Tarea 1 Final Valdebenito

April 25, 2023

Tarea_1_MAA_Diego_Valdebenito

Tarea 1

Instrucciones

Los resultados de los ejericicios propuestos se deben entregar como un notebook por correo electronico a *juancaros@udec.cl* a mas tardar el dia 14/04/23 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 **charls2.csv**.

Las variables tienen la siguiente descripcion:

- retin: 1 si planea planea retirarse
- retage: cuando planea retirarse, medido en años desde la fecha de encuesta (0 implica retirado/a o no planea retirarse)
- 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 diario
- hsize: tamano del hogar
- female: 1 si es mujer, 0 si es hombre
- intmonth: mes en que fue encuestado/a (1-12)
- married: si esta casado/a (binario)
- retired: 1 si esta retirado/a (binario)
- schadj: años de escolaridad
- urban: zona urbana (binario)
- wealth: riqueza neta (miles RMB)
- age: edad al entrar a la encuesta

Preguntas:

1. Cargar la base de datos *charls2.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 de probabilidad lineal (*MCO*) que permita explicar la probabilidad de que una persona que aun trabaja quiera retirarse (*retin*). Seleccione las variables independientes a incluir en el modelo final e interprete su significado.

3. Ejecute un modelo *probit* para responder a la pregunta 2. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.
4. Ejecute un modelo *logit* para responder a la pregunta 2. 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 Poisson para explicar cuando planea retirarse las personas que planean hacerlo. Seleccione las variables independientes a incluir en el modelo final e interprete su significado.
7. Determine sobre dispersion y posible valor optimo de alpha para un modelo Binomial Negativa.
8. Usando la informacion anterior, ejecute un modelo Binomial Negativa para explicar en cuanto tiempo planea retirarse. Seleccione las variables independientes a incluir en el modelo final e interprete su significado.
9. Comente los resultados obtenidos en 6, 7 y 8. ¿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?

Carga de librerias:

Se cargan las librerias a utilizar en la tarea

[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 sklearn
import scipy
from scipy.stats import nbinom
import seaborn as sns

%matplotlib inline
```

[8]:
```
# higher ed data
charls = pd.read_csv('C:/Users/equipo/Desktop/Tarea_1_Laboratorio/data/charls2.
 ↪csv')
charls.dropna(inplace=True)
# Eliminar filas donde retin y retired sean igual a 1 simultaneamente.
charls = charls[((charls['retin'] != 1) | (charls['retired'] != 1))]
charls = charls[~charls['drinkly'].str.contains('.m|.d|.r')]
charls.drinkly.unique()
charls['drinkly'] = charls['drinkly'].astype(int)
```

0.0.1 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%.

0.0.2 Eliminación de filas

Se eliminan las filas de los casos en los cuales tanto se esté en el caso de planear retirarse y estar retirado, ya que no tiene coherencia tener la intención de retirarse si ya no se está trabajando. Además, se eliminaron las filas donde la variable "drinkly" sea distinta de 0 o 1, o que sea texto, ya que es necesario que esta variable sea binaria para poder incorporarla al modelo, y luego se convirtió de objeto a variable entera.

[9]:
```
charls.reset_index(drop=True, inplace=True)
charls.head(10)
```

[9]:

	age	cesd	child	drinkly	female	hrsusu	hsize	intmonth	married	\
0	46	6.0	2	0	1	0.000000	4	7	1	
1	48	0.0	2	1	0	4.143135	4	7	1	
2	56	6.0	1	0	1	0.000000	6	8	1	
3	59	6.0	1	1	0	0.000000	6	8	1	
4	47	4.0	1	1	0	3.806663	3	8	1	
5	46	4.0	1	0	1	4.025352	3	8	1	
6	66	11.0	3	0	1	2.995732	3	8	1	
7	67	2.0	3	1	0	3.178054	3	8	1	
8	80	17.0	7	0	1	0.000000	5	7	1	
9	45	4.0	2	1	0	3.044522	5	7	1	

	retage	retin	retired	schadj	urban	wealth
0	24	1	0	0	0	-5800.0
1	22	1	0	4	0	-5800.0
2	0	0	0	0	0	350.0
3	0	0	0	0	0	350.0
4	11	1	0	4	0	-8100.0
5	24	1	0	8	0	-8100.0
6	7	1	0	0	0	200.0
7	3	1	0	8	0	200.0
8	0	0	1	0	0	0.0
9	0	0	0	8	0	5000.0

[161]:
```
print(charls.dtypes)

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

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

```
age           int64
cesd        float64
child         int64
drinkly       int32
female        int64
hrsusu      float64
hsize         int64
intmonth      int64
married       int64
retage        int64
retin         int64
retired       int64
schadj        int64
urban         int64
wealth      float64
dtype: object
Variables binarias en charls:
['drinkly', 'female', 'married', 'retin', 'retired', 'urban']
```

0.0.3 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 "age", "child", "drinkly", "female", "hsize", "intmont", "married", "retage", "retin", "retired", "schadj" y urban son enteras, de las cuales "drinkly", "female", "married", "retin", "retired", "urban"; y las variables continuas son: "cesd", "hrsusu", "wealth".

Las variables que se consideran a priori como importantes son retin, para saber el porcentaje de personas que desean retirarse, que es la variable de estudio, cesd, porque popularmente se sabe que la salud mental es determinante para que una persona esté planeando dejar de trabajar, hrsusu, porque las horas promedio trabajo diario pueden ser importantes para tomar la decisión de retirarse o no retirarse y la edad (age), ya que a priori se piensa que a mayor edad mayor será la propensión a retirarse. Además, por interés en el análisis, se considerará la variable wealth, para ver si la riqueza está relacionada con el hecho de retirarse.

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

0.0.4 Retin (Si desea Retirarse)

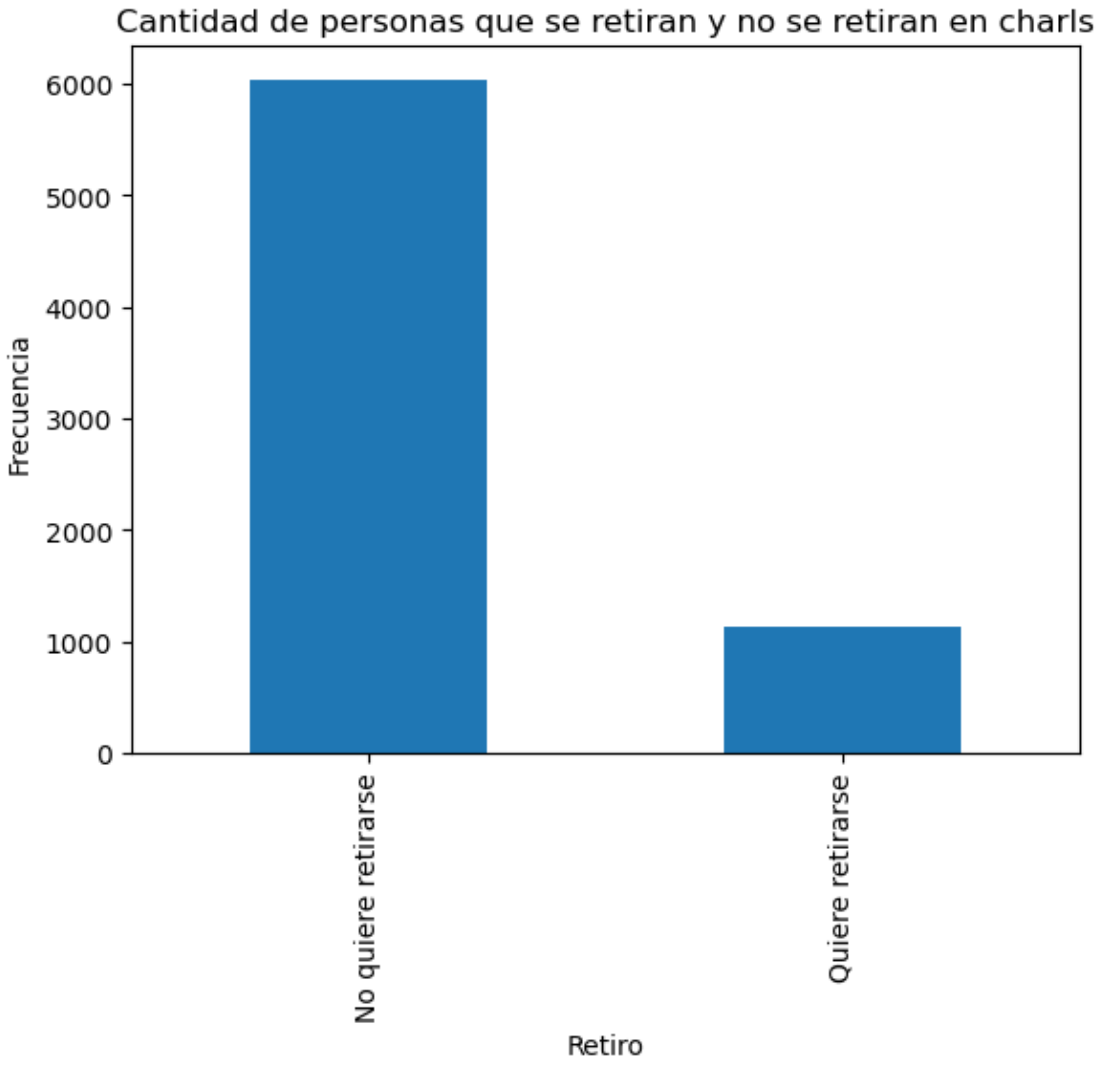
[146]:
```
# Calcular los porcentajes de "retin"
retin_counts = charls['retin'].value_counts(normalize=True)
porcentaje_no_retin = retin_counts[0] * 100
porcentaje_retin = retin_counts[1] * 100

# Imprimir los resultados
```

```
print("El porcentaje de personas que NO planean retirarse es de: {:.2f}%".
  ↪format(porcentaje_no_retin))
print("El porcentaje de personas que planean retirarse es de: {:.2f}%".
  ↪format(porcentaje_retin))
```

```
El porcentaje de personas que NO planean retirarse es de: 84.29%
El porcentaje de personas que planean retirarse es de: 15.71%
```

