tarea1_entregable

April 24, 2024

TAREA 1 - IGNACIO GARRIDO URRA 2022058013

[273]:
```
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 linearmodels.panel as lmp
import sklearn
import scipy
from scipy.stats import nbinom
import seaborn as sns
from statsmodels.stats.diagnostic import het_breuschpagan
import scipy.stats as stats
%matplotlib inline
```

Cabe señalar desde un principio que, para todas las preguntas, el nivel de significatividad utilizado es el clasico, es decir, el de 0.05 o 5%

Preguntas:

1. Cargar la base de datos *disease.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 ajusta Disease_2 como variable numerica, a traves de un cambio con la variable original Disease para que nuestra variable dependiente sea dicotomica en donde 0 nos indica que no tiene ninguna enfermedad y 1 nos indica que al menos tiene una enfermedad. Por ultimo, las estadisticas descriptivas nos indican que existe una normalizacion de ciertas variables.

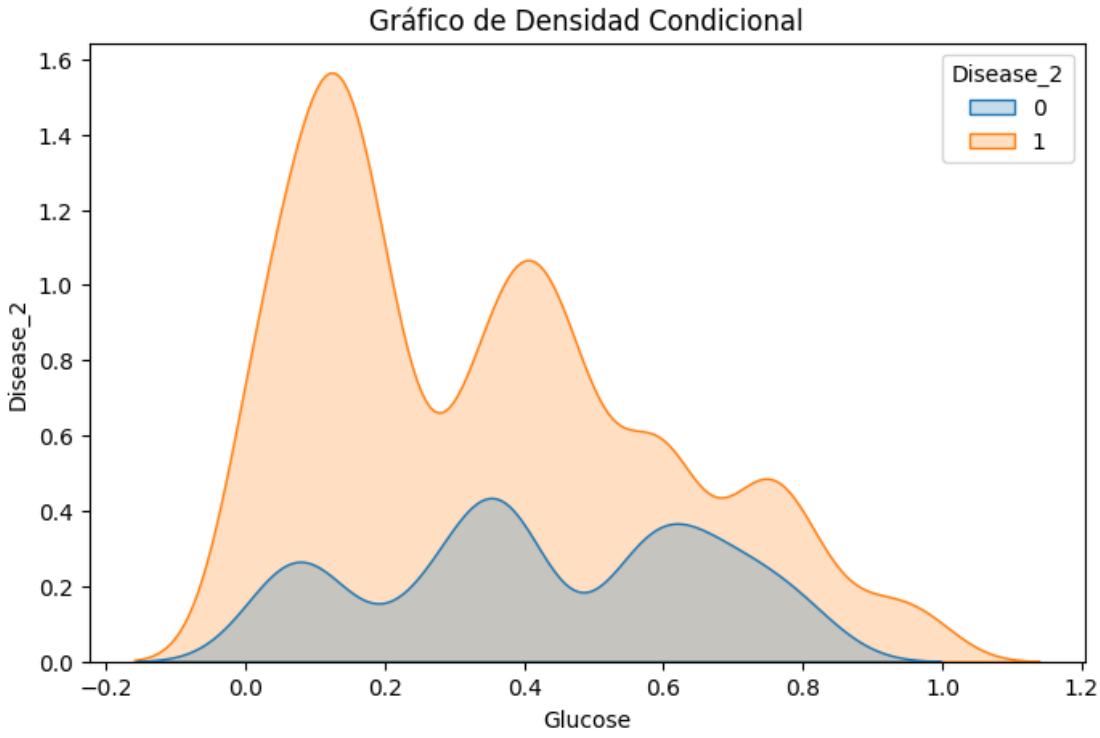
[274]:
```
df = pd.read_csv('disease.csv') #cargamos la BD;
df['Disease_2'] = df['Disease'].apply(lambda x: 1 if x > 0 else 0) #variable␣
↪Desease transformada a dicotomica;
df.dtypes #identificamos los tipos de datos;
df.reset_index(drop=True, inplace=True)
df = df.dropna() #limpiamos los datos faltantes;
df.describe
```

