

# Tarea 2 Final 1 DV

May 23, 2023

## 1 Tarea\_2\_LAB\_MAA\_Diego\_Valdebenito

### Tarea 2

#### *Instrucciones*

Los resultados de los ejercicios propuestos se deben entregar como un notebook por correo electrónico a [juancaros@udec.cl](mailto:juancaros@udec.cl) el día 12/5 hasta las 21:00.

Es importante considerar que el código debe poder ejecutarse en cualquier computadora con la data original del repositorio. Recordar la convención para el nombre de archivo además de incluir en su documento títulos y encabezados por sección. La data a utilizar es **enia.csv**.

Las variables tienen la siguiente descripción:

- *ID*: firm unique identifier
- *year*: survey year
- *tamano*: 1 large, 2 medium, 3 small, 4 micro (función de las ventas y el número de trabajadores)
- *sales*: sales (in log of 1,000 CLP)
- *age*: firm age at time of survey
- *foreign*: non-domestic firm (binary)
- *export*: production for export (binary)
- *workers*: log of number of workers
- *fomento*: firm receives public incentives (binary)
- *iyd*: firm does I+D (binary)
- *impuestos*: taxes (in million US)
- *utilidades*: firm revenue (in million US)

Para este analisis consideraremos tamaño como una variable continua, que identifica el tamaño de la empresa.

Preguntas:

1. Cargar la base de datos *enia.csv* en el ambiente. Identifique los tipos de datos que se encuentran en la base, realice estadísticas descriptivas sobre las variables importantes (Hint: Revisar la distribuciones, datos faltantes, outliers, etc.) y limpie las variables cuando sea necesario. Para las preguntas 2-8 **EXCLUYA LA VARIABLE FOMENTO DE SU ANALISIS**.
2. Ejecute un modelo Pooled OLS para explicar el numero de trabajadores. Seleccione las variables independientes a incluir en el modelo final e interprete su significado.
3. Ejecute un modelo de efectos fijos para explicar el numero de trabajadores. Seleccione las variables independientes a incluir en el modelo final e interprete su significado.
4. Ejecute un modelo de efectos aleatorios para explicar el numero de trabajadores. Seleccione las variables independientes 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 investigación y por qué? ¿Qué variables resultaron ser robustas a la especificación?
6. Ejecute un modelo de efectos aleatorios correlacionados (CRE) para explicar el numero de trabajadores. Seleccione las variables independientes 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 numero de trabajadores?
8. Usando sus respuestas anteriores, que modelo prefiere? que se puede inferir en general respecto del efecto de las variables explicativas sobre el numero de trabajadores?
9. Considere que la variable *fomento* es una politica publica donde aleatoriamente se selecciono un grupo de empresas para recibir recursos financieros dedicados a incentivar I+D. Utilizando fomento como instrumento, estime un modelo en dos etapas para entender el impacto causal de la inversion en I+D sobre el numero de trabajadores, y compare versus el modelo MCO (puntos adicionales para hacerlo en un contexto de panel).

```
[3]: 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 statsmodels.stats.api as sms
import sklearn
import scipy
import linearmodels.panel as lmp
import pytwoway as tw
import bipartitepandas as bpd
```

```
import seaborn as sns
from linearmodels.iv import IV2SLS

%matplotlib inline
```

Intel(R) Extension for Scikit-learn\* enabled (<https://github.com/intel/scikit-learn-intelex>)

### 1.0.1 Pregunta 1

Cabe señalar desde un principio que, para todas las preguntas, el nivel de significatividad utilizado es el clásico, es decir, el de 0.05 o 5%. Se ajusto “tamano” como variable numerica (continua o float) tal como se indica en el enunciado, y se paso utilidades a logaritmo, ajustandola de tal forma de que los valores sean siempre positivos. Cuando la variable “export” tenga celdas vacias, se eliminaran. Se encuentran valores muy extremos en la variable “utilidades” hacia la derecha, por lo cual se decide eliminarlos. Se ignorara la variable “fomento” tal como indican las instrucciones.

```
[7]: # enia data
enia = pd.read_csv('C:/Users/equipo/Desktop/Tarea_2_Laboratorio/
↳LAB-MAA-main_Tarea_2/data/enia.csv')
enia['tamano'] = enia['tamano'].astype(float)
enia = enia.dropna(subset=["export"])

enia['utilidades'] = enia['utilidades']/1000000
enia = enia.loc[enia["utilidades"] <= 0.04]
constante = abs(enia["utilidades"].min()) + 1
enia["lutilidades"] = np.log(enia["utilidades"] + constante)
enia.loc[enia['lutilidades'].isnull(), 'lutilidades'] = 0

enia['P1'] = np.where(enia['year'] == 2007, 1, 0)
enia['P2'] = np.where(enia['year'] == 2009, 1, 0)
enia['P3'] = np.where(enia['year'] == 2013, 1, 0)
enia['P4'] = np.where(enia['year'] == 2015, 1, 0)
enia['P5'] = np.where(enia['year'] == 2017, 1, 0)

enia.dropna(inplace=True)
enia.reset_index(drop=True, inplace=True)

enia.head()
```

```
[7]:
```

	ID	year	tamano	sales	age	foreign	export	workers	fomento	\
0	100003	2007	1.0	7.046558	22	1	1.0	3.486855	0	
1	100003	2009	1.0	7.875563	24	1	1.0	3.504607	0	
2	100003	2013	1.0	7.437399	23	1	1.0	4.691621	1	
3	100003	2015	1.0	7.356472	30	1	1.0	4.682614	1	
4	100003	2017	1.0	4.014772	32	1	1.0	4.611691	1	

```
iyd  impuestos    utilidades  lutilidades  P1  P2  P3  P4  P5
```

0	1	1.231345	7.113891e-06	0.000251	1	0	0	0	0
1	0	8.762233	3.397611e-05	0.000278	0	1	0	0	0
2	1	0.001886	3.137265e-06	0.000247	0	0	1	0	0
3	0	0.411670	1.298413e-06	0.000246	0	0	0	1	0
4	0	0.000721	1.949247e-09	0.000244	0	0	0	0	1

```
[6]: print(enia.dtypes)

# Detectar variables binarias en el DataFrame "charls"
binaries = []
for column in enia.columns:
    if enia[column].nunique() == 2:
        binaries.append(column)

# Imprimir las variables binarias
print("Variables binarias en enia:")
print(binaries)
```

```
ID                int64
year              int64
tamano            float64
sales             float64
age              int64
foreign           int64
export            float64
workers           float64
fomento           int64
iyd               int64
impuestos         float64
utilidades        float64
lutilidades       float64
P1                int32
P2                int32
P3                int32
P4                int32
P5                int32
dtype: object
Variables binarias en enia:
['foreign', 'export', 'fomento', 'iyd', 'P1', 'P2', 'P3', 'P4', 'P5']
```

### 1.0.2 Tipos de variables y variables a analizar

Tal como se ordena en el código anterior, se imprimen los tipos de variables de la base de datos, donde “year”, “age”, “foreign”, “fomento” e “iyd” son enteras, de las cuales “foreign”, “export”, “fomento” e “iyd” son binarias; y las variables continuas son: “tamano”, “sales”, “export”, “workers”, “impuestos”, “utilidades” y la variable creada a partir de una transformacion de utilidades: “lutilidades”.

Las variables que se consideran a priori como importantes son “year”, para saber si hubo perdidas

de datos a través de los años, “workers”, que es la variable de estudio (logaritmo del número de trabajadores), “sales”, porque se desea ver si las ventas de las firmas influyen en el número de trabajadores, lo cual a priori se supone que sí ya que se sabe popularmente que las grandes firmas son las que más trabajadores tienen, “foreign”, porque es de interés saber cuáles, si las firmas locales o las extranjeras, tienen más trabajadores, “iyd”, ya que se desea ver si el hecho de hacer investigación y desarrollo está relacionado con el hecho de que haya más trabajadores en la firma, a priori se cree que sí ya que se piensa que si la firma tiene fondos para hacer investigación y desarrollo es porque es grande en términos de número de trabajadores, también es de interés la variable “utilidades” ya que se desea ver si las empresas con más utilidades son más grandes en términos de personal o no, para ello se analizará utilizando la transformación logarítmica de la variable anterior para que se pueda analizar las utilidades con valores siempre positivos.