[162]:
```
# Filtrar las filas donde "retin" toma el valor de 1 o 0
retin_data = charls[charls['retin'].isin([0, 1])]
retin_data['retin_label'] = retin_data['retin'].map({1: 'Quiere retirarse', 0:␣
  ↪'No quiere retirarse'})
retin_data['retin_label'].value_counts().plot(kind='bar')
plt.xlabel('Retiro')
plt.ylabel('Frecuencia')
plt.title('Cantidad de personas que se retiran y no se retiran en charls')
plt.show()
```



Se puede visualizar que la mayoría de las personas encuestadas no está planeando retirarse, por lo cual es más probable que una persona no esté planeando retirarse.

0.0.5 Cesd (Puntaje de salud mental)

[148]:
```
# Filtrar los datos por "retin" igual a 0 y 1
retin_0_data = charls[charls['retin'] == 0]
retin_1_data = charls[charls['retin'] == 1]

# Realizar la estadística descriptiva para "cesd" en caso de no querer retirarse
hrsusu_retin_0_stats = retin_0_data['cesd'].describe()

# Realizar la estadística descriptiva para "cesd" en caso de querer retirarse
hrsusu_retin_1_stats = retin_1_data['cesd'].describe()

# Imprimir los resultados
print("Estadística Descriptiva para 'cesd' cuando NO quiere retirarse:")
print(hrsusu_retin_0_stats)

print("Estadística Descriptiva para 'cesd' cuando quiere retirarse:")
print(hrsusu_retin_1_stats)
```

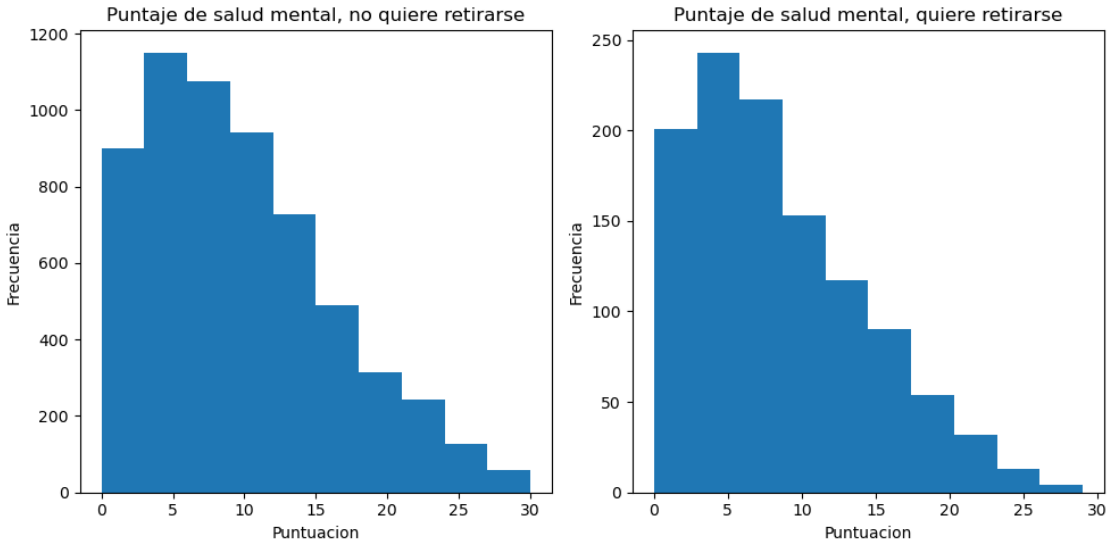
```
Estadística Descriptiva para 'cesd' cuando NO quiere retirarse:
count    6031.000000
mean        9.194661
std         6.516893
min         0.000000
25%         4.000000
50%         8.000000
75%        13.000000
max        30.000000
Name: cesd, dtype: float64
Estadística Descriptiva para 'cesd' cuando quiere retirarse:
count    1124.000000
mean        8.212633
std         6.097763
min         0.000000
25%         3.000000
50%         7.000000
75%        12.000000
max        29.000000
Name: cesd, dtype: float64
```

[149]:
```
# Filtrar las filas donde "retin" toma el valor de 0 o 1
retin_0_data = charls[charls['retin'] == 0]
```

```
retin_1_data = charls[charls['retin'] == 1]
fig, axs = plt.subplots(1, 2, figsize=(10, 5))

# Graficar el histograma de la variable "cesd" cuando "retin" es 0
axs[0].hist(retin_0_data['cesd'], bins=10)
axs[0].set_title('Puntaje de salud mental, no quiere retirarse')
axs[0].set_xlabel('Puntuacion')
axs[0].set_ylabel('Frecuencia')

# Graficar el histograma de la variable "cesd" cuando "retin" es 1
axs[1].hist(retin_1_data['cesd'], bins=10)
axs[1].set_title('Puntaje de salud mental, quiere retirarse')
axs[1].set_xlabel('Puntuacion')
axs[1].set_ylabel('Frecuencia')
plt.tight_layout()
plt.show()
```



Se puede visualizar que en promedio las personas que no quieren retirarse tienen mayor puntuación de salud mental, es decir tienen más patologías que las personas que sí se quieren retirar. Por ello es posible afirmar a priori que el puntaje de salud mental, puede ser significativo para que la persona decida si retirarse o no.

