nonlinear-Copy1

April 25, 2023

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 11/04/23 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
- drinly: bebio 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

[37]:
```
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
```

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.

R: Se ajusto drinkly como variable numerica y se paso wealth a logs, ademas de agregar una variable que indica cuando wealth no existe, dwealth (se podria haber hecho lo mismo con cesd). En general solo se ven valores extremos en wealth, pero no amerita clasificarlos como outliers.

[100]:
```
charls = pd.read_csv('../data/charls2.csv')
charls = charls.replace({'.r': np.nan, '.m': np.nan, '.d': np.nan})
charls['drinkly'] = charls['drinkly'].astype(float)
charls['wealth']=charls['wealth']/100000
charls['dwealth']=0
charls.loc[charls['wealth'].isnull(), 'dwealth'] = 1
charls['lwealth']=np.log(charls['wealth']-charls['wealth'].min()+0.1)
charls.loc[charls['lwealth'].isnull(), 'lwealth'] = 0
charls.reset_index(drop=True, inplace=True)
charls.describe()
```

[100]:

	age	cesd	child	drinkly	female	\
count	9456.000000	8802.000000	9456.000000	9418.000000	9456.000000	
mean	58.087035	9.034992	2.751586	0.334678	0.525275	
std	9.462629	6.462808	1.400139	0.471903	0.499387	
min	21.000000	0.000000	0.000000	0.000000	0.000000	
25%	50.000000	4.000000	2.000000	0.000000	0.000000	
50%	57.000000	8.000000	2.000000	0.000000	1.000000	
75%	64.000000	13.000000	3.000000	1.000000	1.000000	
max	100.000000	30.000000	10.000000	1.000000	1.000000	

	hrsusu	hsize	intmonth	married	retage	\
count	9456.000000	9456.000000	9456.000000	9456.000000	9456.000000	
mean	2.552777	3.758249	7.495347	0.885364	1.390969	
std	1.802885	1.823791	1.009306	0.318599	4.102102	
min	0.000000	1.000000	1.000000	0.000000	0.000000	
25%	0.000000	2.000000	7.000000	1.000000	0.000000	
50%	3.555348	3.000000	7.000000	1.000000	0.000000	
75%	4.025352	5.000000	8.000000	1.000000	0.000000	
max	5.123964	16.000000	12.000000	1.000000	37.000000	

