Tarea2_Cuevas_Reyes

October 14, 2022

[ ]:
```
!pip install linearmodels
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting linearmodels
  Downloading
linearmodels-4.25-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.5
MB)
     |                | 1.5 MB 5.1 MB/s
Requirement already satisfied: statsmodels>=0.11 in
/usr/local/lib/python3.7/dist-packages (from linearmodels) (0.12.2)
Collecting formulaic
  Downloading formulaic-0.5.2-py3-none-any.whl (77 kB)
     |                | 77 kB 5.5 MB/s
Collecting property-cached>=1.6.3
  Downloading property_cached-1.6.4-py2.py3-none-any.whl (7.8 kB)
Requirement already satisfied: Cython>=0.29.21 in /usr/local/lib/python3.7/dist-
packages (from linearmodels) (0.29.32)
Requirement already satisfied: patsy in /usr/local/lib/python3.7/dist-packages
(from linearmodels) (0.5.2)
Collecting mypy-extensions>=0.4
  Downloading mypy_extensions-0.4.3-py2.py3-none-any.whl (4.5 kB)
Requirement already satisfied: scipy>=1.2 in /usr/local/lib/python3.7/dist-
packages (from linearmodels) (1.7.3)
Requirement already satisfied: numpy>=1.16 in /usr/local/lib/python3.7/dist-
packages (from linearmodels) (1.21.6)
Collecting pyhdfe>=0.1
  Downloading pyhdfe-0.1.0-py3-none-any.whl (18 kB)
Requirement already satisfied: pandas>=0.24 in /usr/local/lib/python3.7/dist-
packages (from linearmodels) (1.3.5)
Requirement already satisfied: python-dateutil>=2.7.3 in
/usr/local/lib/python3.7/dist-packages (from pandas>=0.24->linearmodels) (2.8.2)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-
packages (from pandas>=0.24->linearmodels) (2022.2.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-
packages (from python-dateutil>=2.7.3->pandas>=0.24->linearmodels) (1.15.0)
Requirement already satisfied: astor>=0.8 in /usr/local/lib/python3.7/dist-
packages (from formulaic->linearmodels) (0.8.1)
```

```
Collecting typing-extensions>=4.2.0
  Downloading typing_extensions-4.3.0-py3-none-any.whl (25 kB)
Requirement already satisfied: wrapt>=1.0 in /usr/local/lib/python3.7/dist-
packages (from formulaic->linearmodels) (1.14.1)
Collecting interface-meta>=1.2.0
  Downloading interface_meta-1.3.0-py3-none-any.whl (14 kB)
Collecting graphlib-backport>=1.0.0
  Downloading graphlib_backport-1.0.3-py3-none-any.whl (5.1 kB)
Requirement already satisfied: cached-property>=1.3.0 in
/usr/local/lib/python3.7/dist-packages (from formulaic->linearmodels) (1.5.2)
Installing collected packages: typing-extensions, interface-meta, graphlib-
backport, pyhdfe, property-cached, mypy-extensions, formulaic, linearmodels
  Attempting uninstall: typing-extensions
    Found existing installation: typing-extensions 4.1.1
    Uninstalling typing-extensions-4.1.1:
```

[ ]:
```
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 seaborn as sns
import linearmodels.panel as lmp

from sklearn.preprocessing import OneHotEncoder

%matplotlib inline
```

1 Pregunta 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.

[ ]:
```
try:
    charls = pd.read_csv("../data/charls.csv")
except:
    try:
        charls = pd.read_csv("../../data/charls.csv")
    except:
        charls = pd.read_csv("https://raw.githubusercontent.com/juancaros/
 ↪LAB-MAA/main/data/charls.csv")
    charls["inid"] = charls["inid"].astype("int")
```

Los datos se se intentan cargas de dos manera, buscar el archivo llamado “charls.csv” en el com-

putador donde se corra el codigo o directamente desde el repositorio de github del curso

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34371 entries, 0 to 34370
Data columns (total 15 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   cesd       34371 non-null  int64
 1   child      34371 non-null  int64
 2   drinkly    34371 non-null  object
 3   female     34371 non-null  int64
 4   hrsusu     34371 non-null  float64
 5   hsize      34371 non-null  int64
 6   inid       34371 non-null  int64
 7   intmonth   34371 non-null  int64
 8   married    34371 non-null  int64
 9   retired    34371 non-null  int64
 10  schadj     34371 non-null  int64
 11  urban      34371 non-null  int64
 12  wave       34371 non-null  int64
 13  wealth     34371 non-null  float64
 14  age        34371 non-null  int64
dtypes: float64(2), int64(12), object(1)
memory usage: 3.9+ MB
```

[ ]:
```
charls.loc[charls.drinkly == "0.None", "drinkly"] = 0
charls.loc[charls.drinkly == "1.Yes", "drinkly"] = 1
charls.loc[charls.drinkly == ".m:missing", "drinkly"] = None
charls.dropna(inplace = True)


aux = charls.value_counts("inid").to_frame().reset_index()
aux.columns = ["inid", "counts"]
inids = aux[aux.counts == 3]["inid"]
charls = charls[charls.inid.isin(inids)]

categories = pd.get_dummies(charls["intmonth"])
categories.columns = ["jan", "feb", "jun", "jul","ago", "sept", "oct", "nov",␣
 ↪"dic"]
```

[ ]:
```
categories.sum()
```

[ ]:
```
jan       54
feb       42
jun       60
jul     4994
```