0.0.6 Hrsusu (Horas medias de trabajo diario)

[150]:
```
# Filtrar los datos por "retin" igual a 0 y 1
retin_0_data = charls[charls['retin'] == 0]
retin_1_data = charls[charls['retin'] == 1]
```

```
# Realizar la estadística descriptiva para "hrsusu" en caso de no querer␣
 ↪retirarse
hrsusu_retin_0_stats = retin_0_data['hrsusu'].describe()

# Realizar la estadística descriptiva para "hrsusu" en caso de querer retirarse
hrsusu_retin_1_stats = retin_1_data['hrsusu'].describe()

# Imprimir los resultados
print("Estadística Descriptiva para 'hrsusu' cuando NO quiere retirarse:")
print(hrsusu_retin_0_stats)

print("Estadística Descriptiva para 'hrsusu' cuando quiere retirarse:")
print(hrsusu_retin_1_stats)
```

```
Estadística Descriptiva para 'hrsusu' cuando NO quiere retirarse:
count    6031.000000
mean        2.388028
std         1.841145
min         0.000000
25%         0.000000
50%         3.332205
75%         4.025352
max         4.890349
Name: hrsusu, dtype: float64
Estadística Descriptiva para 'hrsusu' cuando quiere retirarse:
count    1124.000000
mean        3.285881
std         1.255859
min         0.000000
25%         3.044522
50%         3.737670
75%         4.025352
max         4.941642
Name: hrsusu, dtype: float64
```

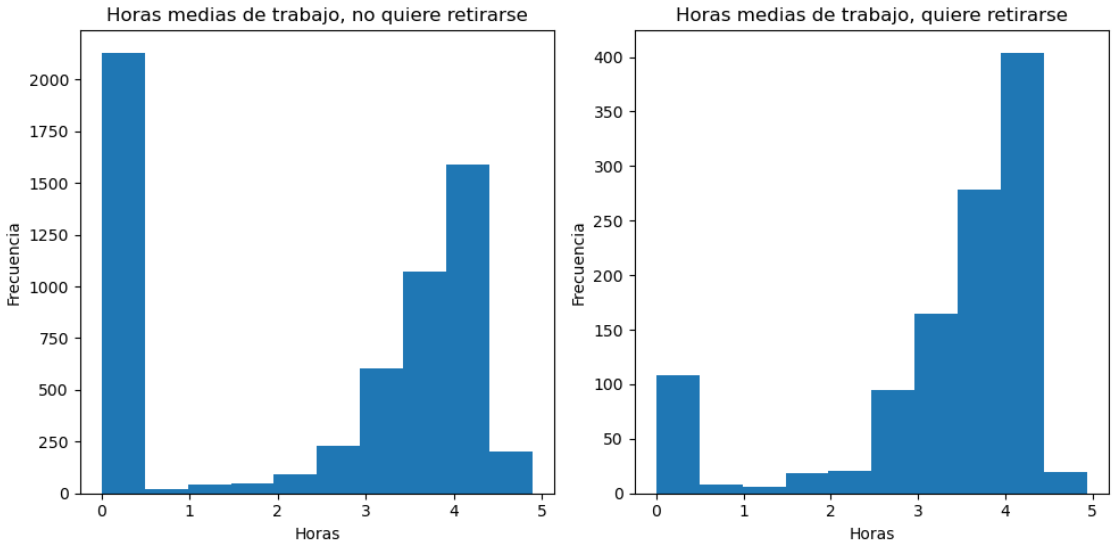
[151]:
```
# Filtrar las filas donde "retin" toma el valor de 0 o 1
retin_0_data = charls[charls['retin'] == 0]
retin_1_data = charls[charls['retin'] == 1]
fig, axs = plt.subplots(1, 2, figsize=(10, 5))

# Graficar el histograma de la variable "hrsusu" cuando no quiere retirarse
axs[0].hist(retin_0_data['hrsusu'], bins=10)
axs[0].set_title('Horas medias de trabajo, no quiere retirarse')
axs[0].set_xlabel('Horas')
axs[0].set_ylabel('Frecuencia')

# Graficar el histograma de la variable "hrsusu" cuando quiere retirarse
```

```
axs[1].hist(retin_1_data['hrsusu'], bins=10)
axs[1].set_title('Horas medias de trabajo, quiere retirarse')
axs[1].set_xlabel('Horas')
axs[1].set_ylabel('Frecuencia')
plt.tight_layout()


plt.show()
```



Se puede visualizar que la mayoría de las personas que no quieren retirarse tienen pocas horas en promedio de trabajo diarias, mientras que las que planean retirarse en su mayoría tienen más horas de trabajo diario medias, en concreto, en promedio las personas que no planean retirarse trabajan 0,9 horas medias diarias menos que las personas que planean retirarse, por lo cual es posible afirmar a priori que las horas de trabajo diario explican el deseo de retirarse o no retirarse por parte del individuo.

0.0.7 Age (edad)

[152]:
```
# Filtrar los datos por "retin" igual a 0 y 1
retin_0_data = charls[charls['retin'] == 0]
retin_1_data = charls[charls['retin'] == 1]

# Realizar la estadística descriptiva para "age" en caso de no querer retirarse
age_retin_0_stats = retin_0_data['age'].describe()

# Realizar la estadística descriptiva para "age" en caso de querer retirarse
age_retin_1_stats = retin_1_data['age'].describe()

# Imprimir los resultados
```

```
print("Estadística Descriptiva para 'age' cuando NO quiere retirarse:")
print(age_retin_0_stats)

print("Estadística Descriptiva para 'age' cuando quiere retirarse:")
print(age_retin_1_stats)
```

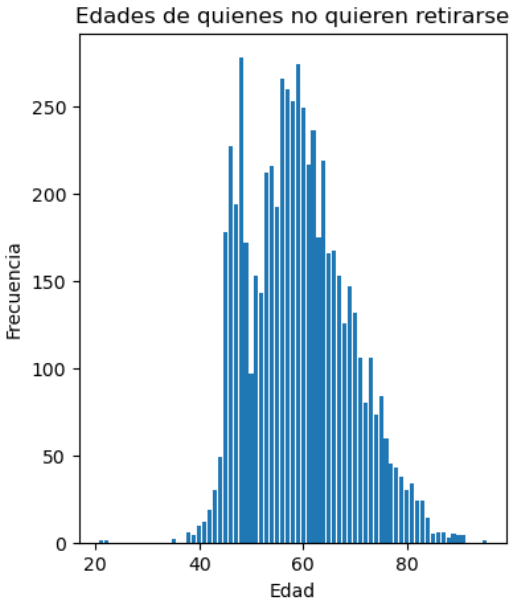
```
Estadística Descriptiva para 'age' cuando NO quiere retirarse:
count    6031.000000
mean       59.249212
std         9.603693
min        21.000000
25%        52.000000
50%        59.000000
75%        66.000000
max        95.000000
Name: age, dtype: float64
Estadística Descriptiva para 'age' cuando quiere retirarse:
count    1124.000000
mean       56.672598
std         8.242297
min        38.000000
25%        50.000000
50%        56.000000
75%        62.000000
max        82.000000
Name: age, dtype: float64
```

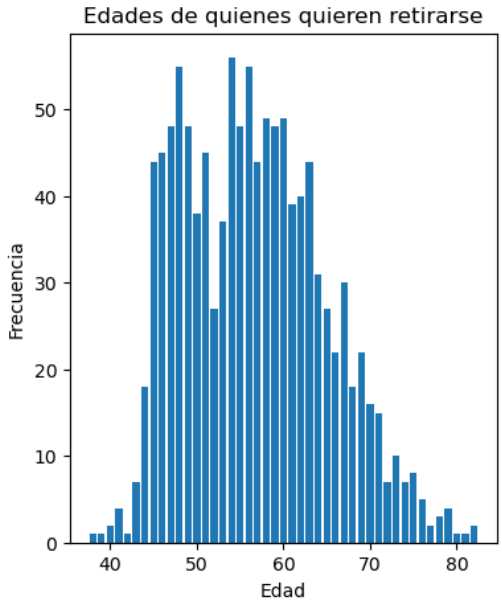
[153]:
```
# Filtrar
age_retin_0 = charls[charls['retin'] == 0]['age']
age_retin_1 = charls[charls['retin'] == 1]['age']

# Crear figura y ejes
fig, axs = plt.subplots(1, 2, figsize=(10, 5))

# Graficar barras para age_retin_0
axs[0].bar(age_retin_0.value_counts().index, age_retin_0.value_counts().values)
axs[0].set_title('Edades de quienes no quieren retirarse')
axs[0].set_xlabel('Edad')
axs[0].set_ylabel('Frecuencia')
# Graficar barras para age_retin_1
axs[1].bar(age_retin_1.value_counts().index, age_retin_1.value_counts().values)
axs[1].set_title('Edades de quienes quieren retirarse')
axs[1].set_xlabel('Edad')
axs[1].set_ylabel('Frecuencia')

plt.subplots_adjust(wspace=0.4)
plt.show()
```





En promedio, las personas que piensan en retirarse tienen 57 años y las que no, 59 años, al ingresar a la encuesta.