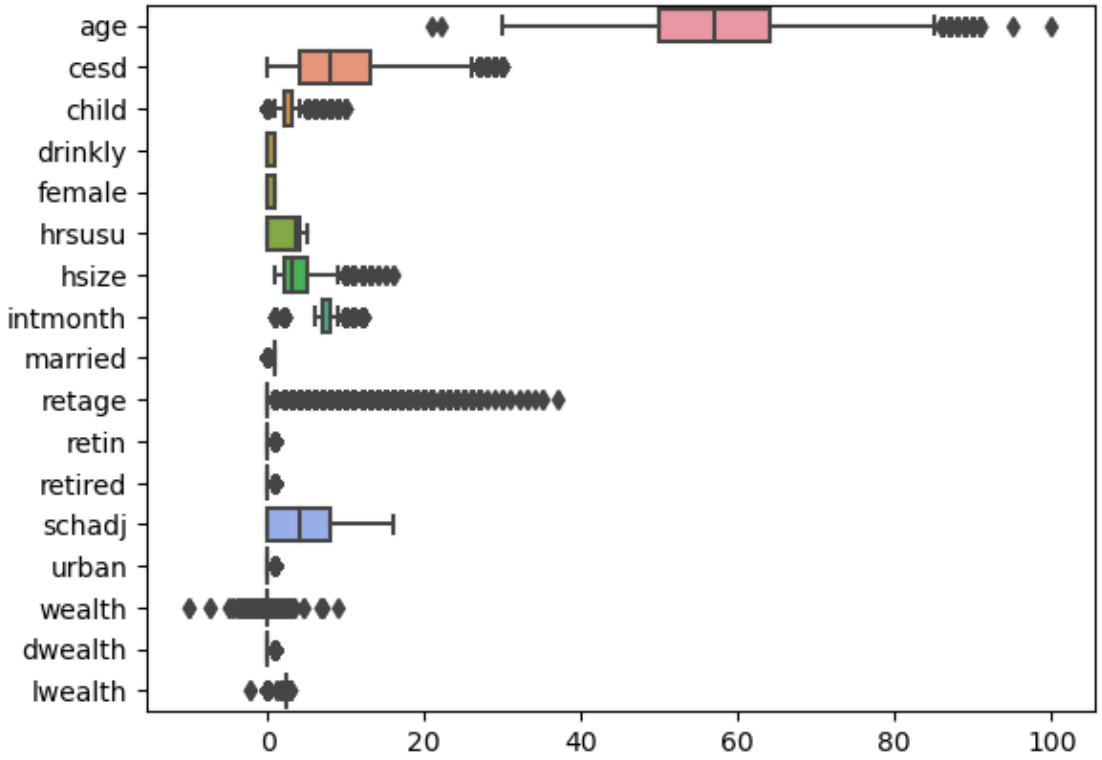
	retin	retired	schadj	urban	wealth	\
count	9456.000000	9456.000000	9456.000000	9456.000000	8590.000000	
mean	0.152602	0.183376	4.100888	0.213832	0.013671	
std	0.359622	0.386995	3.574570	0.410032	0.431023	
min	0.000000	0.000000	0.000000	0.000000	-10.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.000000	

```
50%       0.000000     0.000000     4.000000     0.000000     0.004000
75%       0.000000     0.000000     8.000000     0.000000     0.021875
max       1.000000     1.000000    16.000000     1.000000     9.001000

            dwealth      lwealth
count   9456.000000  9456.000000
mean       0.091582     2.100429
std        0.288450     0.671585
min        0.000000    -2.302585
25%        0.000000     2.312535
50%        0.000000     2.312832
75%        0.000000     2.314514
max        1.000000     2.949741
```

[101]:
```
sns.boxplot(data=charls, orient='h')
```

[101]: <Axes: >



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 dependientes a incluir en el modelo final e interprete su significado.

R: Se excluye intmonth por ser irrelevante, y se excluyen aquellas observaciones que no tienen

valor en wealth (al tratar de agregarlas con una variable dummy dwealth, se encuentra el mismo resultado, por lo cual no aportan al analisis). En vista de lo estimado, Aspectos socioeconomicos y de genero influyen en la probabilidad de querer retirarse (condicional en estar trabajando). Por ejemplo, mujeres tienen 4% menor probabilidad de querer retirarse, y la probabilidad incrementa en 0,78% por año de escolaridad. Otros aspectos demograficos no son relevantes, sin embargo aquellos que declaran beber en el mes pasado tambien son mas probables de querer retirarse (2,3%).

[102]:
```
charls.reset_index(drop=True, inplace=True)
X=charls[charls['retired'] == 0].reset_index(drop=True)
X=X[['retin','age', 'cesd', 'child', 'drinkly', 'female', 'hsize','married',␣
 ↪'hrsusu', 'schadj', 'urban','lwealth']]
X.dropna(inplace=True)
y=X['retin']
X=X[['age', 'cesd', 'child', 'drinkly', 'female', 'hsize','married', 'hrsusu',␣
 ↪'schadj', 'urban','lwealth']]
X=sm.add_constant(X)
model = sm.OLS(y, X)
results = model.fit(cov_type='HC0')
print(results.summary())
```