```
ago     4093
sept     344
oct      147
nov      188
dic      113
dtype: int64
```

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

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10035 entries, 0 to 10055
Data columns (total 15 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   cesd      10035 non-null  int64
 1   child     10035 non-null  int64
 2   drinkly   10035 non-null  object
 3   female    10035 non-null  int64
 4   hrsusu    10035 non-null  float64
 5   hsize     10035 non-null  int64
 6   inid      10035 non-null  int64
 7   intmonth  10035 non-null  int64
 8   married   10035 non-null  int64
 9   retired   10035 non-null  int64
 10  schadj    10035 non-null  int64
 11  urban     10035 non-null  int64
 12  wave      10035 non-null  int64
 13  wealth    10035 non-null  float64
 14  age       10035 non-null  int64
dtypes: float64(2), int64(12), object(1)
memory usage: 1.2+ MB
```

R: Se identifican los siguientes grupos de variables:

- Dicotómicas: drinkly (*), married, retired, urban, female
- Categórica: intmonth (**), INID (***), wave
- Numérica: cesd (****), child, hrsusu, hsize, wealth, age

Además se rescata que la variable intmonth, correspondiente al mes de toma de resultados de la muestra, tienen casi todos sus valores en el mes 7 y 8 (Julio y Agosto) y muy pocos valores en febrero, como se puede ver en el anexo 1, es por eso que se decide omitir el mes de febrero (por lo explicado en **)

Se encontró que habían algunos individuos en la muestra que tenían más valores asociados a un mismo identificador INID (Anexo 3), por los que estos fueron sacados de la muestra, quedando asi un total de 10035 filas, correspondientes a 3345 individuos, los que aparecen en cada uno de los periodos de encuestación quedando una base de datos de panel balanceado (no hay atrición) .

Finalmente se prepararon los datos calculando la media de las variables numéricas, exceptuando cesd, ya que es la variable explicada de los modelos, y age ya que este dato es estático para cada

individuo dependiendo de la edad que tenía al comenzar las encuestas. Luego se le da estructura de panel, como se ve a continuación

(*) Para el caso de drinkly esta variable fue transformada de a dicotomica 0-1 en vez de 0.None-1.Yes, en este proceso se encontraron valores faltantes los cuales fueron eliminados

() **intmonth fue separada en variables dicotómicas donde cada una toma el valor si el dato fue tomado en el mes representado por esta o no, con el objetivo de ver si este tiene algún impacto o no en el modelo.** añadir un poquito mas de explicacion** Es importante notar que para la elaboración de los modelos es importante no poner la variable intmonth y omitir uno de los meses, ya que en caso contrario las variables pasan a formar combinaciones lineales entre si y causa problemas en el modelo por lo explicado anteriormente.

(***) Se transforma la variable INID de float a int para facilitar el arreglo del problema de que el identificador no sea unico

(****) Se investigó el significado del cesd, donde según el articulo adjunto se especifica que a mayor número más sintomas depresivos presentan los encuestados. https://www.sciencedirect.com/science/article/pii/S0033350621002572#appsec1

[ ]:
```
X = charls[['child', 'hrsusu', 'hsize', 'wealth']]
Xm = X.groupby(charls["inid"]).transform("mean")
Xid = charls[['inid', 'wave', 'cesd', 'drinkly', 'married',
              'retired', 'urban', 'female', "intmonth", 'child', 'hrsusu',␣
 ↪'hsize',
              'wealth', 'age']]

Xc=pd.DataFrame(
    np.c_[Xid, categories, Xm],
    columns=['inid', 'wave', 'cesd', 'drinkly', 'married', 'retired', 'urban',
             'female', "intmonth",
             'child', 'hrsusu', 'hsize', 'wealth', 'age',
             'mchild', 'mhrsusu', 'mhsize', 'mwealth',
             "jan", "feb", "jun","jul","ago", "sept", "oct", "nov", "dic"])

#set panel structure
Xc = Xc.set_index(["inid","wave"])
# Xc.describe()
```

2 Pregunta 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.

- Consideraciones previas: Se propone que las variable hrsusu y retired no pueden coexistir, debido a que si los individuos estan retirados la variable retired tomará valor 1 y hara que la variable hrsusu tambien tome el valor 0 ya que indica la cantidad de horas trabajadas a la semana, habiendo una alta correlación, se puede apreciar mejor en el anexo 2, que indica su alta correlacion, por lo que se decide eliminar la variable retired. Lo mismo pasa con la

variable age y child que tienen una alta correlación, donde se elige omitir age

[ ]:
```
y=Xc['cesd']
X=Xc[['drinkly', 'married', 'urban', 'female', 'child',
      'hrsusu', 'hsize', 'wealth']]

X=sm.add_constant(X)

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