[274]:
```
<bound method NDFrame.describe of        Glucose  Cholesterol  Hemoglobin
Platelets  White Blood Cells  \
```

```
0     0.739597     0.650198    0.713631   0.868491           0.687433
1     0.377112     0.391959    0.577246   0.573482           0.685303
2     0.693767     0.730686    0.751196   0.747326           0.742084
3     0.377112     0.391959    0.577246   0.573482           0.685303
4     0.377112     0.391959    0.577246   0.573482           0.685303
...        ...          ...         ...        ...                ...
2346  0.107165     0.603341    0.791215   0.178840           0.718674
2347  0.353734     0.757757    0.755007   0.012594           0.227684
2348  0.353734     0.757757    0.755007   0.012594           0.227684
2349  0.107165     0.603341    0.791215   0.178840           0.718674
2350  0.353734     0.757757    0.755007   0.012594           0.227684

      Red Blood Cells  Hematocrit  Mean Corpuscular Volume  \
0            0.529895    0.290006                 0.631045
1            0.605134    0.472465                 0.098744
2            0.413056    0.820138                 0.140164
3            0.605134    0.472465                 0.098744
4            0.605134    0.472465                 0.098744
...               ...         ...                      ...
2346         0.825769    0.753657                 0.396669
2347         0.425117    0.387461                 0.461418
2348         0.425117    0.387461                 0.461418
2349         0.825769    0.753657                 0.396669
2350         0.425117    0.387461                 0.461418

      Mean Corpuscular Hemoglobin   Insulin  ...  Diastolic Blood Pressure  \
0                        0.001328  0.034129  ...                 0.071455
1                        0.721378  0.700015  ...                 0.445291
2                        0.756092  0.772896  ...                 0.488942
3                        0.721378  0.700015  ...                 0.445291
4                        0.721378  0.700015  ...                 0.445291
...                           ...       ...  ...                      ...
2346                     0.762667  0.232877  ...                 0.474378
2347                     0.305588  0.654441  ...                 0.212859
2348                     0.305588  0.654441  ...                 0.212859
2349                     0.762667  0.232877  ...                 0.474378
2350                     0.305588  0.654441  ...                 0.212859

      Triglycerides     HbA1c  LDL Cholesterol  HDL Cholesterol  Heart Rate  \
0          0.653472  0.502665         0.215560         0.512941    0.939485
1          0.694545  0.646206         0.657711         0.307132    0.599542
2          0.093734  0.755660         0.603351         0.381331    0.518567
3          0.694545  0.646206         0.657711         0.307132    0.599542
4          0.694545  0.646206         0.657711         0.307132    0.599542
...             ...       ...              ...              ...         ...
2346       0.731369  0.489514         0.102679         0.861035    0.979192
2347       0.393263  0.446854         0.729376         0.615543    0.612188
```