0.0.8 Wealth (riqueza neta)

[154]:
```
# Filtrar los datos por "retin" igual a 0 y 1
retin_0_data = charls[charls['retin'] == 0]
retin_1_data = charls[charls['retin'] == 1]

# Realizar la estadística descriptiva para "wealth" en caso de no querer␣
 ↪retirarse
we_retin_0_stats = retin_0_data['wealth'].describe()

# Realizar la estadística descriptiva para "wealth" en caso de querer retirarse
we_retin_1_stats = retin_1_data['wealth'].describe()

# Imprimir los resultados
print("Estadística Descriptiva para 'wealth' cuando NO quiere retirarse:")
print(we_retin_0_stats)

print("Estadística Descriptiva para 'wealth' cuando quiere retirarse:")
print(we_retin_1_stats)
```

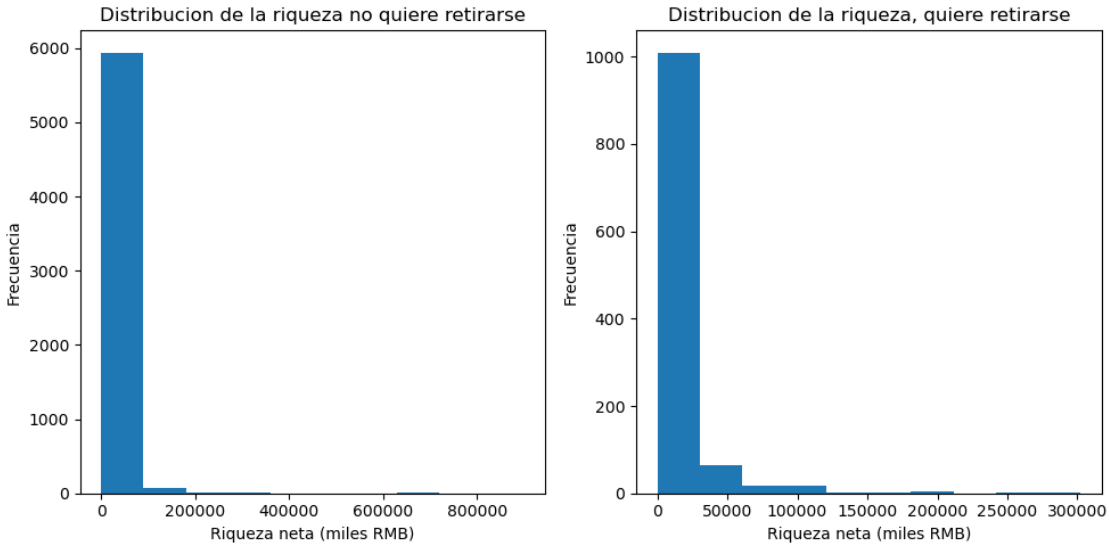
```
Estadística Descriptiva para 'wealth' cuando NO quiere retirarse:
count      6031.000000
mean       7185.616813
std       30363.855926
min           0.000000
```

```
25%         100.000000
50%         500.000000
75%        2625.000000
max      900100.000000
Name: wealth, dtype: float64
Estadística Descriptiva para 'wealth' cuando quiere retirarse:
count      1124.000000
mean      11158.900356
std       28903.154210
min           0.000000
25%         300.000000
50%        1000.000000
75%        7500.000000
max      302000.000000
Name: wealth, dtype: float64
```

[155]:
```python
# Filtrar las filas donde "retin" toma el valor de 0 o 1
retin_0_data = charls[charls['retin'] == 0]
retin_1_data = charls[charls['retin'] == 1]
fig, axs = plt.subplots(1, 2, figsize=(10, 5))

# Graficar el histograma de la variable "wealth" cuando no quiere retirarse
axs[0].hist(retin_0_data['wealth'], bins=10)
axs[0].set_title('Distribucion de la riqueza no quiere retirarse')
axs[0].set_xlabel('Riqueza neta (miles RMB)')
axs[0].set_ylabel('Frecuencia')

# Graficar el histograma de la variable "cesd" cuando quiere retirarse
axs[1].hist(retin_1_data['wealth'], bins=10)
axs[1].set_title('Distribucion de la riqueza, quiere retirarse')
axs[1].set_xlabel('Riqueza neta (miles RMB)')
axs[1].set_ylabel('Frecuencia')
plt.tight_layout()
plt.show()
```



Se puede visualizar que en promedio las personas que tienen mayor riqueza tienen más interés en retirarse que las que tienen menor riqueza.

0.0.9 Pregunta 2

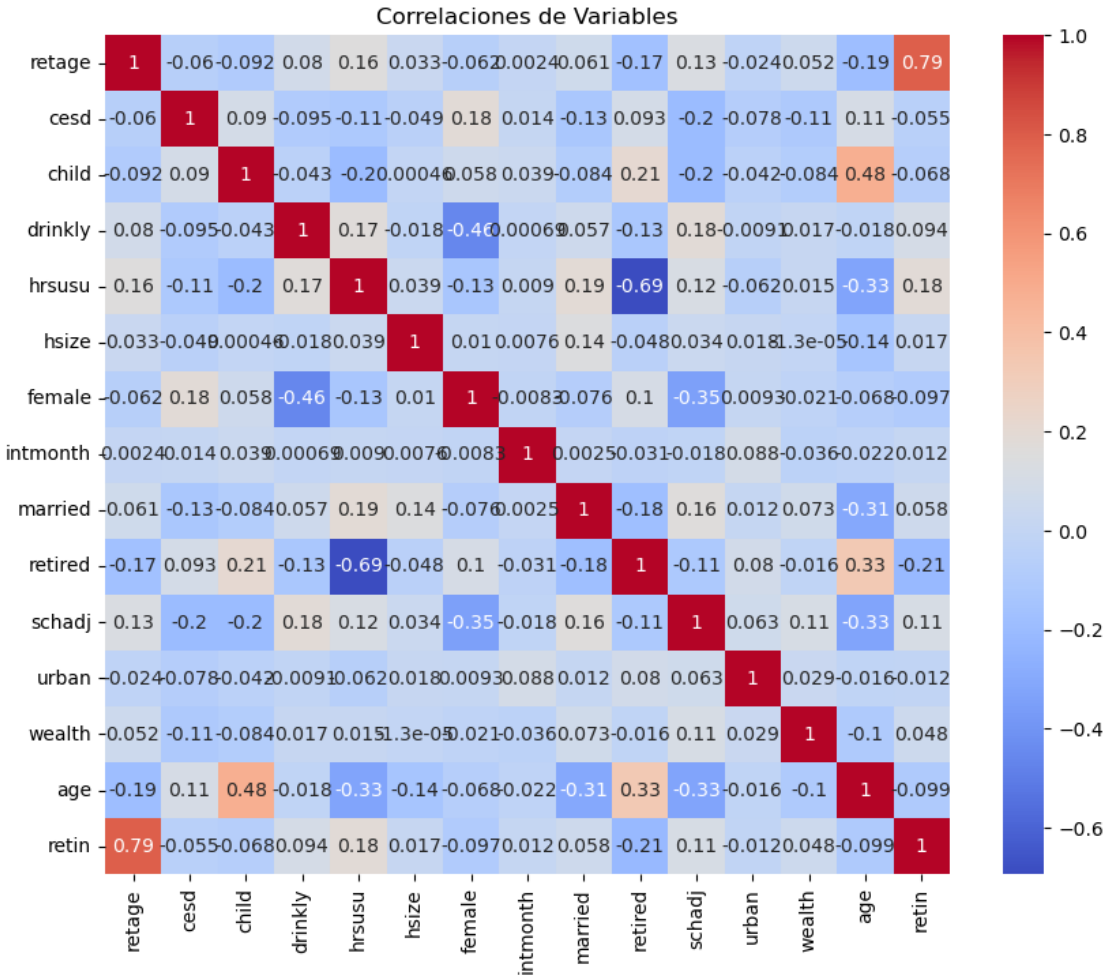
0.0.10 MCO

Se estima en primera instancia una regresión lineal con estimación de mínimos cuadrados ordinarios, y en un principio se pondrán todas las variables que no tengan problemas de multicolinealidad para visualizar cuales son y cuales no son significativas.