```
                          PooledOLS Estimation Summary
================================================================================
Dep. Variable:                   cesd   R-squared:                        0.0616
Estimator:                  PooledOLS   R-squared (Between):              0.0929
No. Observations:               10035   R-squared (Within):              -0.0003
Date:                Mon, Oct 03 2022   R-squared (Overall):              0.0616
Time:                        02:45:04   Log-likelihood                -3.237e+04
Cov. Estimator:                Robust
                                        F-statistic:                      82.276
Entities:                        3345   P-value                           0.0000
Avg Obs:                       3.0000   Distribution:                 F(8,10026)
Min Obs:                       3.0000
Max Obs:                       3.0000   F-statistic (robust):             77.831
                                        P-value                           0.0000
Time periods:                       3   Distribution:                 F(8,10026)
Avg Obs:                       3345.0
Min Obs:                       3345.0
Max Obs:                       3345.0

                             Parameter Estimates
================================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
--------------------------------------------------------------------------------
const          9.5485     0.2781     34.330     0.0000      9.0033      10.094
drinkly       -0.0635     0.1431    -0.4438     0.6572     -0.3441      0.2170
married       -1.5940     0.1956    -8.1481     0.0000     -1.9775     -1.2105
urban         -2.0081     0.1289    -15.581     0.0000     -2.2607     -1.7554
female         1.9443     0.1358     14.319     0.0000      1.6781      2.2104
child          0.1590     0.0441     3.6066     0.0003      0.0726      0.2455
hrsusu         0.0028     0.0023     1.1881     0.2348     -0.0018      0.0074
hsize         -0.0560     0.0342    -1.6366     0.1018     -0.1230      0.0111
wealth     -2.757e-06  2.181e-06    -1.2644     0.2061  -7.032e-06   1.517e-06
================================================================================
```

3 Interpretación

Dada la significancia de las variables se estima que la constante es significativa por lo que se podría representar una tendencia al aumento en la escala de salud mental, las demas variables que son significativas son "married" y "urban", lo que implica que el estar casado y vivir en zona urbana que disminuye en la escala de salud mental. Mientras que las variables "female" y "child", el ser mujer indica aumento en la escala de salud mental y el numero de hijos, a mayor cantidad mayor puntacion en la escala de la salud mental.

4 Pregunta 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.

Asumiendo cov(,X) != 0

[ ]:
```
y=Xc['cesd']
X=Xc[['drinkly', 'married', 'child',
      'hrsusu', 'hsize', '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:                        0.0034
Estimator:                   PanelOLS   R-squared (Between):              0.0142
No. Observations:               10035   R-squared (Within):               0.0034
Date:                Mon, Oct 03 2022   R-squared (Overall):              0.0106
Time:                        02:48:23   Log-likelihood                -2.72e+04
Cov. Estimator:                Robust
                                        F-statistic:                      3.8510
Entities:                        3345   P-value                           0.0008
Avg Obs:                       3.0000   Distribution:                  F(6,6684)
Min Obs:                       3.0000
Max Obs:                       3.0000   F-statistic (robust):             3.2589
                                        P-value                           0.0034
Time periods:                       3   Distribution:                  F(6,6684)
Avg Obs:                       3345.0
Min Obs:                       3345.0
Max Obs:                       3345.0

                             Parameter Estimates
================================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
```

```
------------------------------------------------------------------------------
const          9.9426     0.5373     18.504     0.0000      8.8893      10.996
drinkly        0.2079     0.1886     1.1028     0.2702     -0.1617      0.5776
married       -1.2113     0.5051    -2.3980     0.0165     -2.2016     -0.2211
child          0.1533     0.0960     1.5972     0.1103     -0.0349      0.3415
hrsusu        -0.0026     0.0026    -0.9864     0.3240     -0.0077      0.0026
hsize         -0.1235     0.0440    -2.8080     0.0050     -0.2098     -0.0373
wealth     -5.037e-07  8.412e-07    -0.5988     0.5493  -2.153e-06   1.145e-06
==============================================================================

F-test for Poolability: 3.8462
P-value: 0.0000
Distribution: F(3344,6684)

Included effects: Entity
```

4.1 Inclusión de variables:

Se utilizaron las mismas variables propuestas para la resolución en OLS, con la singularidad de que en esta ocasión se eliminaron las variable urban y female, debido a que estas eran absorvidas completamente por los efectos incluidos.

4.2 Interpretación

Dada la significancia de las variables se estima que el estar casado/a y el tamaño del hogar tienen un impacto en la salud mental de los encuestados con un 95% de confianza, los cuales tienen un impacto negativo para la métrica. Recordar que un impacto negativo para la métrica implica una disminución en los sintomas depresivos (y la gravedad de estos). Distinto es el caso de la constante, ya que si las demas variables se mantienen constantes, existe una tendencia al aumento en la metrica y por lo tanto aumento en sintomas depresivos.

5 Pregunta 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.

Asumiendo cov(,X) = 0

[ ]:
```
y=Xc['cesd']
X=Xc[['drinkly', 'married', 'urban', 'female', 'child',
      'hrsusu', 'hsize', 'wealth']]