OLS Regression Results

Dep. Variable:	retin	R-squared:	0.019
Model:	OLS	Adj. R-squared:	0.017
Method:	Least Squares	F-statistic:	12.25
Date:	Mon, 24 Apr 2023	Prob (F-statistic):	3.33e-23
Time:	12:33:26	Log-Likelihood:	-3541.6
No. Observations:	7194	AIC:	7107.
Df Residuals:	7182	BIC:	7190.
Df Model:	11		
Covariance Type:	HC0		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1719	0.052	3.274	0.001	0.069	0.275
age	-0.0006	0.001	-0.963	0.336	-0.002	0.001
cesd	-0.0010	0.001	-1.420	0.156	-0.002	0.000
child	-0.0064	0.004	-1.652	0.099	-0.014	0.001
drinkly	0.0232	0.012	2.008	0.045	0.001	0.046
female	-0.0386	0.012	-3.293	0.001	-0.062	-0.016
hsize	0.0014	0.003	0.556	0.578	-0.004	0.007
married	0.0117	0.016	0.720	0.471	-0.020	0.044
hrsusu	0.0097	0.003	3.237	0.001	0.004	0.016
schadj	0.0078	0.002	4.995	0.000	0.005	0.011
urban	0.0065	0.012	0.551	0.581	-0.017	0.030
lwealth	0.0095	0.007	1.366	0.172	-0.004	0.023

Omnibus:	1470.173	Durbin-Watson:	1.489

```
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2587.357
Skew:                           1.463   Prob(JB):                         0.00
Kurtosis:                       3.256   Cond. No.                         672.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
```

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.

R: Notar que los resultados son bastante similares a OLS, aunque la precision en algunas variables es menor, como drinkly. Al mirar los efectos marginales, los resultados son arbitrariamente identicos.

[82]:
```
model = sm.Logit(y, X)
results = model.fit(cov_type='HC0')
print(results.summary())
mfx = results.get_margeff()
print(mfx.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.490314
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                   retin   No. Observations:                 7194
Model:                           Logit   Df Residuals:                     7182
Method:                            MLE   Df Model:                           11
Date:                Mon, 24 Apr 2023   Pseudo R-squ.:                 0.01872
Time:                        12:22:13   Log-Likelihood:                -3527.3
converged:                       True   LL-Null:                       -3594.6
Covariance Type:                  HC0   LLR p-value:                 2.027e-23
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.5935      0.346     -4.604      0.000      -2.272      -0.915
age           -0.0044      0.004     -1.001      0.317      -0.013       0.004
cesd          -0.0068      0.005     -1.363      0.173      -0.016       0.003
child         -0.0452      0.027     -1.702      0.089      -0.097       0.007
drinkly        0.1369      0.070      1.954      0.051      -0.000       0.274
female        -0.2539      0.075     -3.391      0.001      -0.401      -0.107
hsize          0.0096      0.017      0.567      0.571      -0.024       0.043
married        0.0961      0.119      0.809      0.418      -0.137       0.329
hrsusu         0.0670      0.021      3.129      0.002       0.025       0.109
schadj         0.0473      0.009      5.034      0.000       0.029       0.066
urban          0.0393      0.073      0.535      0.592      -0.105       0.183
lwealth        0.0629      0.047      1.339      0.180      -0.029       0.155
==============================================================================
        Logit Marginal Effects
```

```
====================================
Dep. Variable:                   retin
Method:                           dydx
At:                            overall
================================================================================
                dy/dx    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
age           -0.0007      0.001     -1.001      0.317      -0.002       0.001
cesd          -0.0011      0.001     -1.363      0.173      -0.003       0.000
child         -0.0071      0.004     -1.703      0.089      -0.015       0.001
drinkly        0.0214      0.011      1.955      0.051    -5.2e-05       0.043
female        -0.0398      0.012     -3.394      0.001      -0.063      -0.017
hsize          0.0015      0.003      0.567      0.571      -0.004       0.007
married        0.0151      0.019      0.809      0.418      -0.021       0.052
hrsusu         0.0105      0.003      3.131      0.002       0.004       0.017
schadj         0.0074      0.001      5.062      0.000       0.005       0.010
urban          0.0062      0.011      0.535      0.592      -0.016       0.029
lwealth        0.0099      0.007      1.339      0.181      -0.005       0.024
================================================================================
```

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.

R: Nuevamente hay diferencias numericas minimas entre cada modelo, pero los resultados son virtualmente identicos entre LPM, Probit y Logit.