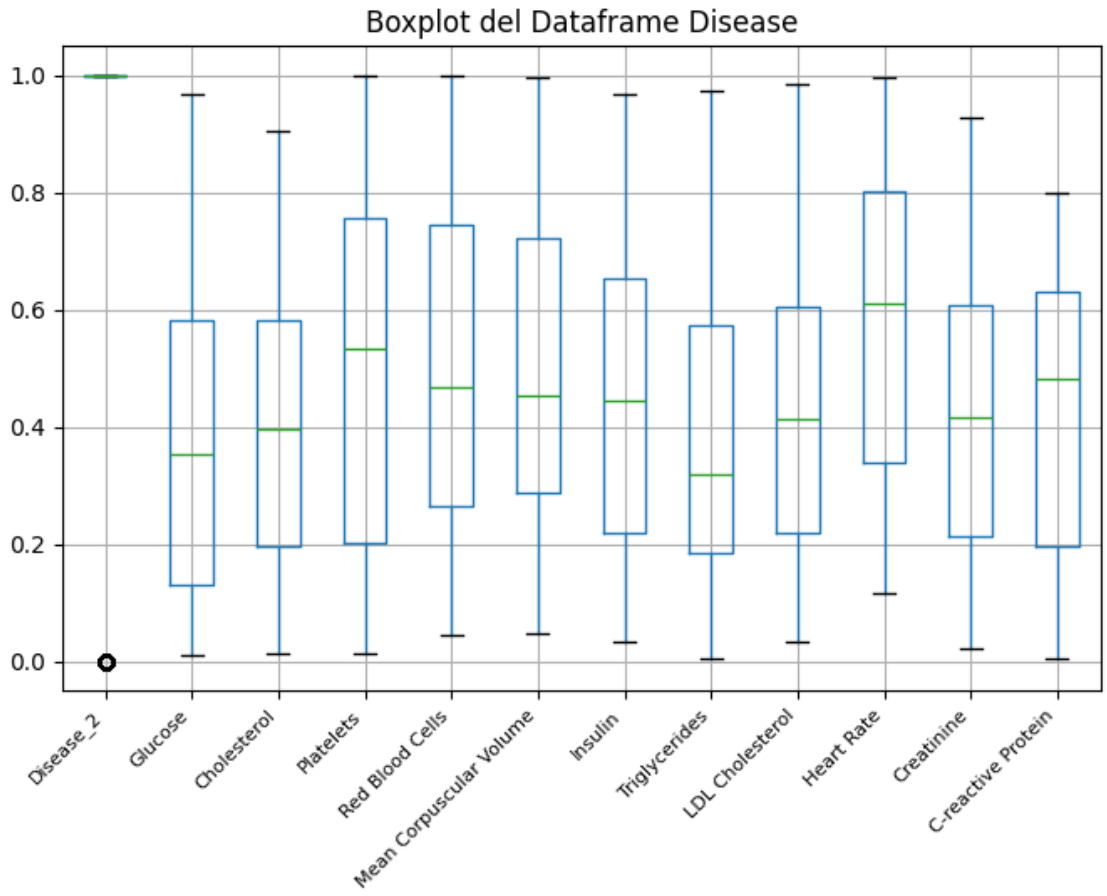
```
2348       0.393263  0.446854         0.729376         0.615543    0.612188
2349       0.731369  0.489514         0.102679         0.861035    0.979192
2350       0.393263  0.446854         0.729376         0.615543    0.612188

      Creatinine  C-reactive Protein  Disease  Disease_2
0       0.095512            0.769230        0          0
1       0.477714            0.607319        0          0
2       0.645247            0.751157        0          0
3       0.477714            0.607319        0          0
4       0.477714            0.607319        0          0
...          ...                 ...      ...        ...
2346    0.554960            0.621687        4          1
2347    0.407891            0.532100        4          1
2348    0.407891            0.532100        4          1
2349    0.554960            0.621687        4          1
2350    0.407891            0.532100        4          1

[2351 rows x 22 columns]>
```

[275]:
```
#Revision de distribuciones; En este caso variable explicativa 'Glucose' (como␣
↪ejemplo)
plt.figure(figsize=(8, 5))
sns.kdeplot(data=df, x='Glucose', hue='Disease_2', fill=True)
plt.xlabel('Glucose')
plt.ylabel('Disease_2')
plt.title('Gráfico de Densidad Condicional')
plt.show()
```