[157]:
```
# Seleccionar las variables de interés
variables = ['retage', 'cesd', 'child', 'drinkly', 'hrsusu', 'hsize',
             'female', 'intmonth', 'married', 'retired', 'schadj', 'urban',␣
 ↪'wealth', 'age', 'retin']
df = charls[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()
```



0.0.11 Multicolinealidad

Como sólo se ve una correlación alta, correspondiente a la variable "retage" con la variable dependiente "retin", no hay multicolinealidad, pues no se visualizan correlaciones altas entre las variables explicativas del modelo. La variable drinkly se quitará a pesar de que no tenga una correlación alta con otra variable explicativa, porque esta causa problemas al querer estimar el modelo de regresión. También se quitará la variable "retired" porque a pesar de haberla filtrado anteriormente, esta causa problemas al ser un vector de ceros.

[21]:
```
y=charls['retin']
X=charls[['retage','cesd','child','hrsusu','hsize','female','intmonth','married','schadj','urb
       'age']]
X=sm.add_constant(X)
model = sm.OLS(y, X)
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   retin   R-squared:                       0.643
Model:                             OLS   Adj. R-squared:                  0.643
Method:                  Least Squares   F-statistic:                     1213.
Date:                 Wed, 12 Apr 2023   Prob (F-statistic):               0.00
Time:                         15:18:21   Log-Likelihood:                731.22
No. Observations:                8080   AIC:                            -1436.
Df Residuals:                    8067   BIC:                            -1345.
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.2266      0.033     -6.952      0.000      -0.290      -0.163
retage         0.0701      0.001    116.471      0.000       0.069       0.071
cesd           0.0002      0.000      0.593      0.553      -0.001       0.001
child         -0.0058      0.002     -2.852      0.004      -0.010      -0.002
hrsusu         0.0151      0.001     10.166      0.000       0.012       0.018
hsize         -0.0005      0.001     -0.357      0.721      -0.003       0.002
female        -0.0199      0.006     -3.604      0.000      -0.031      -0.009
intmonth       0.0044      0.002      1.761      0.078      -0.000       0.009
married        0.0211      0.008      2.596      0.009       0.005       0.037
schadj         0.0018      0.001      2.201      0.028       0.000       0.003
urban          0.0088      0.006      1.442      0.149      -0.003       0.021
wealth      1.114e-07    5.7e-08      1.955      0.051   -2.98e-10    2.23e-07
age            0.0037      0.000     10.506      0.000       0.003       0.004
==============================================================================
Omnibus:                     3078.218   Durbin-Watson:                   1.771
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            20565.567
Skew:                           1.675   Prob(JB):                         0.00
Kurtosis:                      10.062   Cond. No.                     5.80e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 5.8e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

0.0.12 Modelo lineal inicial

En el modelo inicial, se puede visualizar que por el valor grande del estadístico F, el modelo es significativo, lo que quiere decir que existe al menos una variable que explica la variación de la variable dependiente, que en este caso es "retin", lo que se interpreta que las variables en general explican la decisión de retirarse o no retirarse. En concreto, la variabilidad de retirarse o no retirarse es explicada en un 64.3% por las variables del modelo. Sin embargo, tal como los indican las pruebas de significancia individual, las variables cesd, hsize, intmonth, urban no son significativas, por lo

que se procederá a quitarlas del modelo en el modelo lineal final y al mismo tiempo comparar los resultados. Cabe señalar que también se quitara la variable "retired" del modelo, debido a que no tiene sentido aplicar un análisis de si la persona se quiere retirar si ya esta retirada al igual que la variable "retage", ya que esta última causa problemas en el número de iteraciones de máxima verosimilitud, la variable wealth no se quitara del modelo ya que está en el umbral de la significancia. Además, hay evidencia de auto correlación positiva en los errores del modelo estimado por MCO, por lo que es posible afirmar que los errores del modelo están correlacionados positivamente, lo que puede indicar que hay un patrón sistemático de dependencia temporal en los errores del modelo.

[28]:
```
y=charls['retin']
X=charls[['child','hrsusu','female','married','schadj','wealth','age']]
X=sm.add_constant(X)
model = sm.OLS(y, X)
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   retin   R-squared:                       0.044
Model:                             OLS   Adj. R-squared:                  0.043
Method:                  Least Squares   F-statistic:                     52.50
Date:                 Wed, 12 Apr 2023   Prob (F-statistic):           1.16e-73
Time:                         15:51:36   Log-Likelihood:                -3254.9
No. Observations:                 8080   AIC:                             6526.
Df Residuals:                     8072   BIC:                             6582.
Df Model:                            7
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1630      0.040      4.034      0.000       0.084       0.242
child         -0.0030      0.003     -0.919      0.358      -0.010       0.003
hrsusu         0.0305      0.002     12.598      0.000       0.026       0.035
female        -0.0426      0.009     -4.746      0.000      -0.060      -0.025
married        0.0112      0.013      0.846      0.397      -0.015       0.037
schadj         0.0063      0.001      4.797      0.000       0.004       0.009
wealth      3.813e-08   9.29e-08      0.410      0.682   -1.44e-07     2.2e-07
age           -0.0014      0.001     -2.481      0.013      -0.003      -0.000
==============================================================================
Omnibus:                     2127.735   Durbin-Watson:                   1.530
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4223.697
Skew:                           1.681   Prob(JB):                         0.00
Kurtosis:                       4.113   Cond. No.                     4.44e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

```
[2] The condition number is large, 4.44e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

0.0.13 Modelo lineal final

Al igual que en el modelo lineal inicial, el modelo es significativo, y, se puede visualizar que el coeficiente de determinación ajustado ahora es solo de 4.3% como consecuencia de haber quitado del modelo la variable "retage". Sigue habiendo un patrón sistemático de dependencia temporal en los errores del modelo. En cuanto a la interpretación de las variables explicativas: - La probabilidad de que la persona se retire disminuye en 0.31% cuando aumenta en 1 el número de hijos - La probabilidad de que la persona se retire aumenta en un 3.05% cuando aumenta en una unidad las horas promedio de trabajo. -Cuando la persona es mujer, disminuye en un 4.26% la probabilidad de retirarse. - Cuando la persona está casada, aumenta un 1.12% la probabilidad de querer retirarse. - Cuando el individuo tiene un año más de escolaridad, la probabilidad de retirarse aumenta en un 0.63%. - Cuando la riqueza neta aumenta en 1000 RMB, la probabilidad de que la persona se retira aumenta ínfimamente según este modelo - Y finalmente, cuando aumenta la edad en un año al ser encuestado, la probabilidad de querer retirarse disminuye en un 0.14%.

0.0.14 Pregunta 3

0.0.15 Probit

[29]:
```
y = charls['retin']
X = charls[['child','hrsusu','female','married','schadj','wealth','age']]

model = sm.Probit(y, X)
probit_model = model.fit()
print(probit_model.summary())

mfx = probit_model.get_margeff()
print(mfx.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.423831
         Iterations 6
                          Probit Regression Results
==============================================================================
Dep. Variable:                   retin   No. Observations:                 8080
Model:                          Probit   Df Residuals:                     8073
Method:                            MLE   Df Model:                            6
Date:                 Wed, 12 Apr 2023   Pseudo R-squ.:                 0.04975
Time:                         15:51:45   Log-Likelihood:                -3424.6
converged:                       True   LL-Null:                       -3603.8
Covariance Type:            nonrobust   LLR p-value:                 2.257e-74
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
child         -0.0069      0.015     -0.466      0.641      -0.036       0.022
hrsusu         0.1262      0.011     11.914      0.000       0.105       0.147
```