[103]:
```
model = sm.Probit(y, X)
results = model.fit(cov_type='HC0')
print(results.summary())
mfx = results.get_margeff()
print(mfx.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.490366
         Iterations 5
                          Probit Regression Results
================================================================================
Dep. Variable:                   retin   No. Observations:                 7194
Model:                         Probit   Df Residuals:                     7182
Method:                            MLE   Df Model:                           11
Date:                Mon, 24 Apr 2023   Pseudo R-squ.:                 0.01862
Time:                        12:33:51   Log-Likelihood:                -3527.7
converged:                       True   LL-Null:                       -3594.6
Covariance Type:                  HC0   LLR p-value:                 2.886e-23
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const         -0.9547      0.196     -4.880      0.000      -1.338      -0.571
age           -0.0026      0.002     -1.056      0.291      -0.008       0.002
```

cesd	-0.0037	0.003	-1.310	0.190	-0.009	0.002
child	-0.0245	0.015	-1.633	0.103	-0.054	0.005
drinkly	0.0796	0.040	1.985	0.047	0.001	0.158
female	-0.1454	0.042	-3.426	0.001	-0.229	-0.062
hsize	0.0049	0.010	0.504	0.614	-0.014	0.024
married	0.0529	0.066	0.802	0.423	-0.076	0.182
hrsusu	0.0383	0.012	3.167	0.002	0.015	0.062
schadj	0.0266	0.005	4.927	0.000	0.016	0.037
urban	0.0195	0.042	0.461	0.645	-0.063	0.102
lwealth	0.0389	0.027	1.459	0.145	-0.013	0.091

Probit Marginal Effects

Dep. Variable:	retin
Method:	dydx
At:	overall

	dy/dx	std err	z	P>|z|	[0.025	0.975]
age	-0.0007	0.001	-1.057	0.291	-0.002	0.001
cesd	-0.0010	0.001	-1.310	0.190	-0.003	0.001
child	-0.0067	0.004	-1.633	0.102	-0.015	0.001
drinkly	0.0218	0.011	1.986	0.047	0.000	0.043
female	-0.0399	0.012	-3.430	0.001	-0.063	-0.017
hsize	0.0013	0.003	0.504	0.614	-0.004	0.007
married	0.0145	0.018	0.802	0.423	-0.021	0.050
hrsusu	0.0105	0.003	3.170	0.002	0.004	0.017
schadj	0.0073	0.001	4.948	0.000	0.004	0.010
urban	0.0053	0.012	0.461	0.645	-0.017	0.028
lwealth	0.0107	0.007	1.459	0.145	-0.004	0.025

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?

R: Ante las minimas diferencias entre los distintos metodos, no hay preferencia entre Probit y Logit, pero LPM es inferior dado que producira predicciones fuera del intervalo de interes. En cualquier caso, dado que el modelo no es causal, podemos inferir que el set de variables disponibles explica una fraccion muy menor de la intencion de retirarse, y en virtud de aquello, las predicciones del modelo seran poco confiables.