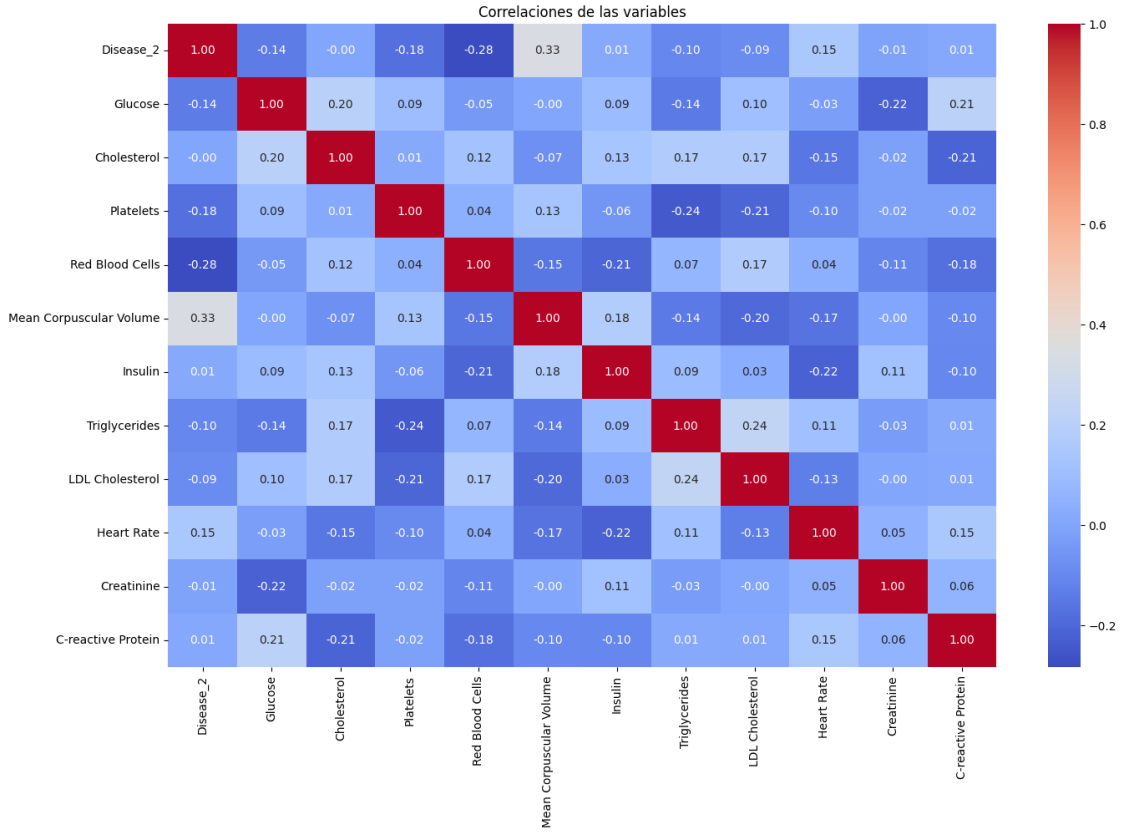
R: Previamente, en la revision de las distribuciones de los datos gracias al grafico de densidad condicional he decidido no incluir en el modelo lineal inicial las siguientes variables: 'Hemoglobin', 'White Blood Cells', 'Hematocrit', 'Mean corpuscular Hemoglobin', 'BMI', 'Systolic Blood Pressure', 'Diastolic Blood Pressure', 'HbA1c' y 'HDL Cholesterol'. Los principales argumentos son: Graficas muy similares, por ende, no hay mucha variacion en el comportamiento de ambas variables entre no tener enfermedades y tener al menos una con respecto a las variables anteriormente mencionadas.

[276]:
```
#Busqueda de Outliers;
X = ['Disease_2','Glucose', 'Cholesterol', 'Platelets', 'Red Blood Cells', 'Mean␣
 ↪Corpuscular Volume', 'Insulin', 'Triglycerides', 'LDL Cholesterol', 'Heart␣
 ↪Rate', 'Creatinine', 'C-reactive Protein']
plt.figure(figsize=(8, 5))
boxplot = df[X].boxplot()
boxplot.set_xticklabels(boxplot.get_xticklabels(), fontsize=8, rotation=45,␣
 ↪ha='right')
plt.title('Boxplot del Dataframe Disease')
plt.show()
```



R: No se encuentran valores extremos en las variables,por ende, no amerita clasificarlos como outliers.

[277]:
```
#Matriz de correlaciones
X = ['Disease_2','Glucose', 'Cholesterol', 'Platelets', 'Red Blood Cells', 'Mean␣
 ↪Corpuscular Volume', 'Insulin', 'Triglycerides', 'LDL Cholesterol', 'Heart␣
 ↪Rate', 'Creatinine', 'C-reactive Protein']
correlation_matrix = df[X].corr()
plt.figure(figsize=(16, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlaciones de las variables')
plt.show()
```



R: Como no se aprecia correlaciones altas presentes en el grafico, podemos hablar de que no existen señales de multicolinealidad, ya que no se visualizan correlaciones altas entre las variables explicativas del modelo.

2. Ejecute un modelo de probabilidad lineal (*MCO*) que permita explicar la probabilidad de que una persona tenga al menos una enfermedad. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.]

