)
mfemale                                                    0.7596
                                                          (0.7402)
mhrsusu                                                    0.0144
                                                          (2.2109)
mhsize                                                     0.0742
                                                          (0.9665)
mmarried                                                  -0.2094
                                                         (-0.3676)
mretired                                                   0.4217
                                                          (1.0368)
mschadj                                                   -0.0954

murban                                                    -0.8932
                                                         (-0.4528)
mwealth                                                  -4.734e-06
                                                         (-1.2967)
mage                                                      -0.0038
                                                         (-0.1686)
------------------------------------------------------------------

T-stats reported in parentheses
```

0.10 Pregunta 8: ¿Qué modelo es mejor?

Dado que el CRE no resultó mayormente significativo comparado con el modelo de efectos aleatorios, este último entregaría una mejor estimación. Sin embargo, el modelo de efectos aleatorios no favorece la estimación como el modelo de efectos fijos, lo cual fue corroborado con el test de Hausman.

Tarea 2

Instrucciones

Los resultados de los ejericicios propuestos se deben entregar como un notebook por correo electronico a *juan.caro@uni.lu* el dia 3/10 hasta las 21:00.

Es importante considerar que el código debe poder ejecutarse en cualquier computadora con la data

original del repositorio. Recordar la convencion para el nombre de archivo ademas de incluir en su documento titulos y encabezados por seccion. La data a utilizar es **charls.csv**.

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)

Preguntas:

1. 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.

2. 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 significado.

3. 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.

4. 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.

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?

6. 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?

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?

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?