```
female        -0.2758      0.035     -7.904      0.000      -0.344      -0.207
married       -0.0704      0.054     -1.299      0.194      -0.177       0.036
schadj         0.0097      0.005      1.979      0.048    9.52e-05       0.019
wealth      2.868e-07   3.91e-07      0.734      0.463   -4.79e-07    1.05e-06
age           -0.0201      0.001    -17.528      0.000      -0.022      -0.018
==============================================================================
       Probit Marginal Effects
====================================
Dep. Variable:                   retin
Method:                           dydx
At:                           overall
==============================================================================
                dy/dx    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
child         -0.0016      0.003     -0.466      0.641      -0.008       0.005
hrsusu         0.0297      0.002     12.064      0.000       0.025       0.034
female        -0.0648      0.008     -7.965      0.000      -0.081      -0.049
married       -0.0166      0.013     -1.299      0.194      -0.042       0.008
schadj         0.0023      0.001      1.979      0.048    2.24e-05       0.005
wealth      6.741e-08   9.19e-08      0.734      0.463   -1.13e-07    2.47e-07
age           -0.0047      0.000    -18.277      0.000      -0.005      -0.004
==============================================================================
```

Se utilizaron las mismas variables que en la estimación por Mínimos Cuadrados Ordinarios. En este modelo, los valores de los coeficientes estimados son irrelevantes para el análisis, lo que sí es importante conocer son la significancia individual y los cambios marginales. Se puede apreciar que ahora el número de hijos (child), el hecho de esta casado o no (married) y la riqueza neta (wealth) no son variables significativas para el modelo. Tras esto, las variables significativas son: hrsusu, female, schadj y age. En cuanto a los cambios marginales, si la persona es mujer y la edad a la que fue encuestado el individuo explican negativamente al hecho de querer retirarse. Por ejemplo, si la persona es mujer, disminuye en un 6.48% la probabilidad de que se retire. En contraste, las horas medias de trabajo diario y los años de escolaridad explican positivamente al hecho de querer retirarse. Por ejemplo, si aumenta en una unidad las horas medias de trabajo diario, la probabilidad de querer retirarse aumenta en un 2.97%. Se observa que el Pseudo Rsquared es superior al R squared de la estimación por MCO, en concreto, de un 4.975%, por lo que es posible afirmar que el Modelo Probit es superior al Modelo de Regresión por estimación MCO.

0.0.16 Pregunta 4

0.0.17 Logit

[30]:
```
y = charls['retin']
X = charls[['child','hrsusu','female','married','schadj','wealth','age']]

model = sm.Logit(y, X)
logit_model = model.fit()
print(logit_model.summary())
```

```
mfx = logit_model.get_margeff()
print(mfx.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.424148
         Iterations 6
                          Logit Regression Results
==============================================================================
Dep. Variable:                   retin   No. Observations:                 8080
Model:                           Logit   Df Residuals:                     8073
Method:                            MLE   Df Model:                            6
Date:                Wed, 12 Apr 2023   Pseudo R-squ.:                 0.04903
Time:                        15:54:46   Log-Likelihood:                -3427.1
converged:                       True   LL-Null:                       -3603.8
Covariance Type:            nonrobust   LLR p-value:                2.851e-73
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
child         -0.0160      0.027     -0.588      0.556      -0.069       0.037
hrsusu         0.2259      0.020     11.448      0.000       0.187       0.265
female        -0.4820      0.063     -7.675      0.000      -0.605      -0.359
married       -0.1138      0.100     -1.142      0.253      -0.309       0.081
schadj         0.0205      0.009      2.346      0.019       0.003       0.038
wealth      4.823e-07   6.87e-07      0.702      0.482   -8.64e-07    1.83e-06
age           -0.0347      0.002    -16.461      0.000      -0.039      -0.031
==============================================================================
        Logit Marginal Effects
=====================================
Dep. Variable:                   retin
Method:                           dydx
At:                           overall
==============================================================================
                dy/dx    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
child         -0.0021      0.004     -0.588      0.556      -0.009       0.005
hrsusu         0.0297      0.003     11.499      0.000       0.025       0.035
female        -0.0635      0.008     -7.732      0.000      -0.080      -0.047
married       -0.0150      0.013     -1.143      0.253      -0.041       0.011
schadj         0.0027      0.001      2.346      0.019       0.000       0.005
wealth       6.35e-08   9.04e-08      0.702      0.482   -1.14e-07    2.41e-07
age           -0.0046      0.000    -17.107      0.000      -0.005      -0.004
==============================================================================
```

Se utilizaron las mismas variables que en la estimación por Mínimos Cuadrados Ordinarios y que en el Modelo Probit. En este modelo (al igual que en Probit), los valores de los coeficientes estimados son irrelevantes para el análisis, lo que sí es importante conocer son la significancia individual y los cambios marginales. En este caso, las variables número de hijos (child), si está casado(a) (married) y riqueza neta (wealth) no son significativas para el modelo. Luego, las variables significativas son

hrsusu, female, schadj y age. En relación a los cambios marginales, si la persona es mujer y la edad a la que fue encuestado el individuo explican negativamente al hecho de querer retirarse. Por ejemplo, si la persona es mujer, la probabilidad de querer retirarse disminuye en un 6.35%. En cambio, las horas medias de trabajo diario y los años de escolaridad explican positivamente al deseo de retirarse. Por ejemplo, si aumenta en una unidad los años de escolaridad, la probabilidad de querer retirarse aumenta en un 0.27%. Se observa que el Pseudo Rsquared es superior al R squared de la estimación por MCO, pero inferior al Pseudo R squared del Modelo Probit, en concreto, de un 4.903%, por lo que es posible afirmar, en este caso, que el Modelo Logit es superior al Modelo de Regresión por estimación MCO pero inferior al Modelo Probit.

0.0.18 Pregunta 5: Comparación de los modelos Probit y Logit

A priori, se sabe que la estimación MCO es débil al querer estimar variables binarias, ya que gráficamente en gran parte de la recta de regresión no están incluidos los valores de 0 y 1 y además se asume que los errores se distribuyen normalmente y tienen una varianza constante. Esta suposición no se cumple cuando se trabaja con variables binarias, ya que los errores tienen una distribución Bernoulli o binomial. Por estas razones, la estimación por mínimos cuadrados ordinarios no es la mejor técnica para modelar variables binarias. En su lugar, se utilizan otras técnicas de modelado estadístico específicas para variables binarias, como la regresión Logit y la regresión Probit. En este caso, al comparar las bondades de ajuste de los modelos se llega a la conclusión de que el mejor modelo para este caso particular es el Probit, lo cual tiene sentido, ya que la regresión Probit es usada mayoritariamente cuando se investiga sobre temas de ciencias sociales, en cambio la regresión Logit se utiliza más en campos relacionados con las ciencias médicas y biológicas y, como en este caso se desea estudiar el deseo de jubilarse, se está tratando de una investigación relacionada con ciencias sociales. Sin embargo, se observa que los 3 modelos presentaron un mal ajuste al necesariamente retirar la variable "retage", por lo cual es necesario buscar otras variables que resulten ser significativas para poder explicar en mayor porcentaje al modelo. En todos los casos, las variables que resultaron ser significativas son las horas medias de trabajo diario, si es mujer, los años de escolaridad y la edad al ser encuestado.

0.0.19 Pregunta 6

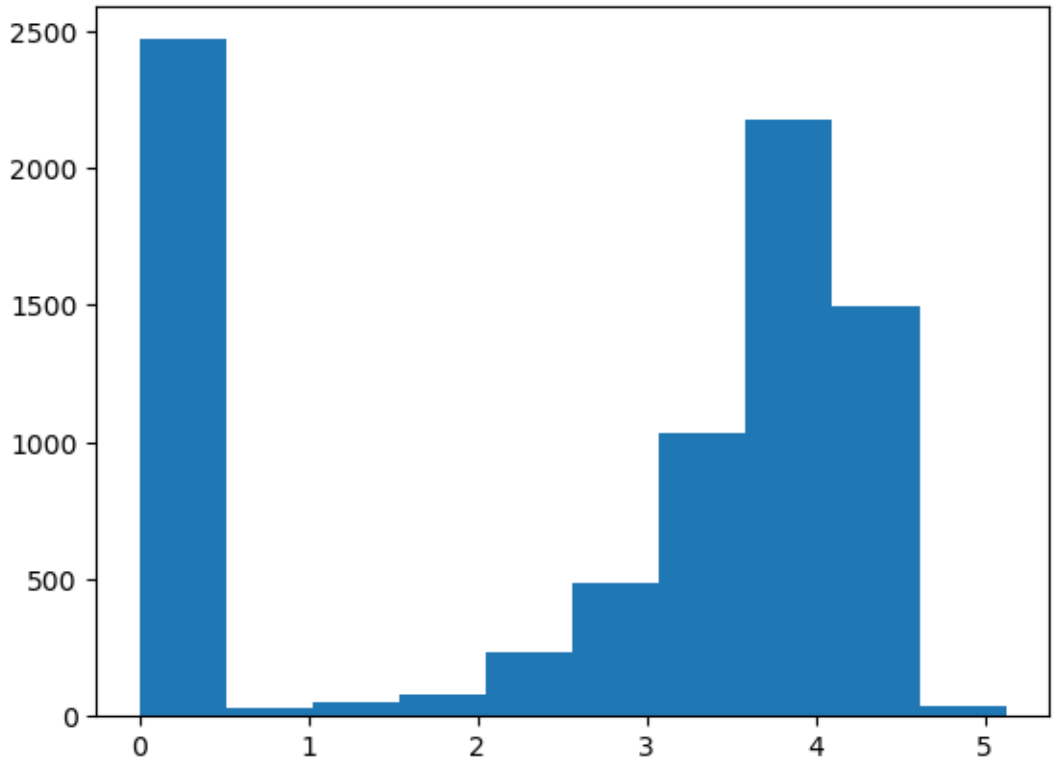
Se decidió quitar la variable "intmonth" ya que no tiene sentido relacionar el mes en que se realizó la encuesta con los años en los que la persona desea retirarse. Al igual que en los modelos anteriores, se decidió no introducir las variables "retin" y "retage" en el mismo modelo, en este caso porque la variación de una infla la variación de otra (el coeficiente estimado es muy superior al 100%).

0.0.20 Poisson inicial

[48]:
```
y=charls['retage']
X=charls[['cesd','child','hrsusu','hsize','female','married','schadj','urban','wealth',
       'age']]
plt.hist(charls.hrsusu)
charls.hrsusu.head()
```

[48]:
```
0    0.000000
1    4.143135
2    0.000000
```