A continuación, se muestran los estadísticos descriptivos y gráficamente los comportamientos variables de estudio y explicativas antes mencionadas:

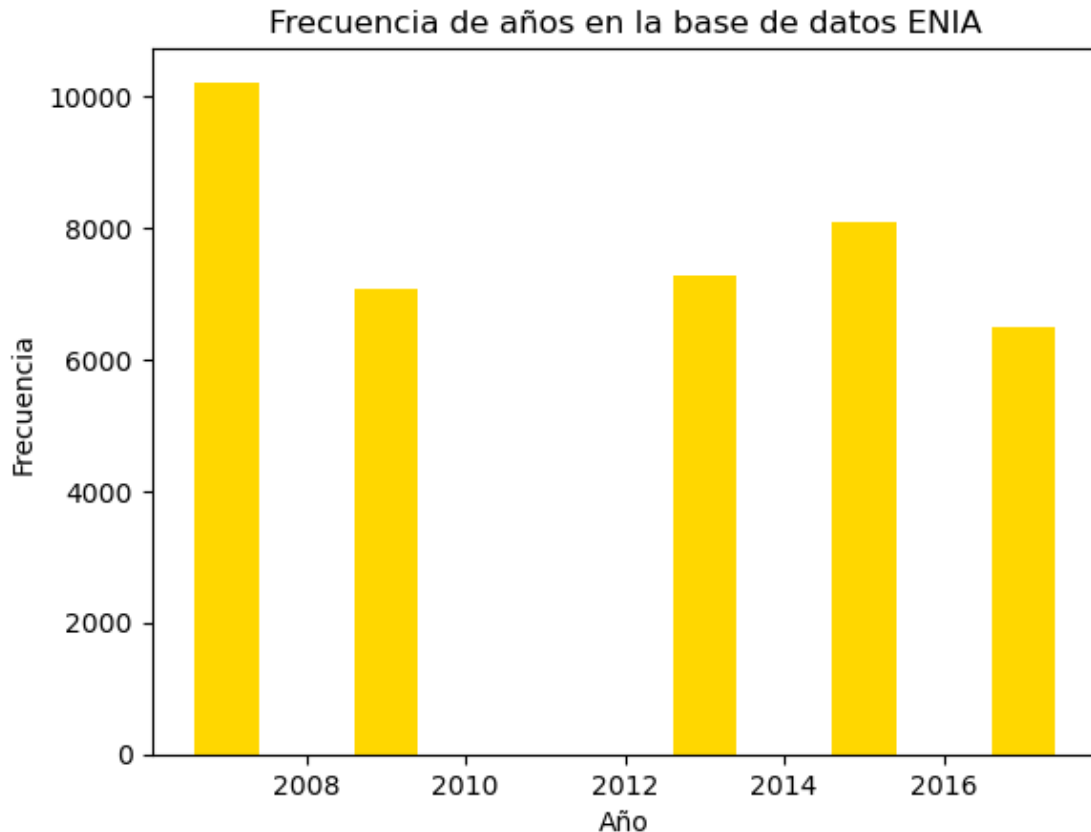
### 1.0.3 Year

```
[8]: # contar la frecuencia de cada año en la variable "year" de la base de datos ↵
      ↪ "enia"
year_count = enia["year"].value_counts()

# crear un gráfico de barras con los años en el eje x y la frecuencia en el eje ↵
      ↪ y
plt.bar(x=year_count.index, height=year_count.values, color="gold")

# añadir etiquetas al gráfico
plt.title("Frecuencia de años en la base de datos ENIA")
plt.xlabel("Año")
plt.ylabel("Frecuencia")

# mostrar el gráfico
plt.show()
```



Se puede visualizar que en el año 2007 hay más datos que en los demás años, bajando en 2009, subiendo de 2013 y un poco más en 2015 para luego volver a bajar en 2017, por lo cual se detecta atrición, es decir, pérdida de observaciones en una muestra de datos a lo largo del tiempo, lo que puede afectar la calidad y validez de los resultados del análisis.

#### 1.0.4 Workers

```
[9]: # set a grey background (use sns.set_theme() if seaborn version 0.11.0 or above)
sns.set(style="darkgrid")

# crear una figura compuesta de dos objetos matplotlib.Axes (ax_box y ax_hist)
f, (ax_box, ax_hist) = plt.subplots(2, sharex=True,
    ↳ gridspec_kw={"height_ratios": (.15, .85)})

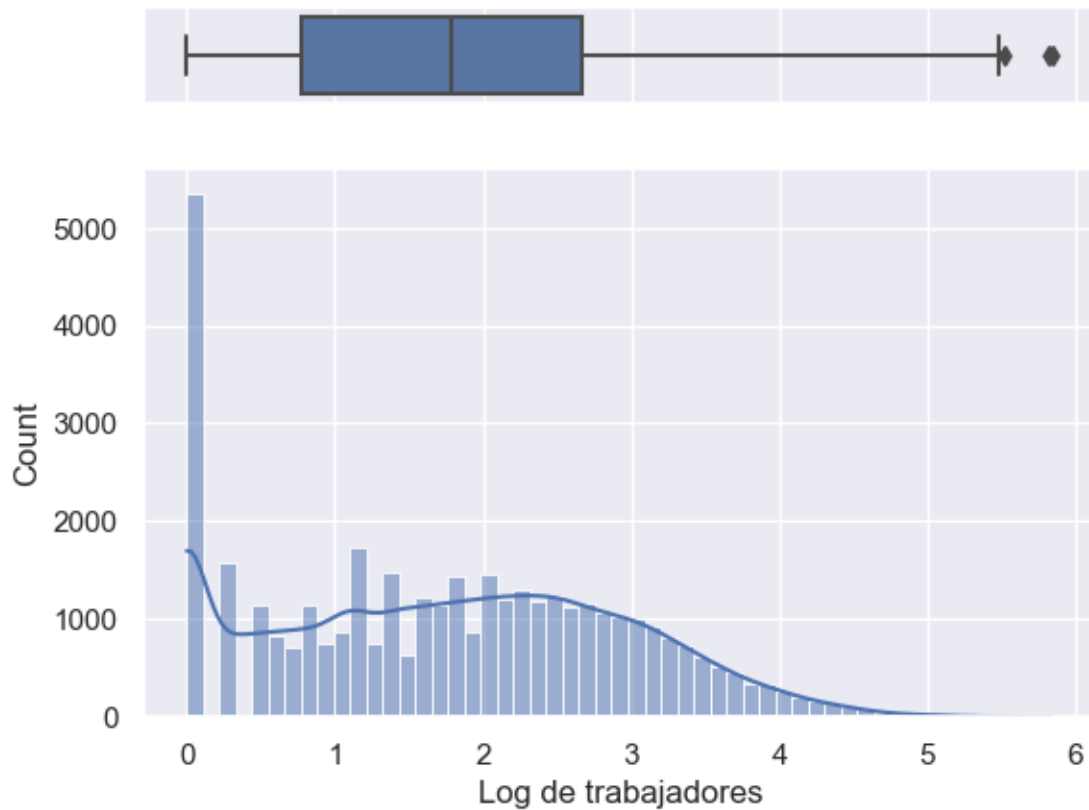
# asignar un gráfico a cada objeto ax
sns.boxplot(enia["workers"], ax=ax_box)
sns.histplot(data=enia, x="workers", ax=ax_hist, kde=True)
ax_hist.set_xlabel('Log de trabajadores')

# Eliminar el nombre del eje x del boxplot
```

```
ax_box.set(xlabel='')

# mostrar el gráfico
plt.show()
```

C:\ProgramData\Anaconda3\lib\site-packages\seaborn\\_decorators.py:36:  
FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.  
warnings.warn(



El boxplot muestra una mediana menor a 2 log trabajadores, es decir, hay un sesgo positivo en la distribución, y la mayoría de los datos se encuentran aproximadamente entre log 0.8 trabajadores y log 2.6 trabajadores, el valor mínimo es de 0 log trabajadores y el valor máximo es de 5.85 log trabajadores, presentándose datos atípicos en el orden de los 5 a 6 log trabajadores.

### 1.0.5 Foreign

```
[10]: # filtrar por foreign == 0
eniaf1 = enia[enia["foreign"] == 0]
eniaf2 = enia[enia["foreign"] == 1]

# set a grey background (use sns.set_theme() if seaborn version 0.11.0 or above)
sns.set(style="darkgrid")

# crear una figura compuesta de dos objetos matplotlib.Axes (ax_box y ax_hist)
f, ((ax_box1, ax_box2), (ax_hist1, ax_hist2)) = plt.subplots(2, 2, sharex=True,
    ↪ gridspec_kw={"height_ratios": (.15, .85)})

# asignar un gráfico a cada objeto ax
sns.boxplot(eniaf1["workers"], ax=ax_box1)
ax_box1.set_xlabel('')
sns.histplot(data=eniaf1, x="workers", ax=ax_hist1, kde=True)
ax_hist1.set_xlabel('Log de trabajadores, firma local')
sns.boxplot(eniaf2["workers"], ax=ax_box2)
ax_box2.set_xlabel('')
sns.histplot(data=eniaf2, x="workers", ax=ax_hist2, kde=True)
ax_hist2.set_xlabel('Log de trabajadores, firma extranjera')

# Eliminar el nombre del eje x del boxplot
ax_box.set(xlabel='')

# mostrar el gráfico
plt.show()
```

C:\ProgramData\Anaconda3\lib\site-packages\seaborn\\_decorators.py:36:

FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.