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:	0.0337
Estimator:	RandomEffects	R-squared (Between):	0.0902
No. Observations:	10035	R-squared (Within):	0.0027
Date:	Mon, Oct 03 2022	R-squared (Overall):	0.0608
Time:	02:45:16	Log-likelihood	-2.924e+04
Cov. Estimator:	Robust		
		F-statistic:	43.765
Entities:	3345	P-value	0.0000
Avg Obs:	3.0000	Distribution:	F(8,10026)
Min Obs:	3.0000		
Max Obs:	3.0000	F-statistic (robust):	42.582
		P-value	0.0000
Time periods:	3	Distribution:	F(8,10026)
Avg Obs:	3345.0		
Min Obs:	3345.0		
Max Obs:	3345.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	9.5882	0.3382	28.353	0.0000	8.9253	10.251
drinkly	0.0628	0.1503	0.4178	0.6761	-0.2319	0.3575
married	-1.4898	0.2424	-6.1449	0.0000	-1.9650	-1.0145
urban	-2.0460	0.1772	-11.544	0.0000	-2.3934	-1.6986
female	1.9895	0.1815	10.964	0.0000	1.6338	2.3452
child	0.1585	0.0540	2.9334	0.0034	0.0526	0.2643
hrsusu	-7.508e-05	0.0022	-0.0336	0.9732	-0.0045	0.0043
hsize	-0.0875	0.0347	-2.5181	0.0118	-0.1555	-0.0194
wealth	-1.381e-06	1.127e-06	-1.2261	0.2202	-3.59e-06	8.271e-07

5.1 Inclusión de variables:

Se utilizaron las mismas variables propuestas para la resolución en Pooled OLS.

5.2 Interpretación

Dada la significancia de las variables se estima que el estar casado/a (-), el tamaño del hogar (-), que este tenga una ubicación urbana (-), ser mujer (+) y el numero de hijos (+) tienen un impacto en la salud mental de los encuestados con un 95% de confianza. En donde el impacto que tienen estos en el puntaje esperado para el CESD fue descrito en parentesis, cabe recalcar que un mayor CESD implica mayores numero y gravedad de sintomas depresivos. Para la constante ocurren los mismo efectos que en el modelo de efectos fijos, en este caso la constante(+) toma un valor un poco mas alta.

6 Pregunta 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?

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

```
                          Model Comparison
==========================================================================
                                    FE                RE         Pooled
--------------------------------------------------------------------------
Dep. Variable                     cesd              cesd           cesd
Estimator                     PanelOLS    RandomEffects      PooledOLS
No. Observations                 10035             10035          10035
Cov. Est.                       Robust            Robust         Robust
R-squared                       0.0034            0.0337         0.0616
R-Squared (Within)              0.0034            0.0027        -0.0003
R-Squared (Between)             0.0142            0.0902         0.0929
R-Squared (Overall)             0.0106            0.0608         0.0616
F-statistic                     3.8510            43.765         82.276
P-value (F-stat)                0.0008            0.0000         0.0000
=====================     ============   ===============   ============
const                       9.9426***         9.5882***      9.5485***
                            (18.504)          (28.353)       (34.330)
drinkly                        0.2079            0.0628        -0.0635
                            (1.1028)          (0.4178)      (-0.4438)
married                    -1.2113**         -1.4898***     -1.5940***
                           (-2.3980)         (-6.1449)      (-8.1481)
child                          0.1533         0.1585***      0.1590***
                            (1.5972)          (2.9334)       (3.6066)
hrsusu                        -0.0026        -7.508e-05         0.0028
                           (-0.9864)         (-0.0336)       (1.1881)
hsize                      -0.1235***         -0.0875**        -0.0560
                           (-2.8080)         (-2.5181)      (-1.6366)
wealth                     -5.037e-07        -1.381e-06     -2.757e-06
                           (-0.5988)         (-1.2261)      (-1.2644)
urban                                       -2.0460***     -2.0081***
                                            (-11.544)      (-15.581)
female                                       1.9895***      1.9443***
                                             (10.964)       (14.319)
====================== ============== ================ ==============
Effects                        Entity
--------------------------------------------------------------------------

T-stats reported in parentheses
```

6.1 Comentarios

Algunas diferencias que se pueden observar son que, por un lado el modelo de efectos fijos permite un menor número de variables, dado que "urban" y "female" generan conflictos en dicho modelo al ser estas fijas para cada individuo. Esto se debe además a la forma en que cada uno está formulado, el FE resta la media de cada individuo a los datos de este, por lo que elimina variables fijas, el Pooled OLS en cambio resuelve directamente los minimos cuadrados pero agregando variables dummy para cada periodo, con lo que se estima una especie de efecto promedio, mientras que el RE si bien aplica una transformación a las variables, no pierde la capacidad de trabajar variables fijas en el tiempo, como lo hace el FE.