```
3    0.000000
4    3.806663
Name: hrsusu, dtype: float64
```



[49]:
```
poisson=sm.GLM(y,X,family=sm.families.Poisson()).fit()
print(poisson.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 retage   No. Observations:                 8080
Model:                            GLM   Df Residuals:                     8070
Model Family:                 Poisson   Df Model:                            9
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -24377.
Date:                Wed, 12 Apr 2023   Deviance:                       43746.
Time:                        18:46:27   Pearson chi2:                 8.56e+04
No. Iterations:                     6   Pseudo R-squ. (CS):             0.4758
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
cesd          -0.0021      0.002     -1.320      0.187      -0.005       0.001
```

```
child          -0.0429      0.009     -4.763      0.000      -0.060      -0.025
hrsusu          0.2726      0.007     39.544      0.000       0.259       0.286
hsize           0.0574      0.005     11.181      0.000       0.047       0.067
female         -0.0454      0.019     -2.333      0.020      -0.084      -0.007
married         0.6760      0.044     15.514      0.000       0.591       0.761
schadj          0.0647      0.003     25.118      0.000       0.060       0.070
urban          -0.1324      0.024     -5.627      0.000      -0.178      -0.086
wealth     -6.153e-07   1.87e-07     -3.283      0.001   -9.83e-07   -2.48e-07
age            -0.0258      0.001    -30.636      0.000      -0.027      -0.024
==============================================================================
```

Dado que la variable cesd ahora no es significativa en este modelo, se procederá a quitarla y a comparar con el segundo modelo de Poisson. El Pseudo R-squared indica que la edad a retirarse es explicada en un 47.58% por las variables explicativas del modelo.

0.0.21 Poisson final

[63]:
```
y2=charls['retage']
X2=charls[['child','hrsusu','hsize','female','married','schadj','urban','wealth','age']]
poisson2=sm.GLM(y2,X2,family=sm.families.Poisson()).fit()
print(poisson2.summary())
print("fitted lambda")
print(poisson2.mu)
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                  retage   No. Observations:                 8080
Model:                            GLM   Df Residuals:                     8071
Model Family:                 Poisson   Df Model:                            8
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -24378.
Date:                Wed, 12 Apr 2023   Deviance:                       43747.
Time:                        21:01:42   Pearson chi2:                 8.55e+04
No. Iterations:                     6   Pseudo R-squ. (CS):             0.4757
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
child          -0.0434      0.009     -4.834      0.000      -0.061      -0.026
hrsusu          0.2725      0.007     39.565      0.000       0.259       0.286
hsize           0.0573      0.005     11.155      0.000       0.047       0.067
female         -0.0503      0.019     -2.627      0.009      -0.088      -0.013
married         0.6758      0.044     15.527      0.000       0.590       0.761
schadj          0.0650      0.003     25.309      0.000       0.060       0.070
urban          -0.1306      0.023     -5.561      0.000      -0.177      -0.085
wealth     -5.95e-07   1.87e-07     -3.180      0.001   -9.62e-07   -2.28e-07
age            -0.0260      0.001    -31.829      0.000      -0.028      -0.024
==============================================================================
fitted lambda
```

```
[0.6530125  2.61462956 0.58736359 … 0.9134986  1.63530732 0.38756004]
```

Se puede visualizar que el valor del Pseudo R-Squared es prácticamente el mismo que en el modelo Poisson inicial, por lo cual es posible afirmar que el tiempo en el que planea retirarse el individuo es explicada de igual manera por las variables que quedaron seleccionadas. - El aumento del número de hijos explica una disminución en 4.34% en el tiempo en el que planea retirarse el individuo. - El aumento de las horas medias de trabajo diario implica un aumento en un 27.25% en el tiempo en el que planea retirarse el individuo. - El aumento del tamaño del hogar implica un aumento en un 5.73% en el tiempo en el que planea retirarse el individuo. - El hecho de ser mujer disminuye en un 5.03% el tiempo en el que planea retirarse el individuo. - El hecho de estar casado implica un aumento del 67.58% del tiempo en el que planea retirarse el individuo. - En este caso, un aumento en los años de escolaridad implica una disminución porcentual de 6.5% en el tiempo en el que planea retirarse el individuo. - El hecho de que el individuo viva en una zona urbana implica una disminución del 13.06% del tiempo en el que planea retirarse el individuo. - Un aumento en la riqueza neta implica una pequeña disminución en el tiempo en el que planea retirarse el individuo. - El aumento de la edad al ser encuestado implica una disminución del 2.6% en el tiempo en el que planea retirarse el individuo.

0.0.22 Pregunta 7

0.0.23 Test sobre dispersion

A simple test for overdispersion can be determined with the results of the Poisson model, using the ratio of Pearson chi2 / Df Residuals. A value larger than 1 indicates overdispersion. In the case above (6), data suggets overdispersion.

The Negative Binomial model estimated above is using a value of θ (or $\alpha = 1/\theta$) equal to 1. In order to determine the appropiate value of α, you can estimate a simple regression using the output of the Poisson model:

1. Construct the following variable aux=$[(y-\lambda)^2 - \lambda]/\lambda$
2. Regress the variable aux with λ as the only explanatory variable (no constant)
3. The estimated value is an appropiate guess for $\alpha = 1/\theta$

In the model of the previous section, just use the options on sm.families.NegativeBinomial, in order to manually enter the value of alpha. See example below.

Se aplicará este test sobre dispersión al modelo Poisson inicial.

[60]:
```
y=charls['retage']
X=charls[['cesd','child','hrsusu','hsize','female','married','schadj','urban','wealth','age']]
aux=((y2-poisson.mu)**2-poisson.mu)/poisson.mu
auxr=sm.OLS(aux,poisson.mu).fit()
print(auxr.summary())
print(auxr.params)
```