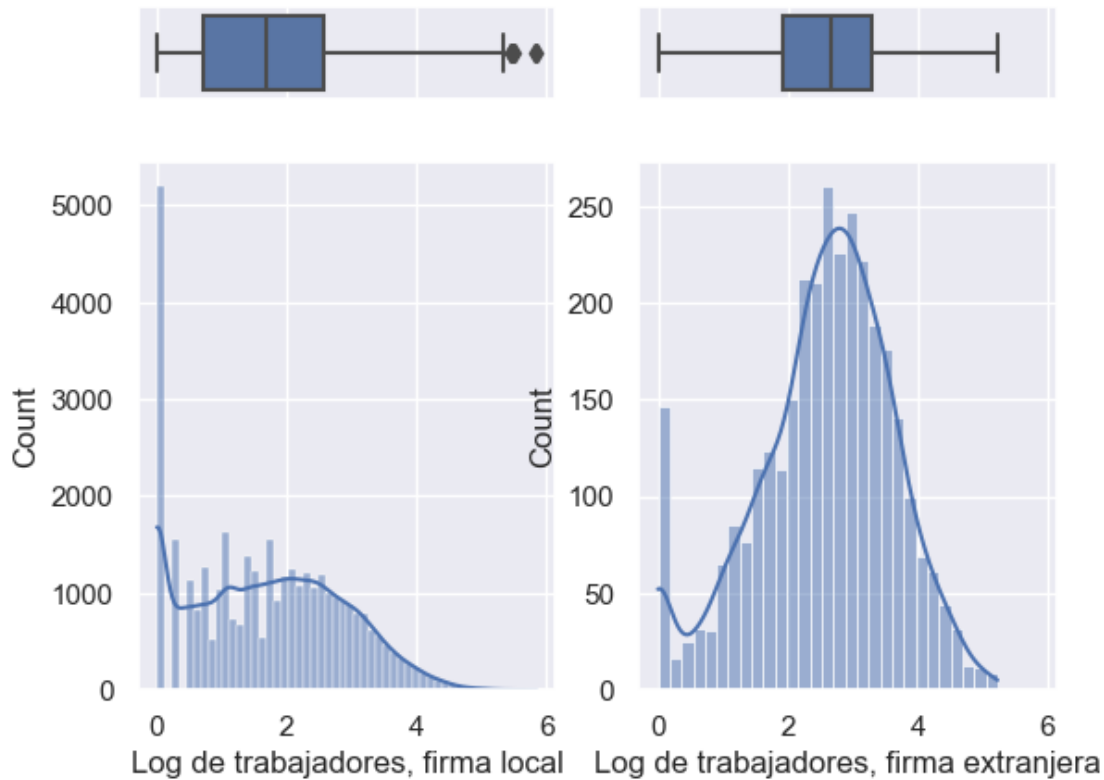
warnings.warn(

C:\ProgramData\Anaconda3\lib\site-packages\seaborn\\_decorators.py:36:

FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.

warnings.warn(





Se puede visualizar claramente que hay diferencias en los casos donde la firma es extranjera o no, pues, en el caso donde las firmas son locales, el boxplot muestra una mediana menor a 2 log trabajadores, es decir, hay un sesgo positivo en la distribución, y la mayoría de los datos se encuentran aproximadamente entre log 0.7 trabajadores y log 2.6 trabajadores, el valor mínimo es de 0 log trabajadores y el valor máximo es de 5.85 log trabajadores, presentándose datos atípicos en el orden de los 5 a 6 log trabajadores (bastante similar al análisis de la variable workers en general, debido a que la mayoría de las observaciones fueron de firmas locales). En contraste, cuando la firma es extranjera, se muestra una distribución asintóticamente simétrica, donde la mediana esta en casi 3 log trabajadores, la mayoría de los datos se encuentran entre los 1.9 y los 3.3 trabajadores aproximadamente y no se observan datos atípicos, el valor mínimo es de 0 log trabajadores y el valor máximo es de 5.2 trabajadores.

Luego de esto, es posible afirmar que cuando la firma es extranjera, hay más trabajadores, por lo cual es a priori una variable importante para explicar el número de trabajadores en los modelos que se verán mas adelante.

### 1.0.6 I+D

```
[11]: # filtrar por foreign == 0
eniai1 = enia[enia["iyd"] == 0]
eniai2 = enia[enia["iyd"] == 1]
```

```

# set a grey background (use sns.set_theme() if seaborn version 0.11.0 or above)
sns.set(style="darkgrid")

# crear una figura compuesta de dos objetos matplotlib.Axes (ax_box y ax_hist)
f, ((ax_box1, ax_box2), (ax_hist1, ax_hist2)) = plt.subplots(2, 2, sharex=True,
↳ gridspec_kw={"height_ratios": (.15, .85)})

# asignar un gráfico a cada objeto ax
sns.boxplot(eniai1["workers"], ax=ax_box1)
ax_box1.set_xlabel('')
sns.histplot(data=eniai1, x="workers", ax=ax_hist1, kde=True)
ax_hist1.set_xlabel('Log de trabajadores, NO I+D')
sns.boxplot(eniai2["workers"], ax=ax_box2)
ax_box2.set_xlabel('')
sns.histplot(data=eniai2, x="workers", ax=ax_hist2, kde=True)
ax_hist2.set_xlabel('Log de trabajadores, SI I+D')

# Eliminar el nombre del eje x del boxplot
ax_box.set_xlabel='')

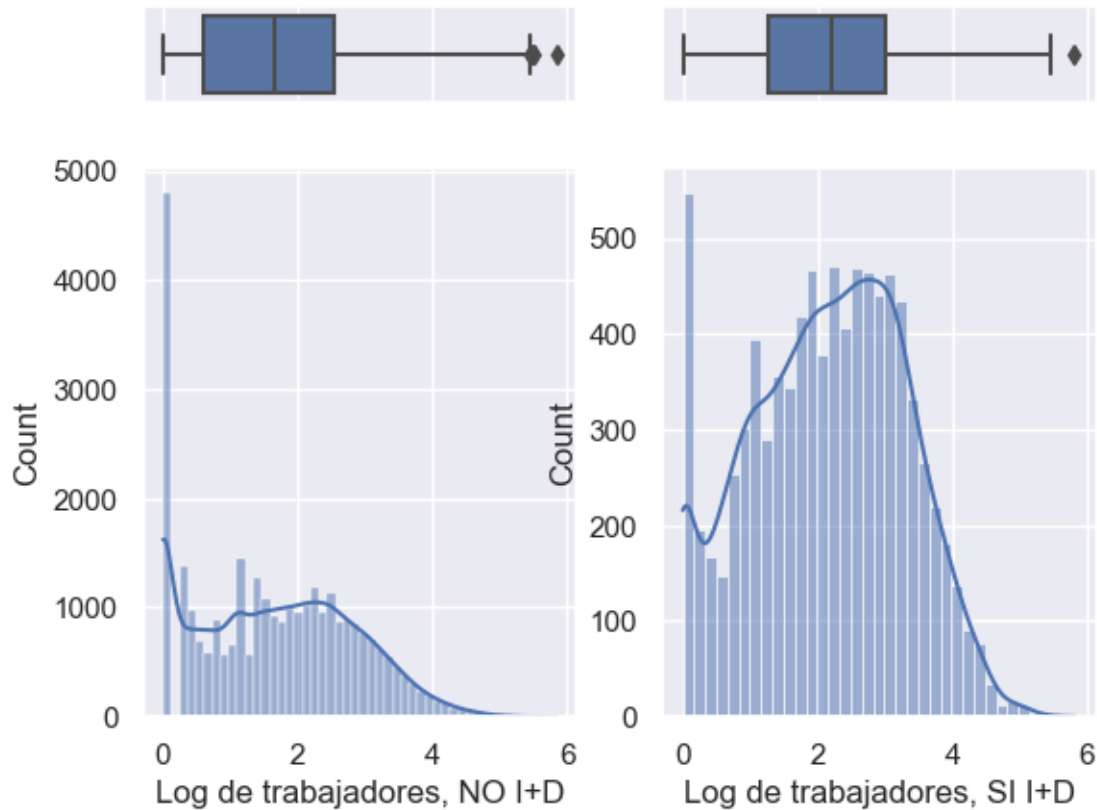
# mostrar el gráfico
plt.show()

```

```

C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36:
FutureWarning: Pass the following variable as a keyword arg: x. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or
misinterpretation.
  warnings.warn(
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36:
FutureWarning: Pass the following variable as a keyword arg: x. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or
misinterpretation.
  warnings.warn(

```



Se observa un comportamiento parecido en la distribución de log trabajadores cuando se hace o no se hace Investigación y Desarrollo (I+D) y también en los valores mínimos y máximos y en los cuartiles, sin embargo, la mediana de cuando si se realiza I+D está más a la derecha que cuando no se realiza I+D, por lo que es posible afirmar que el hecho de que haya I+D en la empresa implica que esta tenga levemente más trabajadores.