El más adecuado para responder la pregunta de investigación puede ser el Fixed Effects , tal como indica el test de hausman. lo que sugiere que no se está cumpliendo el supuesto del random efects. Cabe destacat que para realizar el test se tienen que eliminar algunas variables de la especificaicón del modelo Random Effects, las cuales además son significativas en este, por lo que puede que este modelo aporte información nueva que de todas formas sea relevante.

Existe un concenso entre los 3 modelos en que el estar casado genera una reducción en los sintomas depresivos de los encuestados. Además para los Pooled OLS y Random Effects el vivir en una zona urbana, el numero de hijos y ser mujer también tienen impactos significativos.

Adicionalmente, se trató de modelar que pasaba con el mes en que se tomaron las muestras, como se muestra en el anexo 4, sin embargo no se tomó en cuenta para los modelos finales esta especificación debido a que se producían muchas diferencias entre los modelos, además de existir una disparidad en la cantdiad de encuestados por mes

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

y=Xc['cesd']
X=Xc[['drinkly', 'married', 'child',
      'hrsusu', 'hsize', 'wealth']]

X=sm.add_constant(X)

model=lmp.RandomEffects(y,X)
re_aux=model.fit(cov_type="robust")

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

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

```
/usr/local/lib/python3.7/dist-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)

Hausman Test: chi-2 = 37.385122013511406, df = 6, p-value =
1.4812104449515861e-06
```

7 Pregunta 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?

[ ]:
```
y=Xc['cesd']
X=Xc[['drinkly', 'married', 'urban', 'female', 'child',
      'hrsusu', 'hsize', 'wealth','mchild', 'mhrsusu', 'mhsize', '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:                        0.0343
Estimator:                RandomEffects   R-squared (Between):              0.0901
No. Observations:                 10035   R-squared (Within):               0.0035
Date:                Mon, Oct 03 2022   R-squared (Overall):              0.0610
Time:                        03:31:27   Log-likelihood                -2.924e+04
Cov. Estimator:                Robust
                                        F-statistic:                      29.625
Entities:                        3345   P-value                           0.0000
Avg Obs:                       3.0000   Distribution:                F(12,10022)
Min Obs:                       3.0000
Max Obs:                       3.0000   F-statistic (robust):             28.909
                                        P-value                           0.0000
Time periods:                       3   Distribution:                F(12,10022)
Avg Obs:                       3345.0
Min Obs:                       3345.0
Max Obs:                       3345.0

                             Parameter Estimates
================================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
--------------------------------------------------------------------------------
```

```
const          9.5479     0.3464     27.563     0.0000      8.8689      10.227
drinkly        0.0595     0.1504     0.3958     0.6922     -0.2353      0.3543
married       -1.4977     0.2424    -6.1793     0.0000     -1.9728     -1.0226
urban         -2.0461     0.1779    -11.502     0.0000     -2.3948     -1.6974
female         1.9894     0.1814     10.968     0.0000      1.6339      2.3450
child          0.1570     0.0540     2.9059     0.0037      0.0511      0.2628
hrsusu      3.377e-05     0.0022     0.0151     0.9879     -0.0043      0.0044
hsize         -0.0816     0.0350    -2.3297     0.0198     -0.1503     -0.0129
wealth     -1.393e-06  1.133e-06    -1.2297     0.2188  -3.613e-06   8.274e-07
mchild        -0.9394     0.6385    -1.4712     0.1413     -2.1911      0.3122
mhrsusu       -1.2091     0.6943    -1.7416     0.0816     -2.5701      0.1518
mhsize        -0.1889     0.7443    -0.2538     0.7996     -1.6479      1.2700
mwealth        0.0792     0.1095     0.7235     0.4694     -0.1354      0.2939
================================================================================
```

##Inclusión de variables: Se utilizaron las mismas variables propuestas para la resolución en Pooled OLS, agregando la media individual para "child", "hrsusu", "hsize" y "wealth" a fin de estimar la correlación entre el factor heterogeneo no observado.

##Interpretación Al igual que en el modelo random effects se ecnuentra que estar casado/a (-), el tamaño del hogar (-), que este tenga una ubicación urbana (-), ser mujer (+) y el numero de hijos (+) tienen un impacto en la salud mental de los encuestados con un 95% de confianza. En donde el impacto que tienen estos en el puntaje esperado para el CESD fue descrito en parentesis, cabe recalcar que un mayor CESD implica mayores numero y gravedad de sintomas depresivos.