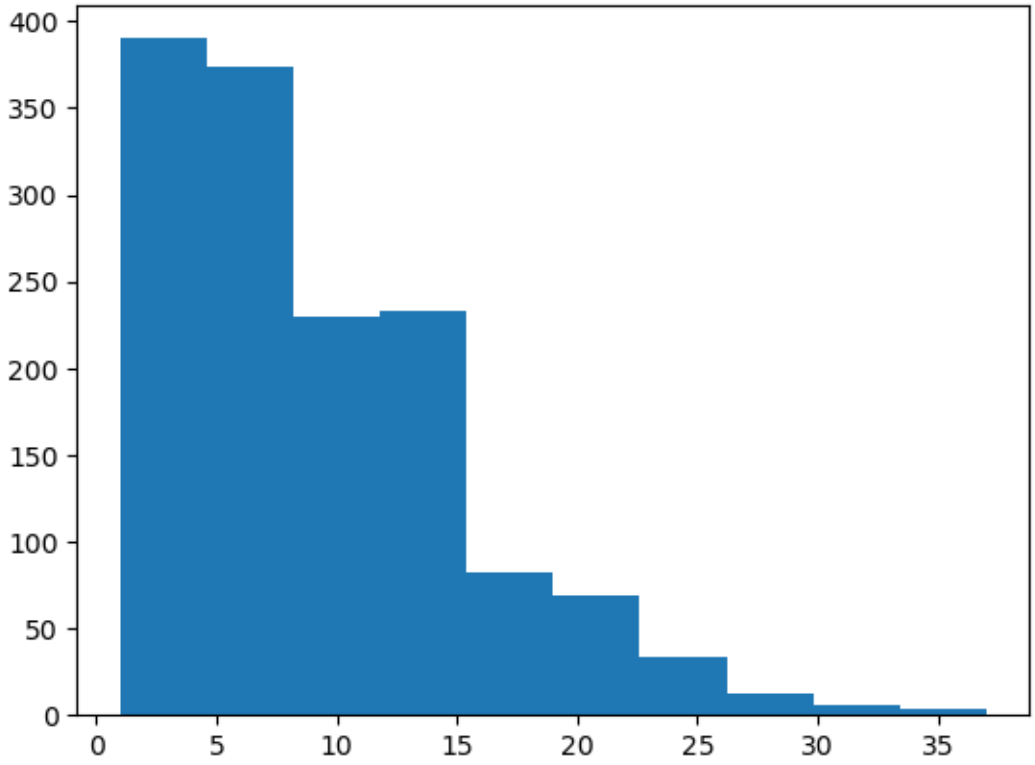
6. Ejecute un modelo Poisson para explicar cuando planea retirarse las personas que planean hacerlo. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: Nos quedamos solo con aquellos que desean retirarse (retin=1), y luego usamos el mismo set de datos de las preguntas anteriores. A diferencia de la probabilidad de querer retirarse, multiples factores tienen impacto en la edad esperada de retiro, algunos obvios como edad y caracteristicas laborales, y otros menos obvios como la zona urbana y tamano del hogar. Por ejemplo, la edad

esperada de retiro (en años) se reduce en 0.17 (años) si las personas viven en zona rural, y disminuye 0.05 anios por cada anio que el individuo es mayor (que el promedio).

[110]:
```
charls.reset_index(drop=True, inplace=True)
X=charls[charls['retired'] == 0].reset_index(drop=True)
X=charls[charls['retin'] != 0].reset_index(drop=True)
X=X[['retage','age', 'cesd', 'child', 'drinkly', 'female', 'hsize','married',␣
 ↪'hrsusu', 'schadj', 'urban','lwealth']]
X.dropna(inplace=True)
y=X['retage']
X=X[['age', 'cesd', 'child', 'drinkly', 'female', 'hsize','married', 'hrsusu',␣
 ↪'schadj', 'urban','lwealth']]
X=sm.add_constant(X)
plt.hist(y)
y.describe()
```

[110]:
```
count    1435.000000
mean        9.108711
std         6.325825
min         1.000000
25%         4.000000
50%         8.000000
75%        13.000000
max        37.000000
Name: retage, dtype: float64
```



[107]:
```
poisson=sm.GLM(y,X,family=sm.families.Poisson()).fit()
print(poisson.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                  retage   No. Observations:                 1435
Model:                             GLM   Df Residuals:                     1423
Model Family:                  Poisson   Df Model:                           11
Link Function:                     Log   Scale:                          1.0000
Method:                           IRLS   Log-Likelihood:                -4816.1
Date:                Mon, 24 Apr 2023   Deviance:                       4188.1
Time:                        13:50:19   Pearson chi2:                 4.22e+03
No. Iterations:                     5   Pseudo R-squ. (CS):             0.7425
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.9125      0.102     48.327      0.000       4.713       5.112
age           -0.0503      0.001    -36.603      0.000      -0.053      -0.048
cesd          -0.0029      0.002     -1.854      0.064      -0.006       0.000
child          0.0123      0.009      1.398      0.162      -0.005       0.030
drinkly        0.0528      0.020      2.626      0.009       0.013       0.092
female         0.0296      0.022      1.357      0.175      -0.013       0.072
hsize          0.0164      0.005      3.074      0.002       0.006       0.027
married       -0.0332      0.040     -0.830      0.407      -0.112       0.045
hrsusu         0.0186      0.007      2.610      0.009       0.005       0.032
schadj        -0.0028      0.003     -1.059      0.290      -0.008       0.002
urban         -0.1706      0.023     -7.576      0.000      -0.215      -0.126
lwealth       -0.0284      0.014     -2.082      0.037      -0.055      -0.002
==============================================================================
```