### 1.0.7 Log utilidades

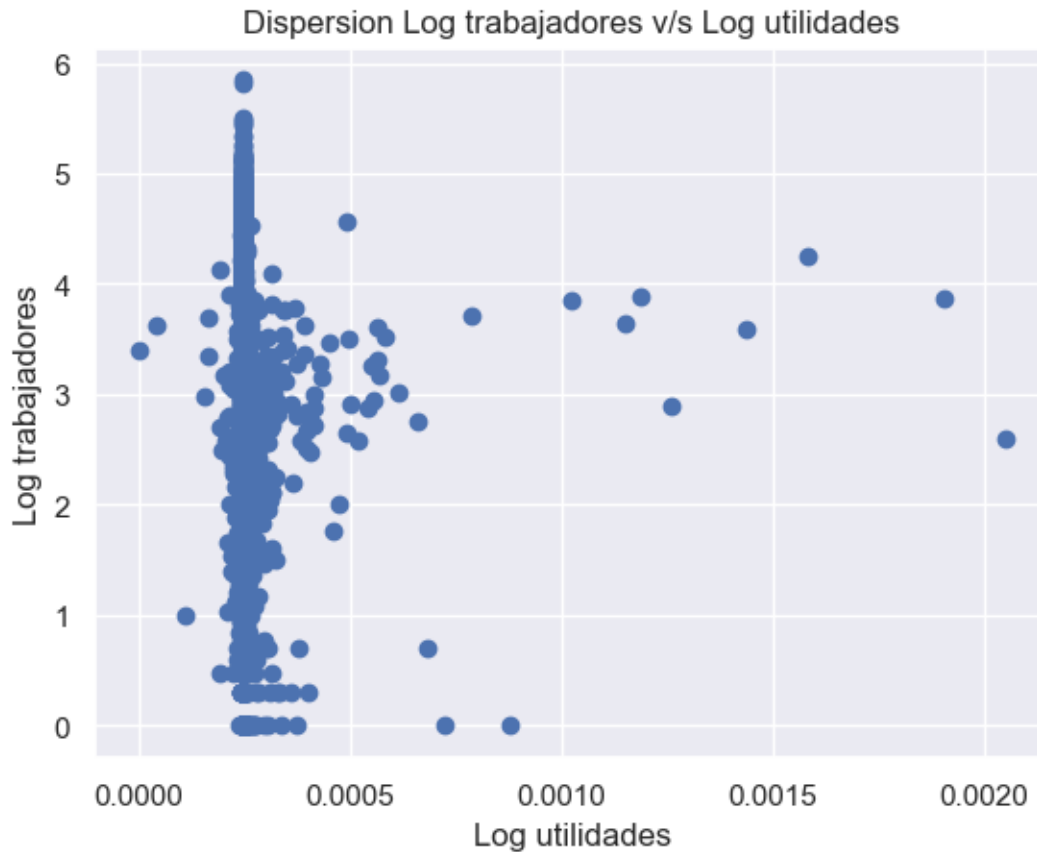
```
[13]: # Seleccionar las variables de interés
x = enia["lutilidades"]
y = enia["workers"]

# Crear el gráfico de dispersión
plt.scatter(x, y)

# Establecer los títulos de los ejes y del gráfico
plt.xlabel("Log utilidades")
plt.ylabel("Log trabajadores")
plt.title("Dispersion Log trabajadores v/s Log utilidades")

# Mostrar el gráfico
```

```
plt.show()
```



En el gráfico de dispersión mostrado anteriormente se puede visualizar que la mayoría de las firmas con menos utilidades tienen un número de mediano a bajo de log trabajadores, es posible afirmar que, a baja utilidad, la cantidad de trabajadores en las firmas es menor. Luego, se observa que una cantidad de trabajadores moderada implica mayores utilidades. Finalmente, se observan casos aislados donde las utilidades son muy altas en donde el número de trabajadores es moderado.

### 1.0.8 Sales

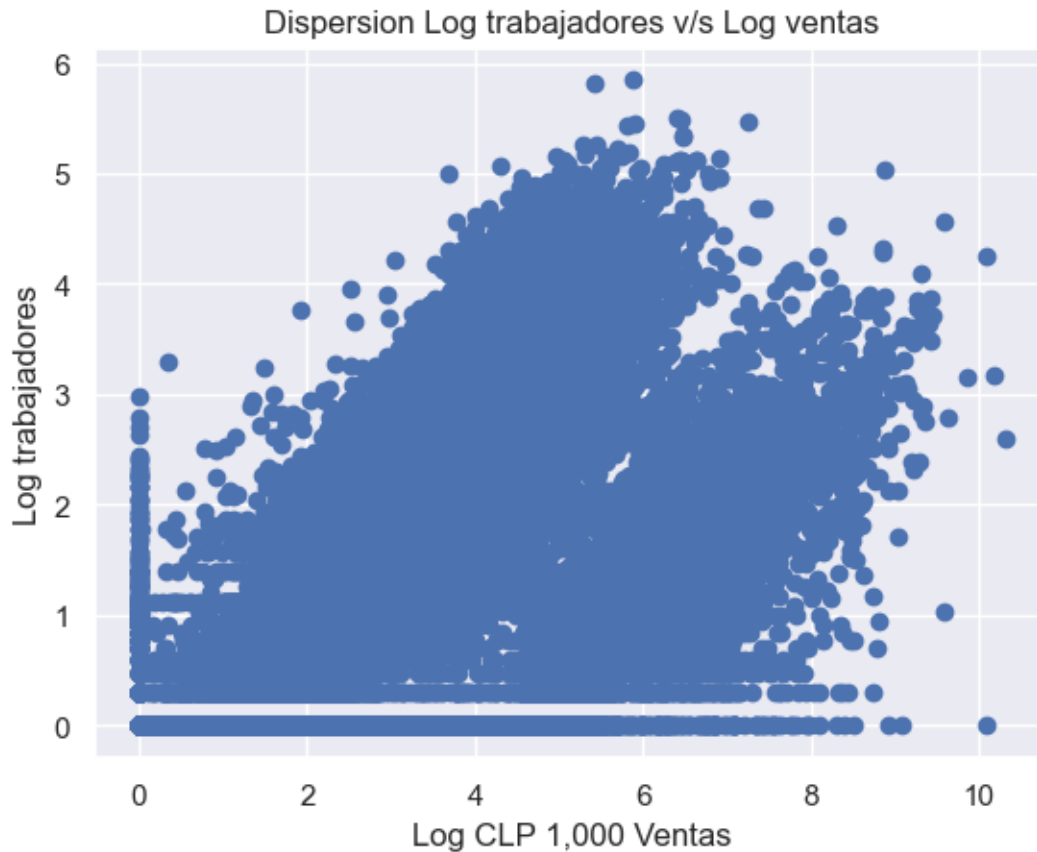
```
[14]: # Seleccionar las variables de interés
x = enia["sales"]
y = enia["workers"]

# Crear el gráfico de dispersión
plt.scatter(x, y)

# Establecer los títulos de los ejes y del gráfico
plt.xlabel("Log CLP 1,000 Ventas")
plt.ylabel("Log trabajadores")
```

```
plt.title("Dispersion Log trabajadores v/s Log ventas")

# Mostrar el gráfico
plt.show()
```



Se puede visualizar una nube de puntos en el diagrama de dispersión anterior, lo que es un indicio de que existe poca correlación entre el logaritmo de número de trabajadores y el logaritmo de ventas. A continuación, se verán las correlaciones entre variables para corroborar esta información y ver si hay indicio de multicolinealidad.