```
                                 OLS Regression Results
==============================================================================
=======
Dep. Variable:                    retage   R-squared (uncentered):
0.035
```

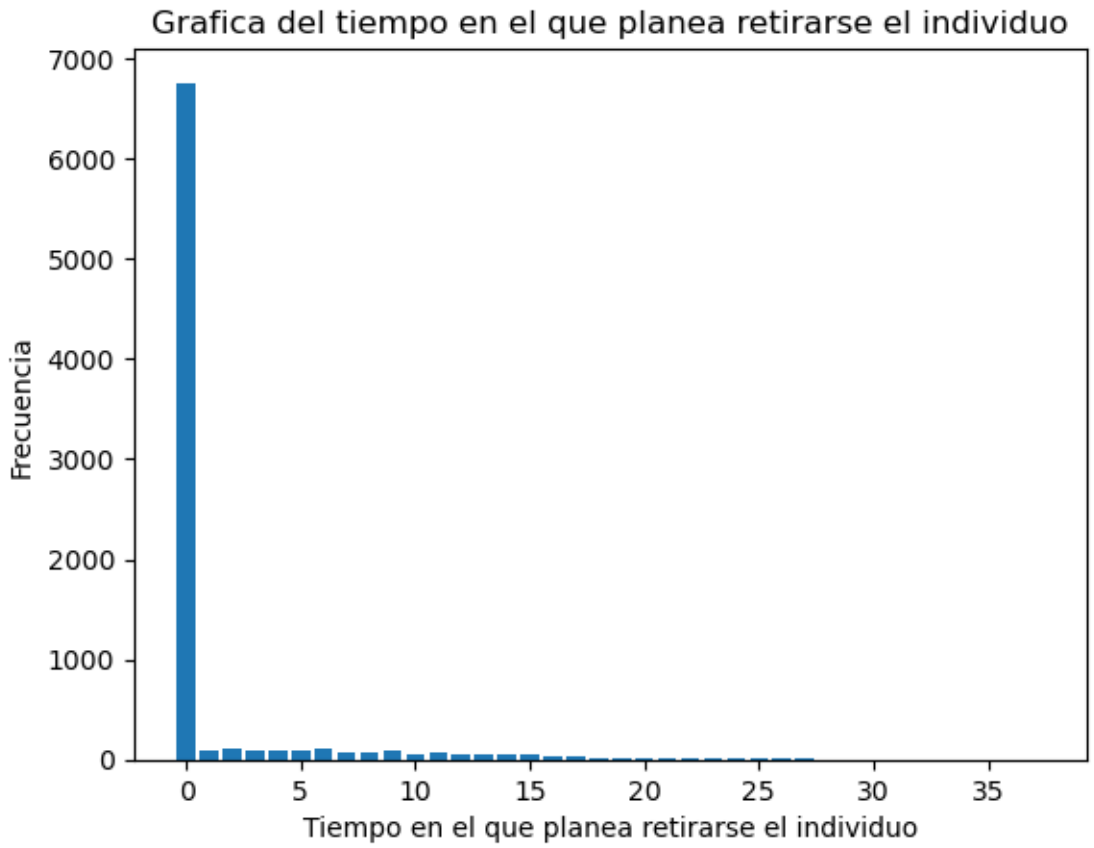
```
Model:                            OLS   Adj. R-squared (uncentered):
0.035
Method:                 Least Squares   F-statistic:
290.2
Date:                Wed, 12 Apr 2023   Prob (F-statistic):
5.88e-64
Time:                        20:15:01   Log-Likelihood:
-42451.
No. Observations:                8080   AIC:
8.490e+04
Df Residuals:                    8079   BIC:
8.491e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             5.0052      0.294     17.035      0.000       4.429       5.581
==============================================================================
Omnibus:                    13181.877   Durbin-Watson:                   1.659
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         10512107.605
Skew:                          10.783   Prob(JB):                         0.00
Kurtosis:                     178.382   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
x1    5.005199
dtype: float64
```

Como alpha es mayor que 1, se considera que hay sobredispersión, lo que significa que la varianza es mayor que la media, lo que a su vez indica que los datos no siguen una distribución de Poisson, por lo que la distribución de Poisson no es adecuada para modelar estos datos. Por ello, es necesario utilizar un modelo como la distribución binomial negativa, para tener en cuenta la sobredispersión y obtener resultados más precisos y fiables.

[62]:
```
retage = charls['retage']
freq = retage.value_counts()
freq = freq.sort_index()

plt.bar(freq.index, freq.values)
plt.xlabel('Tiempo en el que planea retirarse el individuo')
plt.ylabel('Frecuencia')
plt.title('Grafica del tiempo en el que planea retirarse el individuo')
plt.show()
```



Gráficamente se verifica que la varianza es mayor que la media ya que existe una gran proporción de ceros, por lo cual se confirma que la solución es utilizar la distribución binomial negativa.

0.0.24 Pregunta 8

0.0.25 Binomial Negativa

[61]:
```
y=charls['retage']
X=charls[['cesd','child','hrsusu','hsize','female','married','schadj','urban','wealth','age']]
negbin=sm.GLM(y,X,family=sm.families.NegativeBinomial()).fit()
print(negbin.summary())
print("fitted lambda")
print(negbin.mu)
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 retage   No. Observations:                 8080
Model:                            GLM   Df Residuals:                     8070
Model Family:        NegativeBinomial   Df Model:                            9
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -12328.
```

```
Date:                Wed, 12 Apr 2023   Deviance:                       16697.
Time:                        20:15:59   Pearson chi2:                 3.84e+04
No. Iterations:                     9   Pseudo R-squ. (CS):             0.2547
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
cesd          -0.0036      0.003     -1.429      0.153      -0.009       0.001
child         -0.0507      0.014     -3.654      0.000      -0.078      -0.024
hrsusu         0.3004      0.010     30.688      0.000       0.281       0.320
hsize          0.0697      0.008      8.255      0.000       0.053       0.086
female        -0.0409      0.032     -1.270      0.204      -0.104       0.022
married        0.5452      0.058      9.478      0.000       0.432       0.658
schadj         0.0690      0.004     15.669      0.000       0.060       0.078
urban         -0.2003      0.039     -5.143      0.000      -0.277      -0.124
wealth     -8.057e-07   3.37e-07     -2.393      0.017   -1.47e-06   -1.46e-07
age           -0.0255      0.001    -21.083      0.000      -0.028      -0.023
==============================================================================
fitted lambda
[0.60194095 2.78512029 0.56145027 … 0.83552925 1.66815709 0.47427989]
```

En este caso no resultan significativas las variables cesd y female. Todas las demás variables expuestas en el sumario resultaron significativas, correspondientes a child, hrsusu, hsize, married, schadj, urban, wealth y age. Para el caso de las relaciones positivas, por ejemplo, un aumento de las horas medias de trabajo diario implica un aumento en el tiempo en el que planea retirarse el individuo en un 30.04%, en contraste, en el caso las relaciones negativas, por ejemplo, un aumento en el número de hijos implica una disminución del tiempo en el que planea retirarse el individuo en un 5.07%.

0.0.26 Pregunta 9: Comparacion entre Poisson y Binomial Negativa

Claramente, debido a que hay sobre dispersión y una gran proporción de ceros, es posible afirmar que es mejor ocupar la distribución binomial negativa, la cual dice que las variables cesd (puntaje de salud mental) y female (si es mujer la persona) son no significativas a diferencia del modelo Poisson que solo consideraba no significativa a la variable cesd.

Tras lo dicho anteriormente, solo se considerarán los resultados de la distribución Binomial Negativa, los cuales dicen que las variables número de hijos (child), horas medias de trabajo diario (hrsusu), tamaño del hogar (hsize), si está casado(a) (married), años de escolaridad (schadj), si está en zona urbana o no (urban), la riqueza neta (wealth) y la edad a ser encuestado; son las que explican al tiempo en el que planea retirarse el individuo. El Pseudo R-Squared sugiere que el tiempo en el que planea retirarse el individuo es verdaderamente explicado (ya que el modelo Binomial Negativa se considera más apropiado) en un 25.47% por las variables antes mencionadas.