Además se encuentra que mhrsusu tiene una correlación significativa en un 90% con el factor heterogeneo no observado

8 Pregunta 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?

Tomando en cuenta que:

y = X + [+ X] + U

Y con la estimación realizada anteriormente, se identifica que la heterogeneidad fija en el tiempo existe y tiene una correlación significativa (90%) con las horas de trabajo de los encuestados

9 Pregunta 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?

Preferimos el modelo de efectos aleatorios correlacionados, ya que este busca modelar la correlación con la heterogeneidad no observada, la cual es distinta de cero, como se mostró en pasos anteriores con e test de hausman; sin dejar de lado las variables fijas como lo habría hecho un modelo de efectos fijos.

```
[ ]: print(lmp.compare({"FE": fe, "RE": re, "cre": cre, "Pooled": OLS}, stars =␣
       ↪True))
```

```
                                Model Comparison
================================================================================
=========
                                   FE                RE               cre
Pooled
--------------------------------------------------------------------------------
---------
Dep. Variable                    cesd              cesd              cesd
cesd
Estimator                   PanelOLS     RandomEffects     RandomEffects
PooledOLS
No. Observations               10035             10035             10035
10035
Cov. Est.                     Robust            Robust            Robust
Robust
R-squared                     0.0034            0.0337            0.0343
0.0616
R-Squared (Within)            0.0034            0.0027            0.0035
-0.0003
R-Squared (Between)           0.0142            0.0902            0.0901
0.0929
R-Squared (Overall)           0.0106            0.0608            0.0610
0.0616
F-statistic                   3.8510            43.765            29.625
82.276
P-value (F-stat)              0.0008            0.0000            0.0000
0.0000
====================     ============   ===============   ===============
============
const                      9.9426***         9.5882***         9.5479***
9.5485***
                            (18.504)          (28.353)          (27.563)
(34.330)
drinkly                       0.2079            0.0628            0.0595
-0.0635
                            (1.1028)          (0.4178)          (0.3958)
(-0.4438)
married                    -1.2113**         -1.4898***        -1.4977***
-1.5940***
                           (-2.3980)         (-6.1449)         (-6.1793)
(-8.1481)
child                         0.1533         0.1585***         0.1570***
0.1590***
                            (1.5972)          (2.9334)          (2.9059)
```

```
(3.6066)
hrsusu                         -0.0026        -7.508e-05         3.377e-05
0.0028
                              (-0.9864)         (-0.0336)          (0.0151)
(1.1881)
hsize                        -0.1235***         -0.0875**          -0.0816**
-0.0560
                              (-2.8080)         (-2.5181)          (-2.3297)
(-1.6366)
wealth                       -5.037e-07        -1.381e-06         -1.393e-06
-2.757e-06
                              (-0.5988)         (-1.2261)          (-1.2297)
(-1.2644)
urban                                          -2.0460***         -2.0461***
-2.0081***
                                               (-11.544)          (-11.502)
(-15.581)
female                                          1.9895***          1.9894***
1.9443***
                                                (10.964)           (10.968)
(14.319)
mchild                                                              -0.9394
                                                                   (-1.4712)
mhrsusu                                                             -1.2091*
                                                                   (-1.7416)
mhsize                                                              -0.1889
                                                                   (-0.2538)
mwealth                                                              0.0792
                                                                    (0.7235)
======================== ============== ================= =================
==============
Effects                            Entity
-------------------------------------------------------------------------------
---------

T-stats reported in parentheses
```