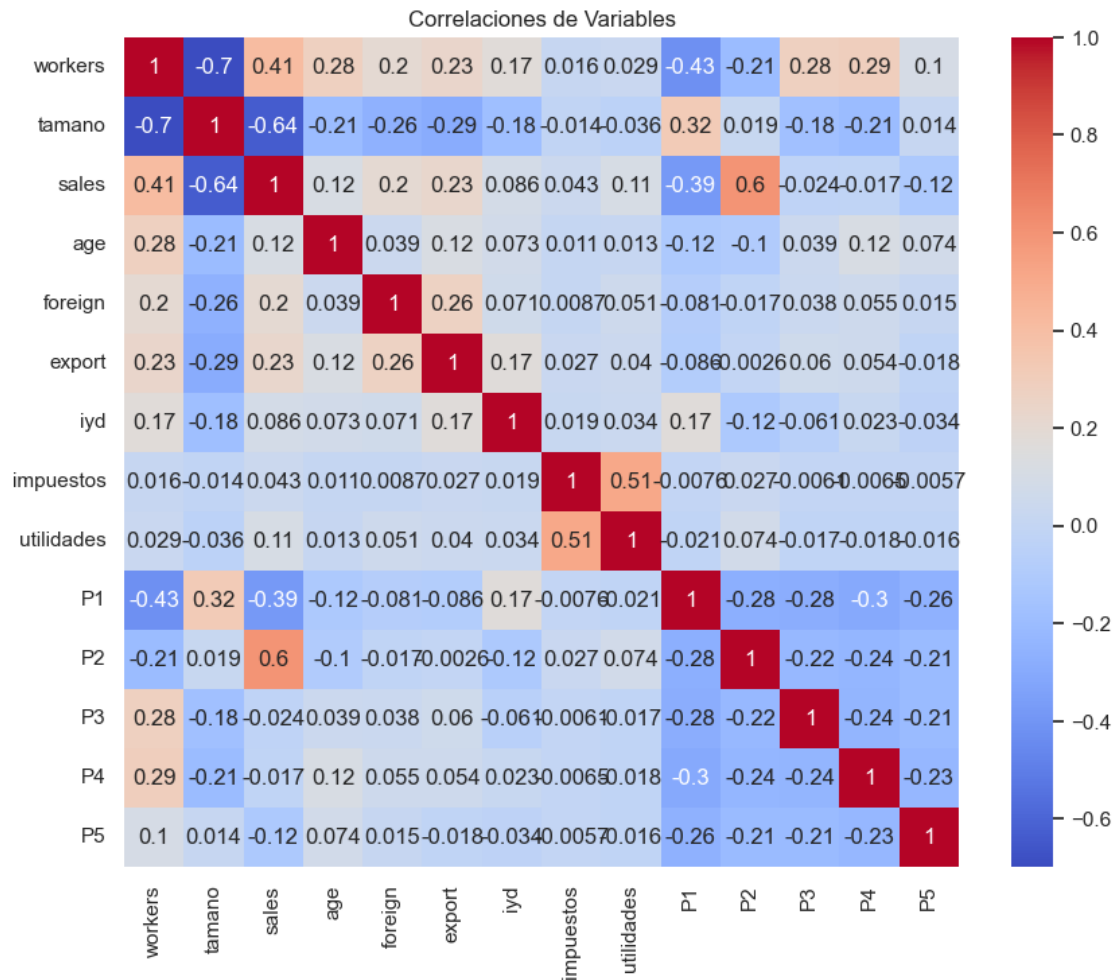
### 1.0.9 Pregunta 2

En la siguiente figura, se muestran las correlaciones entre las variables, se pondrán en el modelo todas las variables que no tengan problemas de multicolinealidad.

```
[24]: # Seleccionar las variables de interés
variables = ['workers', 'tamano', 'sales', 'age', 'foreign', 'export', 'iyd',
            'impuestos', 'utilidades', 'P1', 'P2', 'P3', 'P4',
            'P5']
df = enia[variables]
```

```
# Calcular las correlaciones
correlations = df.corr()

# Crear un heatmap de las correlaciones
plt.figure(figsize=(10, 8))
sns.heatmap(correlations, annot=True, cmap='coolwarm')
plt.title('Correlaciones de Variables')
plt.show()
```



En la matriz de correlaciones se puede observar claramente una correlación negativa entre las variables “tamaño” y “workers”, y era de esperarse que ocurriera esto, ya que la variable “tamaño” es función del número de trabajadores y mientras más grande sea el valor de esta variable, significa que la empresa es más pequeña en términos del número de trabajadores; tras esto, se quitará esta variable de los modelos que se analizarán más abajo. Además, cabe señalar que es conveniente quitar “tamaño” de los modelos porque también está altamente relacionada con la variable ‘sales’, ya que, por definición, también es función de las ventas de las firmas. Se ve que entre el Logaritmo

de trabajadores y el Logaritmo de Ventas existe una correlación que a pesar de que no es tan baja, no deja de serlo, que puede ser el motivo del aspecto grafico del diagrama de dispersión mostrado anteriormente en la descripción de los datos.

A continuación, se construirán las variables necesarias para los distintos modelos, considerando que se está trabajando con data de panel

### 1.0.10 Construcción de variables

```
[16]: X = enia.drop('workers', axis=1, inplace=False)
Xm = enia.groupby(by = 'ID').transform('mean').drop('year', axis=1)
Xm.columns = ['m{}'.format(column) for column in Xm.columns]
Xc = pd.merge(enia,Xm, left_index=True, right_index=True)
Xc = Xc.set_index(['ID','year'])

Xc.describe()
```

```
[16]:
```

	tamano	sales	age	foreign	export \
count	39103.000000	39103.000000	39103.000000	39103.000000	39103.000000
mean	2.248804	3.574035	15.305066	0.081861	0.111194
std	1.153087	1.692547	12.488489	0.274156	0.314376
min	1.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	2.337597	7.000000	0.000000	0.000000
50%	2.000000	3.553239	14.000000	0.000000	0.000000
75%	3.000000	4.538933	20.000000	0.000000	0.000000
max	4.000000	10.309005	190.000000	1.000000	1.000000

	workers	fomento	iyd	impuestos	utilidades \
count	39103.000000	39103.000000	39103.000000	39103.000000	3.910300e+04
mean	1.757771	0.076107	0.224893	0.203861	7.132256e-07
std	1.186489	0.265172	0.417517	15.869669	2.037774e-05
min	0.000000	0.000000	0.000000	-180.992528	-2.443698e-04
25%	0.778151	0.000000	0.000000	0.000000	9.050000e-13
50%	1.785330	0.000000	0.000000	0.000007	8.080000e-11
75%	2.661813	0.000000	0.000000	0.000167	1.283147e-09
max	5.845915	1.000000	1.000000	2981.494528	1.806377e-03

	...	mfomento	miyd	mimpuestos	mutilidades \
count	...	39103.000000	39103.000000	39103.000000	3.910300e+04
mean	...	0.076107	0.224893	0.203861	7.132256e-07
std	...	0.210187	0.346859	9.065499	1.031957e-05
min	...	0.000000	0.000000	-60.319504	-1.221692e-04
25%	...	0.000000	0.000000	0.000000	1.907500e-12
50%	...	0.000000	0.000000	0.000024	2.118090e-10
75%	...	0.000000	0.400000	0.001182	1.418978e-08
max	...	1.000000	1.000000	995.749564	6.335315e-04

	mlutilidades	mP1	mP2	mP3	mP4 \
--	--------------	-----	-----	-----	-------

count	39103.000000	39103.000000	39103.000000	39103.000000	39103.000000
mean	0.000245	0.261131	0.180574	0.185843	0.206736
std	0.000010	0.371470	0.276818	0.251676	0.250901
min	0.000122	0.000000	0.000000	0.000000	0.000000
25%	0.000244	0.000000	0.000000	0.000000	0.000000
50%	0.000244	0.000000	0.000000	0.000000	0.200000
75%	0.000244	0.333333	0.250000	0.333333	0.333333
max	0.000878	1.000000	1.000000	1.000000	1.000000

mP5

count	39103.000000
mean	0.165716
std	0.265710
min	0.000000
25%	0.000000
50%	0.000000
75%	0.250000
max	1.000000

[8 rows x 32 columns]

## 1.1 Minimos Cuadrados Ordinarios Agrupados (Pooled OLS)

```
[26]: y=Xc['workers']
X=Xc[['sales', 'age', 'foreign', 'export', 'iyd', 'impuestos',
      'lutilidades', 'P2', 'P3', 'P4', 'P5']]
X=sm.add_constant(X)
model = lmp.PanelOLS(y, X)
mco = model.fit()
print(mco)
```

### PanelOLS Estimation Summary

Dep. Variable:	workers	R-squared:	0.6173
Estimator:	PanelOLS	R-squared (Between):	0.6322
No. Observations:	39103	R-squared (Within):	0.4112
Date:	Fri, May 12 2023	R-squared (Overall):	0.6173
Time:	13:39:01	Log-likelihood	-4.339e+04
Cov. Estimator:	Unadjusted		
		F-statistic:	5732.6
Entities:	24128	P-value	0.0000
Avg Obs:	1.6206	Distribution:	F(11,39091)
Min Obs:	1.0000		
Max Obs:	5.0000	F-statistic (robust):	5732.6
		P-value	0.0000
Time periods:	5	Distribution:	F(11,39091)
Avg Obs:	7820.6		



Min Obs: 6480.0  
Max Obs: 1.021e+04

Parameter Estimates						
	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-0.3634	0.0528	-6.8789	0.0000	-0.4670	-0.2599
sales	0.4533	0.0032	143.75	0.0000	0.4471	0.4595
age	0.0077	0.0003	24.817	0.0000	0.0071	0.0083
foreign	0.0608	0.0144	4.2320	0.0000	0.0326	0.0890
export	0.0893	0.0127	7.0191	0.0000	0.0643	0.1142
iyd	0.2459	0.0094	26.051	0.0000	0.2274	0.2644
impuestos	0.0001	0.0003	0.4447	0.6565	-0.0004	0.0007
lutilidades	-173.73	213.79	-0.8126	0.4164	-592.76	245.30
P2	-1.0991	0.0157	-70.227	0.0000	-1.1298	-1.0684
P3	1.0899	0.0119	91.599	0.0000	1.0666	1.1132
P4	1.0349	0.0115	89.816	0.0000	1.0123	1.0575
P5	0.8318	0.0120	69.281	0.0000	0.8083	0.8553

Se puede observar en el reporte que el modelo es significativo, por lo cual es posible afirmar que las variables permiten realizar una buena estimacion para la variable dependiente, que es el número de trabajadores. Además, es posible afirmar que las variables explicativas explican en un 61.73% la variación del número de trabajadores.