7. Determine sobre dispersion y posible valor optimo de alpha para un modelo Binomial Negativa.

R: En virtud de los resultados, podemos ver que existe cierta evidencia de sobredispersion (Pearson Chi2 sobre los Df residuos da un valor de 2.96). Al correr el test de sobredispersion vemos que el valor es estadisticamente distinto de 1, confirmando lo anterior.

[112]:
```
aux=((y-poisson.mu)**2-poisson.mu)/poisson.mu
auxr=sm.OLS(aux,poisson.mu).fit()
print(auxr.summary())
```

```
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:                  retage   R-squared (uncentered):
0.127
```

```
Model:                            OLS   Adj. R-squared (uncentered):
0.126
Method:                 Least Squares   F-statistic:
208.8
Date:                Mon, 24 Apr 2023   Prob (F-statistic):
2.76e-44
Time:                        14:01:56   Log-Likelihood:
-4375.3
No. Observations:                1435   AIC:
8753.
Df Residuals:                    1434   BIC:
8758.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.1995      0.014     14.450      0.000       0.172       0.227
==============================================================================
Omnibus:                     1335.800   Durbin-Watson:                   1.762
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            52368.658
Skew:                           4.355   Prob(JB):                         0.00
Kurtosis:                      31.284   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.
```

8. Usando la informacion anterior, ejecute un modelo Binomial Negativa para explicar el número de personas que hay dentro de un hogar. (*n_personas*). Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: El modelo de Binomial Negativa entrega resultados en general muy similares a Poisson, sin embargo hay diferencias significativas en algunas variables como (log) wealth, drinkly y hrsusu. Dados los resultados, se observa que las variables demograficas son aquellas que influencian la decision de edad de retiro.

[113]:
```
negbin=sm.GLM(y,X,family=sm.families.NegativeBinomial(alpha=0.2)).fit()
print(negbin.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                  retage   No. Observations:                 1435
Model:                             GLM   Df Residuals:                     1423
Model Family:        NegativeBinomial   Df Model:                           11
Link Function:                     Log   Scale:                          1.0000
```

Method:	IRLS	Log-Likelihood:	-4206.9
Date:	Mon, 24 Apr 2023	Deviance:	1590.6
Time:	14:02:13	Pearson chi2:	1.53e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3977
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.0289	0.173	29.065	0.000	4.690	5.368
age	-0.0513	0.002	-22.465	0.000	-0.056	-0.047
cesd	-0.0039	0.003	-1.467	0.142	-0.009	0.001
child	0.0042	0.014	0.289	0.773	-0.024	0.032
drinkly	0.0619	0.034	1.811	0.070	-0.005	0.129
female	0.0244	0.037	0.656	0.512	-0.048	0.097
hsize	0.0224	0.009	2.533	0.011	0.005	0.040
married	-0.0655	0.064	-1.028	0.304	-0.190	0.059
hrsusu	0.0152	0.012	1.268	0.205	-0.008	0.039
schadj	-0.0043	0.005	-0.949	0.343	-0.013	0.005
urban	-0.1643	0.038	-4.357	0.000	-0.238	-0.090
lwealth	-0.0324	0.024	-1.350	0.177	-0.079	0.015

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?

R: Dada la sobredispersion, los modelos Poisson y Binomial Negativa producen diferencias importantes, afectando la significancia de algunos parametros. En virtud de aquello, se favorece la Binomial Negativa, ya que el supuesto de media igual a a varianza no se cumpliria. En cualquier caso, las variables de edad, tamano de hogar y zona urbana son robustas a la especificacion (significativas en ambos modelos).