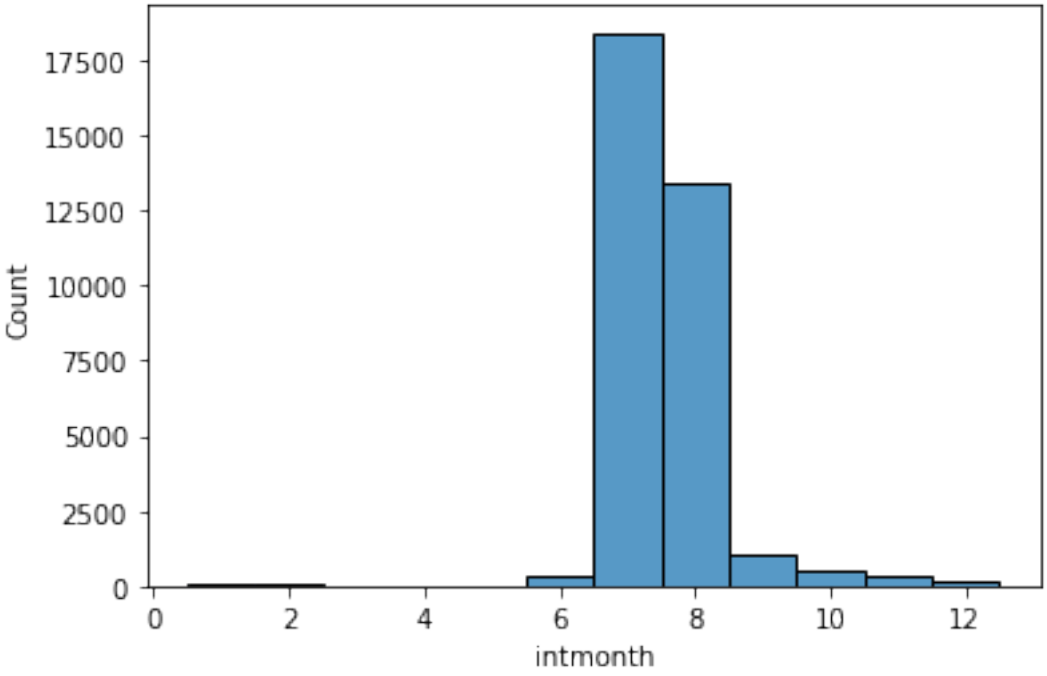
10 Anexos

10.1 Anexo 1: Variable intmonth

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

[ ]:
```
intmonth
7     18414
8     13424
```

```
9       1013
10       479
11       366
6        298
12       195
1        105
2         54
3         23
dtype: int64
```

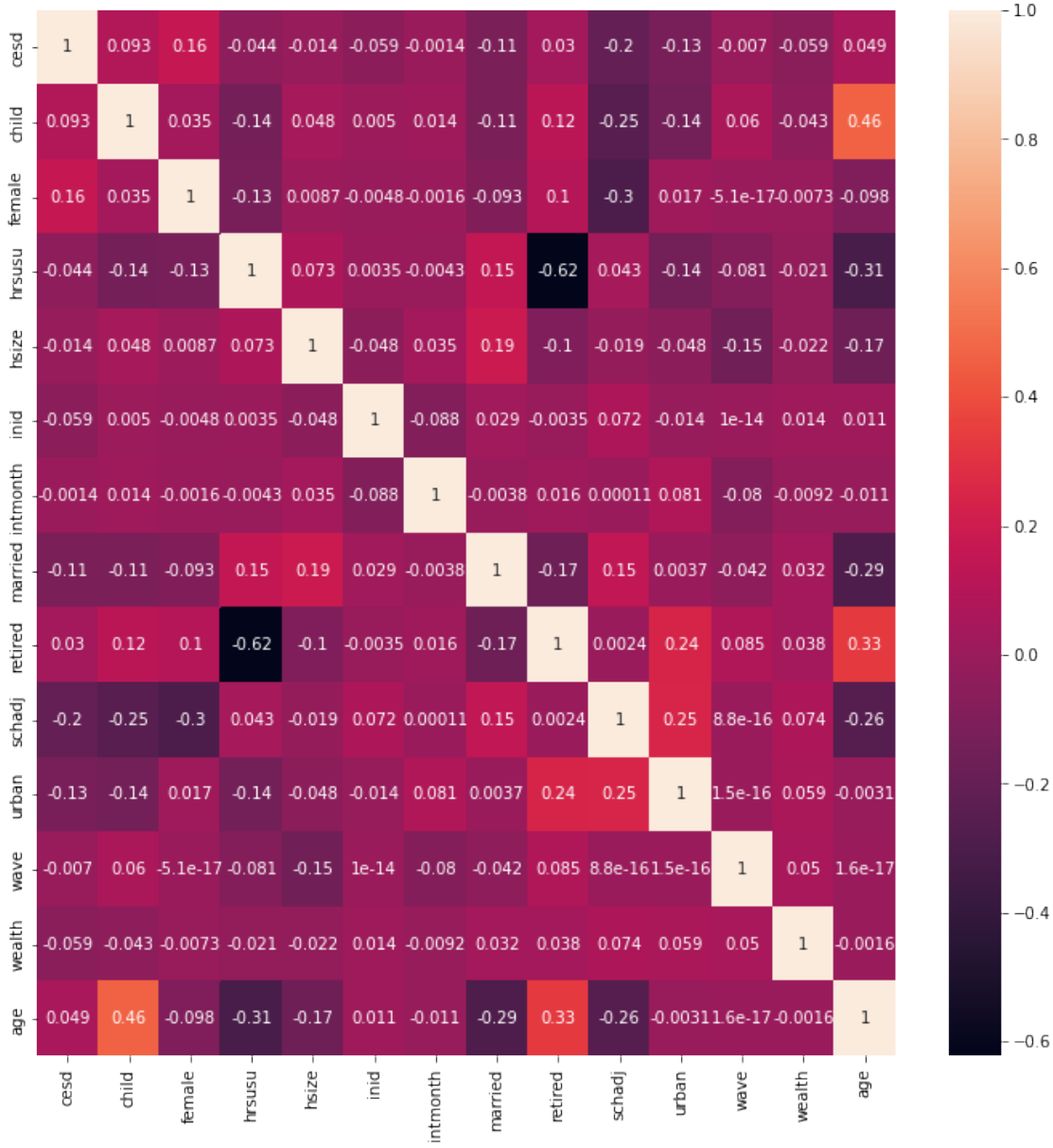


11 Anexo 2: Correlación entre variables

Se encuentra una correlación relevante entre retired-hsusu y entre child-age

[ ]:
```
f, ax = plt.subplots(figsize = (12,12))
sns.heatmap(charls.corr(), annot=True)
```