Podemos observar que la mayoría de los factores son significativos a excepción de los impuestos de las firmas (impuestos) y el logaritmo de las utilidades (lutilidades).

Con respecto a las variables significativas, es posible afirmar lo siguiente:

- La cantidad de trabajadores aumenta un 45.33% cuando las ventas aumentan en log(1000) unidades monetarias.
- La cantidad de trabajadores aumenta un 0.77% cuando la edad de la firma aumenta en 1 año.
- La cantidad de trabajadores aumenta en un 6.08% cuando la firma es extranjera.
- La cantidad de trabajadores aumenta en un 8.93% cuando la firma exporta.
- La cantidad de trabajadores aumenta en un 24.59% cuando la firma realiza investigación y desarrollo.

Debido a que el periodo base es 2007, la dummy “P1” no puede coexistir con las demás en el modelo, por lo cual se decidió hacer la estimación solamente con las variables dummies de tiempo “P2”, “P3”, “P4” y “P5”, correspondientes a los años 2009, 2013, 2015 y 2017, respectivamente. En esta estimación, no se pueden interpretar el coeficiente de P2, ya que no tiene sentido que el número de trabajadores disminuya un 109% de 2007 a 2009 cuando el periodo es 2009, pero si se puede interpretar, en este caso, que el número de trabajadores, en el año 2009 fue más del doble que en 2007.

## 1.2 Modelo MCO Agrupado Robusto

### 1.3 First differences

```
[27]: y=Xc['workers']
X=Xc[['sales', 'age', 'foreign','export', 'iyd', 'impuestos',
      'lutilidades','P2','P3','P4','P5']]
model=lm.PooledOLS(y,X)
fd=model.fit(cov_type="robust")
print(fd)
```

#### PooledOLS Estimation Summary

```
=====
Dep. Variable:          workers    R-squared:                0.8801
Estimator:             PooledOLS  R-squared (Between):      0.8688
No. Observations:      39103      R-squared (Within):       0.4100
Date:                  Fri, May 12 2023  R-squared (Overall):     0.8801
Time:                  13:46:55      Log-likelihood            -4.341e+04
Cov. Estimator:        Robust

F-statistic:                2.608e+04
Entities:                  24128      P-value                  0.0000
Avg Obs:                   1.6206     Distribution:             F(11,39092)
Min Obs:                   1.0000
Max Obs:                   5.0000     F-statistic (robust):     2.552e+04
P-value                    0.0000
Time periods:              5      Distribution:             F(11,39092)
Avg Obs:                   7820.6
Min Obs:                   6480.0
Max Obs:                   1.021e+04
```

#### Parameter Estimates

```
=====
Parameter  Std. Err.    T-stat    P-value    Lower CI    Upper CI
-----
sales      0.4519    0.0056    80.078    0.0000    0.4408    0.4629
age        0.0076    0.0003    23.295    0.0000    0.0069    0.0082
foreign    0.0661    0.0157    4.2121    0.0000    0.0353    0.0969
export     0.0915    0.0124    7.3743    0.0000    0.0672    0.1158
iyd        0.2454    0.0088    27.892    0.0000    0.2282    0.2626
impuestos  0.0011    0.0002    4.6173    0.0000    0.0006    0.0015
lutilidades -1613.8    53.560    -30.131    0.0000    -1718.8    -1508.9
P2         -1.0970    0.0257    -42.711    0.0000    -1.1474    -1.0467
P3         1.0843    0.0121    89.502    0.0000    1.0605    1.1080
P4         1.0295    0.0117    87.649    0.0000    1.0064    1.0525
P5         0.8259    0.0123    67.317    0.0000    0.8019    0.8500
=====
```

Se puede observar que cuando se asume que hay heterocedasticidad, disminuyen los errores estándar de los estimadores y, además, ahora todas las variables del modelo seleccionadas anteriormente son

significativas. Adicionalmente, ahora la bondad de ajuste es muy superior al del modelo MCO Agrupado que no asume heterocedasticidad, siendo ahora de un 88.01%. Se detecta sesgo en el estimador del logaritmo de las utilidades, debido a que ahora el valor es muy inferior al valor cuando no se trabaja con heterocedasticidad. No se detecta sesgo en los demás estimadores debido a que los valores de estos son prácticamente idénticos a los del modelo cuando los errores estándar no son robustos.

### 1.3.1 Pregunta 3

## 1.4 Estimación por Efectos Fijos

Ahora se pasa a la aplicación del modelo con estimación de efectos fijos, el cual tal y como su nombre lo indica se caracteriza por hacer que los efectos de las variables se mantengan fijos a través del tiempo, lo que a diferencia del modelo con estimación MCO Agrupado, permite una comparación individual para cada uno de los datos, es decir, una comparación de individuos contra sí mismos. Es importante notar que este modelo implica menos suposiciones sobre el comportamiento de los residuos. Para la ejecución de este modelo se consideraron las mismas variables que para el modelo MCO Agrupado.

```
[32]: X=Xc[['sales', 'age', 'foreign','export', 'iyd', 'impuestos',
        ↪'utilidades', 'P2', 'P3', 'P4', 'P5']]
X=sm.add_constant(X)
model=lm.PanelOLS(y,X, entity_effects=True)
fe=model.fit(cov_type="robust")
print(fe)
```

### PanelOLS Estimation Summary

```
=====
Dep. Variable:          workers    R-squared:                0.5559
Estimator:              PanelOLS   R-squared (Between):      0.4193
No. Observations:       39103      R-squared (Within):       0.5559
Date:                   Fri, May 12 2023  R-squared (Overall):     0.4232
Time:                   14:21:27    Log-likelihood            -3478.7
Cov. Estimator:         Robust

                               F-statistic:          1702.8
Entities:                24128    P-value                  0.0000
Avg Obs:                  1.6206   Distribution:             F(11,14964)
Min Obs:                  1.0000
Max Obs:                  5.0000   F-statistic (robust):     813.56
                               P-value                0.0000
Time periods:              5      Distribution:             F(11,14964)
Avg Obs:                   7820.6
Min Obs:                   6480.0
Max Obs:                   1.021e+04
```

### Parameter Estimates

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

const	0.8043	0.0649	12.396	0.0000	0.6771	0.9314
sales	0.1207	0.0136	8.9076	0.0000	0.0942	0.1473
age	-0.0011	0.0008	-1.4099	0.1586	-0.0026	0.0004
foreign	0.0889	0.0357	2.4924	0.0127	0.0190	0.1589
export	0.0295	0.0240	1.2259	0.2203	-0.0177	0.0766
iyd	0.0461	0.0124	3.7138	0.0002	0.0218	0.0705
impuestos	-0.0001	0.0001	-0.9533	0.3405	-0.0004	0.0001
lutilidades	87.821	206.89	0.4245	0.6712	-317.71	493.35
P2	-0.3468	0.0404	-8.5838	0.0000	-0.4260	-0.2676
P3	1.0278	0.0152	67.839	0.0000	0.9981	1.0575
P4	1.0002	0.0164	60.993	0.0000	0.9680	1.0323
P5	0.9702	0.0189	51.457	0.0000	0.9333	1.0072
=====						

F-test for Poolability: 4.1562

P-value: 0.0000

Distribution: F(24127,14964)

Included effects: Entity

Al analizar los resultados generales del modelo con estimación de efectos fijos, tenemos en primer lugar que el R cuadrado obtenido de la aplicación del modelo es inferior, lo que quiere decir que las variables logran explicar en menor medida a la variable dependiente, sin embargo, en función del valor p observado de la regresión se observa que ésta es significativa.

Ahora las variables de edad de la firma, si la firma exporta, impuestos y el logaritmo de utilidades son no significativas.

En este caso, asumiendo que hay correlación entre la heterogeneidad no observada y las variables explicativas, esta es la interpretación de los estimadores de las variables explicativas que son significativas: - La cantidad de trabajadores aumenta un 12.07% cuando las ventas aumentan en log(1000) unidades monetarias. - La cantidad de trabajadores aumenta en un 8.89% cuando la firma es extranjera. - La cantidad de trabajadores aumenta en un 4.61% cuando la firma realiza investigación y desarrollo.

Con respecto a las variables binarias de tiempo:

- La cantidad de trabajadores disminuye en un 34.68% cuando se pasa de 2007 a 2009.
- La cantidad de trabajadores aumenta en un 102.78% cuando se pasa de 2007 a 2013.
- La cantidad de trabajadores aumenta en un 100.02% cuando se pasa de 2007 a 2015.
- La cantidad de trabajadores aumenta en un 97.02% cuando se pasa de 2007 a 2017.

#### 1.4.1 Pregunta 4

### 1.5 Estimación por Efectos Aleatorios

El modelo con estimación de efectos aleatorios tiene la misma especificación que el de efectos fijos, pero con la diferencia de que los valores ya no son fijos para cada uno de los datos estudiados, sino que se comportan de manera aleatoria con un valor medio y una varianza distinta de cero. Este modelo resulta ser más eficiente pero menos consistente que el de efectos fijos.

```
[34]: model=lmp.RandomEffects(y,X)
re=model.fit(cov_type="robust")
print(re)
```

```

RandomEffects Estimation Summary
=====
Dep. Variable:          workers    R-squared:                0.5343
Estimator:             RandomEffects  R-squared (Between):      0.6195
No. Observations:      39103    R-squared (Within):       0.4799
Date:                  Fri, May 12 2023  R-squared (Overall):      0.6036
Time:                  14:26:22    Log-likelihood            -2.421e+04
Cov. Estimator:        Robust
                               F-statistic:                4077.2
Entities:              24128    P-value                  0.0000
Avg Obs:               1.6206    Distribution:             F(11,39091)
Min Obs:               1.0000
Max Obs:               5.0000    F-statistic (robust):     4431.8
                               P-value                  0.0000
Time periods:          5    Distribution:             F(11,39091)
Avg Obs:               7820.6
Min Obs:               6480.0
Max Obs:               1.021e+04

```

```

Parameter Estimates
=====

```

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	-0.1202	0.0510	-2.3589	0.0183	-0.2201	-0.0203
sales	0.3677	0.0062	59.156	0.0000	0.3555	0.3798
age	0.0056	0.0003	16.192	0.0000	0.0049	0.0062
foreign	0.1446	0.0164	8.8224	0.0000	0.1124	0.1767
export	0.1334	0.0124	10.774	0.0000	0.1091	0.1576
iyd	0.1743	0.0076	23.005	0.0000	0.1595	0.1892
impuestos	-2.75e-05	0.0002	-0.1101	0.9123	-0.0005	0.0005
lutilidades	-155.09	200.33	-0.7742	0.4388	-547.74	237.55
P2	-0.9032	0.0247	-36.635	0.0000	-0.9515	-0.8549
P3	1.0897	0.0095	114.58	0.0000	1.0711	1.1084
P4	1.0348	0.0094	110.09	0.0000	1.0164	1.0532
P5	0.9163	0.0104	88.058	0.0000	0.8959	0.9367

```

=====

```

El valor del R cuadrado disminuyo con respecto al modelo anterior, siendo ahora de 0,5197, y el valor p de la prueba F sigue indicando que el modelo es significativo, por lo que todas las variables en conjunto explican estadísticamente con un 95% de confianza la variación en el número de trabajadores de la firma.

En este caso, solo las variables de impuestos y logaritmo de utilidades resultaron no significativas. Con respecto a las variables significativas, ahora es posible afirmar lo siguiente: - La cantidad de trabajadores aumenta un 36.77% cuando las ventas aumentan en  $\log(1000)$  unidades monetarias.

- La cantidad de trabajadores aumenta un 0.56% cuando la edad de la firma aumenta en 1 año. - La cantidad de trabajadores aumenta en un 14.46% cuando la firma es extranjera. - La cantidad de trabajadores aumenta en un 13.34% cuando la firma exporta. - La cantidad de trabajadores aumenta en un 17.43% cuando la firma realiza investigación y desarrollo.

Con respecto a las variables binarias de tiempo: - La cantidad de trabajadores disminuye en un 90.32% cuando se pasa de 2007 a 2009. - La cantidad de trabajadores aumenta en un 108.97% cuando se pasa de 2007 a 2013. - La cantidad de trabajadores aumenta en un 103.48% cuando se pasa de 2007 a 2015. - La cantidad de trabajadores aumenta en un 91.63% cuando se pasa de 2007 a 2017.

```
[35]: re.variance_decomposition #Que tanta variacion depende de los individuos entre_
      ↪ si
```

```
[35]: Effects                0.314335
      Residual                0.182793
      Percent due to Effects  0.632302
      Name: Variance Decomposition, dtype: float64
```

### 1.5.1 Pregunta 5

## 1.6 Comparación de modelos

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

Model Comparison			
	FE	RE	Pooled
Dep. Variable	workers	workers	workers
Estimator	PanelOLS	RandomEffects	PanelOLS
No. Observations	39103	39103	39103
Cov. Est.	Robust	Robust	Unadjusted
R-squared	0.1734	0.0309	0.2580
R-Squared (Within)	0.1734	-0.1010	-0.5335
R-Squared (Between)	-0.3118	0.2145	0.3042
R-Squared (Overall)	-0.1992	0.1904	0.2580
F-statistic	448.51	177.94	1941.5
P-value (F-stat)	0.0000	0.0000	0.0000
const	2.1671 (29.955)	1.1601 (13.837)	0.8062 (11.039)
sales	-0.1688 (-27.799)	0.0777 (26.558)	0.2391 (74.676)
age	0.0127 (9.0815)	0.0217 (33.373)	0.0204 (48.610)
foreign	0.1054 (2.1379)	0.5024 (21.954)	0.4109 (20.765)
export	0.0588	0.3885	0.3359

	(1.7464)	(21.527)	(19.085)
iyd	-0.1733	0.1327	0.3050
	(-9.6421)	(11.932)	(24.212)
impuestos	-0.0001	0.0002	0.0005
	(-0.4023)	(0.5992)	(1.1969)
lutilidades	98.209	-1054.6	-1446.7
	(0.3585)	(-3.0828)	(-4.8624)
=====			
Effects	Entity		
-----			

T-stats reported in parentheses

### 1.6.1 Test de Hausman

```
[43]: 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

[44]: 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 = 843.2784278644602, df = 11, p-value = 9.603787577909946e-174

**El test de Hausman indica lo siguiente:**

- Hipótesis nula (H0): Los efectos fijos son no correlacionados con los regresores.
- Hipótesis alternativa (HA): Los efectos fijos están correlacionados con los regresores.

Como el valor-p del test de Hausman es inferior al nivel de significancia del 5%, es posible afirmar que existe suficiente evidencia estadística que indica que la estimación por efectos fijos es superior a la estimación por efectos aleatorios.

**Comparación:** La estimación por MCO agrupado tiene una gran desventaja, que es que asume que, al agrupar las observaciones, los coeficientes son iguales para todos los individuos, es decir, no hay distinción entre ellos, uno es tan bueno como el otro y además supone exogeneidad, lo que significa que las perturbaciones no están correlacionadas con las variables explicativas. Además, cabe señalar que, al comparar los modelos, se observa sesgo en los estimadores de MCO agrupado, ya que, en muchos casos, los estimadores de Efectos aleatorios y de efectos fijos son inferiores a los de MCO agrupado, por lo cual estos modelos, en este caso, son superiores. Además, se confirma

que la estimación por efectos fijos es superior a la estimación de MCO Agrupado, debido a que la razón F de esta estimación es significativa con un 95% de confianza. Como se indicó anteriormente la estimación de efectos fijos también es superior a la de efectos aleatorios (lo que indica que las heterogeneidades no observadas están relacionadas con las variables explicativas), por lo cual, es posible afirmar que es mejor utilizar efectos fijos para estimar el número de trabajadores si solo se tienen estos tres modelos de MCO Agrupado, efectos aleatorios y efectos fijos.

## 1.6.2 Pregunta 6

## 1.7 Efectos aleatorios correlacionados

```
[45]: y = Xc['workers']
X=Xc[['sales', 'age', 'foreign','export', 'iyd', 'impuestos',
      ↪'lutilidades','msales','mage','mforeign','mexport',
      'miyd','mimpuestos','mlutilidades', 'P2','P3','P4','P5']]
X=sm.add_constant(X)
model=lmpr.RandomEffects(y,X)
cre=model.fit(cov_type="robust")
print(cre)
```

### RandomEffects Estimation Summary

```
=====
Dep. Variable:          workers    R-squared:                0.5621
Estimator:             RandomEffects    R-squared (Between):      0.6251
No. Observations:      39103    R-squared (Within):       0.5421
Date:                  Fri, May 12 2023    R-squared (Overall):      0.6229
Time:                  15:29:59    Log-likelihood            -2.301e+04
Cov. Estimator:        Robust

                                F-statistic:                2786.7
Entities:              24128    P-value                  0.0000
Avg Obs:               1.6206    Distribution:            F(18,39084)
Min Obs:               1.0000
Max Obs:               5.0000    F-statistic (robust):    3614.5
                                P-value                  0.0000
Time periods:          5    Distribution:            F(18,39084)
Avg Obs:               7820.6
Min Obs:               6480.0
Max Obs:               1.021e+04
```

### Parameter Estimates