[ ]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb1f3e0afd0>
```



12 Anexo 3: inid que se repiten más de 3 veces

[ ]:
```
charls.value_counts("inid")
```

[ ]:
```
inid
296331000000    468
108259000000    396
270402000000    393
298245000000    378
```

```
244604000000    369
                …
46746323001       3
46746324001       3
46746324002       3
46746325001       3
56346116002       3
Length: 3459, dtype: int64
```

[ ]:
```
aux = charls.value_counts("inid").to_frame().reset_index()
aux.columns = ["inid", "counts"]
inids = aux[aux.counts < 4]["inid"]
inids
```

[ ]:
```
107     60440220002
108     60440214001
109     60440207001
110     60440207002
111     60440209001
           …
3454    46746323001
3455    46746324001
3456    46746324002
3457    46746325001
3458    56346116002
Name: inid, Length: 3352, dtype: int64
```

12.1 Anexo 4: Modelos incluyendo los meses

[ ]:
```
y=Xc['cesd']
X=Xc[['drinkly', 'married', 'urban', 'female', 'child',
      'hrsusu', 'hsize', 'wealth', "jan", "feb", "jun", "ago", "sept", "oct",␣
 ↪"nov", "dic"]]
X=sm.add_constant(X)

model=lmp.PooledOLS(y,X)
OLS2=model.fit(cov_type="robust")


X=Xc[['drinkly', 'married', 'child',
      'hrsusu', 'hsize', 'wealth', "jan", "feb", "jun", "ago", "sept", "oct",␣
 ↪"nov", "dic"]]

X=sm.add_constant(X)

model=lmp.PanelOLS(y,X, entity_effects=True)
fe2=model.fit(cov_type="robust")
```

```
X=Xc[['drinkly', 'married', 'urban', 'female', 'child',
      'hrsusu', 'hsize', 'wealth', "jan", "feb", "jun","ago", "sept", "oct",␣
 ↪"nov", "dic"]]
X=sm.add_constant(X)

model=lmp.RandomEffects(y,X)
re2=model.fit(cov_type="robust")

print(lmp.compare({"Pooled": OLS2, "FE": fe2, "RE": re2}, stars = True))
```

Model Comparison

	Pooled	FE	RE
Dep. Variable	cesd	cesd	cesd
Estimator	PooledOLS	PanelOLS	RandomEffects
No. Observations	10035	10035	10035
Cov. Est.	Robust	Robust	Robust
R-squared	0.0653	0.0074	0.0370
R-Squared (Within)	-0.0002	0.0074	0.0061
R-Squared (Between)	0.0984	0.0126	0.0928
R-Squared (Overall)	0.0653	0.0109	0.0637
F-statistic	43.743	3.5781	24.035
P-value (F-stat)	0.0000	0.0000	0.0000
const	9.5402***	10.030***	9.6289***
	(33.931)	(18.549)	(28.349)
drinkly	-0.0577	0.2357	0.0762
	(-0.4032)	(1.2477)	(0.5073)
married	-1.5850***	-1.2060**	-1.4929***
	(-8.0824)	(-2.3817)	(-6.1642)
urban	-2.1246***		-2.1087***
	(-16.213)		(-11.840)
female	1.9489***		1.9974***
	(14.367)		(11.030)
child	0.1543***	0.1390	0.1568***
	(3.4962)	(1.4423)	(2.9062)
hrsusu	0.0028	-0.0025	-0.0001
	(1.1757)	(-0.9608)	(-0.0456)
hsize	-0.0606*	-0.1108**	-0.0825**
	(-1.7540)	(-2.4907)	(-2.3538)
wealth	-2.711e-06	-5.041e-07	-1.393e-06
	(-1.2518)	(-0.5980)	(-1.2264)
jan	-1.5264**	-0.7575	-1.0003
	(-2.0475)	(-1.0440)	(-1.5591)
feb	-0.7181	-1.6754**	-1.2756*

	(-0.9240)	(-1.9910)	(-1.8315)
jun	-0.8855	0.1980	-0.2701
	(-1.0126)	(0.2688)	(-0.3651)
ago	0.1363	-0.1850	-0.0502
	(1.0446)	(-1.4286)	(-0.4321)
sept	0.4172	0.0039	0.1560
	(1.2464)	(0.0129)	(0.5583)
oct	-2.0034***	-2.0441***	-2.0337***
	(-4.7205)	(-4.4627)	(-5.4211)
nov	0.8480*	0.2764	0.4898
	(1.9285)	(0.7059)	(1.3525)
dic	2.0631***	0.0377	0.7775
	(3.2102)	(0.0637)	(1.3965)
Effects		Entity	

T-stats reported in parentheses