```
=====
Parameter  Std. Err.    T-stat    P-value    Lower CI    Upper CI
-----
const      -0.3054    0.1987   -1.5373    0.1242   -0.6948    0.0840
sales       0.2162    0.0081   26.814    0.0000    0.2004    0.2320
age        -0.0017    0.0005  -3.3714    0.0007   -0.0027   -0.0007
foreign     0.0842    0.0255    3.3076    0.0009    0.0343    0.1341
export      0.0217    0.0177    1.2285    0.2193   -0.0129    0.0563
iyd         0.0322    0.0094    3.4202    0.0006    0.0138    0.0507
```



impuestos	-0.0001	0.0001	-1.1338	0.2569	-0.0003	8.974e-05
lutilidades	112.66	153.71	0.7329	0.4636	-188.62	413.93
msales	0.1947	0.0058	33.683	0.0000	0.1834	0.2061
mage	0.0106	0.0006	16.658	0.0000	0.0093	0.0118
mforeign	0.0177	0.0330	0.5354	0.5924	-0.0470	0.0823
mexport	0.1293	0.0245	5.2758	0.0000	0.0813	0.1773
miyd	0.3072	0.0145	21.134	0.0000	0.2787	0.3357
mimpuestos	0.0009	0.0005	1.8766	0.0606	-4.206e-05	0.0019
mlutilidades	-544.57	829.81	-0.6563	0.5117	-2171.0	1081.9
P2	-0.6826	0.0241	-28.367	0.0000	-0.7298	-0.6355
P3	1.0380	0.0090	114.99	0.0000	1.0203	1.0557
P4	1.0154	0.0090	112.29	0.0000	0.9977	1.0331
P5	0.9138	0.0101	90.457	0.0000	0.8940	0.9336

Utilizar el modelo de efectos aleatorios correlacionados permite considerar la heterogeneidad de cada individuo encuestado, por lo que da más información considerando la caracterización de cada individuo con variables como sales, age, foreign. En este caso se observa que 3 variables resultaron no significativas: “export”, “impuestos” y “lutilidades” y las 3 variables de media: “mforeign”, “mimpuestos”, “mlutilidades”. Con respecto a las variables significativas:

- La cantidad de trabajadores aumenta un 21.62% cuando las ventas aumentan en log(1000) unidades monetarias.
- La cantidad de trabajadores disminuye en un 0.17% cuando la edad de la firma aumenta en 1 año.
- La cantidad de trabajadores aumenta en un 8.42% cuando la firma es extranjera.
- La cantidad de trabajadores aumenta en un 3.22% cuando la firma realiza investigación y desarrollo.

Con respecto a las dummies de tiempo: - La cantidad de trabajadores disminuye en un 68.26% cuando se pasa de 2007 a 2009. - La cantidad de trabajadores aumenta en un 103.8% cuando se pasa de 2007 a 2013. - La cantidad de trabajadores aumenta en un 101.54% cuando se pasa de 2007 a 2015. - La cantidad de trabajadores aumenta en un 91.38% cuando se pasa de 2007 a 2017.

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.

### 1.7.1 Pregunta 7: EAC para predecir la distribución del componente no observado

```
[41]: Xpred = X
Xpred['sales']=0
Xpred['age']=0
Xpred['foreign']=0
Xpred['export']=0
Xpred['iyd']=0
Xpred['impuestos']=0
Xpred['lutilidades']=0
Xpred['P2']=0
```

```

Xpred['P3']=0
Xpred['P4']=0
Xpred['P5']=0
yhat = cre.predict(Xpred)

sns.histplot(data=y, color="skyblue", label="workers(observed)", kde=True)
sns.histplot(data=yhat, color="red", label="unobserved heterogeneity", kde=True)

plt.legend()
plt.show()

#Comparar la ui con el Y estimado, ver correlacion en cada periodo. Si hay
→heterogeneidad no observada,
# la diferencia entre individuos en el tiempo, la variacion del numero de
→trabajadores se explica en gran o poca medida por la
# heterogeneidad no observada, hacer anova

```



Se puede visualizar que la heterogeneidad no observada se distribuye aproximadamente normal entre 0 y 2 log trabajadores.

Adicionalmete, es posible afirmar que debido a que el histograma de la heterogeneidad no observada

no esta tapando casi nada totalidad del histograma de lo observado, que la variacion de trabajadores en el tiempo es explicada en poca medida por factores no medidos.

### 1.7.2 Pregunta 8: Elección de un modelo

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

Model Comparison			
	FE	RE	CRE
Dep. Variable	workers	workers	workers
Estimator	PanelOLS	RandomEffects	RandomEffects
No. Observations	39103	39103	39103
Cov. Est.	Robust	Robust	Robust
R-squared	0.5559	0.5343	0.5621
R-Squared (Within)	0.5559	0.4799	0.5421
R-Squared (Between)	0.4193	0.6195	0.6251
R-Squared (Overall)	0.4232	0.6036	0.6229
F-statistic	1702.8	4077.2	2786.7
P-value (F-stat)	0.0000	0.0000	0.0000
const	0.8043 (12.396)	-0.1202 (-2.3589)	-0.3054 (-1.5373)
sales	0.1207 (8.9076)	0.3677 (59.156)	0.2162 (26.814)
age	-0.0011 (-1.4099)	0.0056 (16.192)	-0.0017 (-3.3714)
foreign	0.0889 (2.4924)	0.1446 (8.8224)	0.0842 (3.3076)
export	0.0295 (1.2259)	0.1334 (10.774)	0.0217 (1.2285)
iyd	0.0461 (3.7138)	0.1743 (23.005)	0.0322 (3.4202)
impuestos	-0.0001 (-0.9533)	-2.75e-05 (-0.1101)	-0.0001 (-1.1338)
lutilidades	87.821 (0.4245)	-155.09 (-0.7742)	112.66 (0.7329)
P2	-0.3468 (-8.5838)	-0.9032 (-36.635)	-0.6826 (-28.367)
P3	1.0278 (67.839)	1.0897 (114.58)	1.0380 (114.99)
P4	1.0002 (60.993)	1.0348 (110.09)	1.0154 (112.29)
P5	0.9702 (51.457)	0.9163 (88.058)	0.9138 (90.457)
msales			0.1947 (33.683)

mage	0.0106 (16.658)
mforeign	0.0177 (0.5354)
mexport	0.1293 (5.2758)
miyd	0.3072 (21.134)
mimpuestos	0.0009 (1.8766)
mlutilidades	-544.57 (-0.6563)
=====	
Effects	Entity
-----	

T-stats reported in parentheses

Tomando en consideración los resultados obtenidos para cada uno de los modelos, el modelo por el cual hay una preferencia es el modelo con estimación de Efectos Aleatorios Correlacionados, esto debido a que comparte una gran cantidad de características con el modelo de efectos fijos, pero además, integra las medias de cada una de las variables (con la excepción de las variables que son fijas en el tiempo), lo cual brinda una mayor cantidad de información a la hora de tener que estudiar la significancia de las variables. Con respecto a las variables explicativas, las variables que explican al modelo son: Las ventas, la edad de la firma, si la firma es extranjera, y si la firma realiza investigación y desarrollo. Las variables que tienen un efecto positivo en el aumento del número de trabajadores en la firma son: las ventas, si la firma es extranjera y si la firma realiza investigación y desarrollo; y la que tiene un efecto negativo es la edad de la firma.

### 1.7.3 Pregunta 9: Utilizar fomento como instrumento

```
[63]: from linearmodels.iv import compare


# Seleccionar las variables explicativas de tu conjunto de datos
exog_variablesOLS = ['sales', 'age',
                    ↪ 'foreign', 'export', 'impuestos', 'lutilidades', 'iyd']
exogOLS = pd.DataFrame(enia[exog_variablesOLS])
exog_variablesIV = ['sales', 'age',
                    ↪ 'foreign', 'export', 'impuestos', 'lutilidades']
exogIV = pd.DataFrame(enia[exog_variablesIV])

res_ols = IV2SLS(dependent=enia['workers'], exog=exogOLS, endog=None,
                    ↪ instruments=None).fit(cov_type="unadjusted")
res_IV = IV2SLS(dependent=enia['workers'], exog=exogIV, endog=enia['iyd'],
                    ↪ instruments=enia['fomento']).fit(
                    cov_type="unadjusted")

print(compare({"OLS": res_ols, "2SLS": res_IV}))
```

Model Comparison		
	OLS	2SLS
Dep. Variable	workers	workers
Estimator	OLS	IV-2SLS
No. Observations	39103	39103
Cov. Est.	unadjusted	unadjusted
R-squared	0.7670	0.7655
Adj. R-squared	0.7670	0.7655
F-statistic	1.287e+05	1.273e+05
P-value (F-stat)	0.0000	0.0000
sales	0.2409 (75.264)	0.2430 (73.089)
age	0.0207 (49.501)	0.0211 (47.525)
foreign	0.4020 (20.299)	0.4081 (20.369)
export	0.3330 (18.894)	0.3714 (15.428)
impuestos	-0.0017 (-5.1305)	-0.0017 (-5.1440)
lutilidades	1783.1 (32.895)	1895.3 (26.211)
iyd	0.3057 (24.232)	0.1073 (1.2591)
Instruments		fomento

T-stats reported in parentheses

Cuando se utiliza la variable “fomento” como instrumento, se llega a la conclusión de que el hecho de poseer investigación y desarrollo no tiene un impacto significativo en el número de trabajadores, cosa muy distinta a la estimación MCO, ya que esta establece que cuando la firma tiene investigación y desarrollo, hay un aumento significativo de  trabajadores del 30%, con un nivel de significancia del 5% en ambos casos.