[278]:
```
#OLS - Modelo lineal inicial;
X_1 = df[['Glucose', 'Cholesterol', 'Platelets', 'Red Blood Cells', 'Mean␣
 ↪Corpuscular Volume', 'Insulin', 'Triglycerides', 'LDL Cholesterol', 'Heart␣
 ↪Rate', 'Creatinine', 'C-reactive Protein']]
y = df['Disease_2']; #Variable dependiente
X_1 = sm.add_constant(X_1)
model_1 = sm.OLS(y, X_1)
model1_1 = model_1.fit(cov_type='HC0')  # Aquí especifica el tipo de covarianza␣
 ↪HC que deseas utilizar
print(model1_1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Disease_2   R-squared:                       0.318
```

Model:	OLS	Adj. R-squared:	0.315
Method:	Least Squares	F-statistic:	137.9
Date:	Tue, 23 Apr 2024	Prob (F-statistic):	2.43e-244
Time:	21:11:14	Log-Likelihood:	-873.85
No. Observations:	2351	AIC:	1772.
Df Residuals:	2339	BIC:	1841.
Df Model:	11		
Covariance Type:	HC0		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.7805	0.043	18.196	0.000	0.696	0.865
Glucose	-0.3793	0.025	-15.038	0.000	-0.429	-0.330
Cholesterol	0.3205	0.032	10.017	0.000	0.258	0.383
Platelets	-0.2962	0.023	-12.661	0.000	-0.342	-0.250
Red Blood Cells	-0.4227	0.028	-14.834	0.000	-0.479	-0.367
Mean Corpuscular Volume	0.5679	0.026	22.216	0.000	0.518	0.618
Insulin	-0.0669	0.034	-1.947	0.051	-0.134	0.000
Triglycerides	-0.3060	0.033	-9.359	0.000	-0.370	-0.242
LDL Cholesterol	0.0813	0.028	2.937	0.003	0.027	0.136
Heart Rate	0.4011	0.034	11.739	0.000	0.334	0.468
Creatinine	-0.1975	0.033	-5.992	0.000	-0.262	-0.133
C-reactive Protein	0.0905	0.032	2.864	0.004	0.029	0.152

Omnibus:	194.266	Durbin-Watson:	0.616
Prob(Omnibus):	0.000	Jarque-Bera (JB):	202.130
Skew:	-0.672	Prob(JB):	1.28e-44
Kurtosis:	2.492	Cond. No.	14.6

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

Modelo Lineal Inicial.

En el modelo inicial, se puede visualizar que por el valor grande del estadistico F, el modelo es significativo, lo que quiere decir que existe al menos una variable que explica la variacion de la variable dependiente, que en este caso es 'Disease_2', lo que se interpreta que las variables en general explican si el individuo tiene al menos una enfermedad o no. En concreto, la variabilidad de tener al menos una enfermedad es explicada en un 31.8% por las variables del modelo. Ademas, todas las variables menos Insulina son significativas, por lo tanto, el modelo final en terminos de variables será todas menos la ya mencionada.

[279]:
```
#OLS - Modelo lineal inicial;
X = df[['Glucose', 'Cholesterol', 'Platelets', 'Red Blood Cells', 'Mean␣
 ↪Corpuscular Volume', 'Triglycerides', 'LDL Cholesterol', 'Heart Rate',␣
 ↪'Creatinine', 'C-reactive Protein']]
y = df['Disease_2']; #Variable dependiente
X = sm.add_constant(X)
model_2 = sm.OLS(y, X)
model1_2 = model_2.fit(cov_type='HC0')  # Aquí especifica el tipo de covarianza␣
 ↪HC que deseas utilizar
print(model1_2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Disease_2   R-squared:                       0.317
Model:                            OLS   Adj. R-squared:                  0.314
Method:                 Least Squares   F-statistic:                     143.7
Date:                Tue, 23 Apr 2024   Prob (F-statistic):          1.01e-234
Time:                        21:11:14   Log-Likelihood:                -875.94
No. Observations:                2351   AIC:                             1774.
Df Residuals:                    2340   BIC:                             1837.
Df Model:                          10
Covariance Type:                  HC0
==============================================================================
                                 coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
const                          0.7490      0.043     17.546      0.000       0.665
0.833
Glucose                       -0.3890      0.024    -16.131      0.000      -0.436
-0.342
Cholesterol                    0.3162      0.032     10.011      0.000       0.254
0.378
Platelets                     -0.2926      0.023    -12.613      0.000      -0.338
-0.247
Red Blood Cells               -0.4111      0.030    -13.797      0.000      -0.470
-0.353
Mean Corpuscular Volume        0.5598      0.024     22.943      0.000       0.512
```

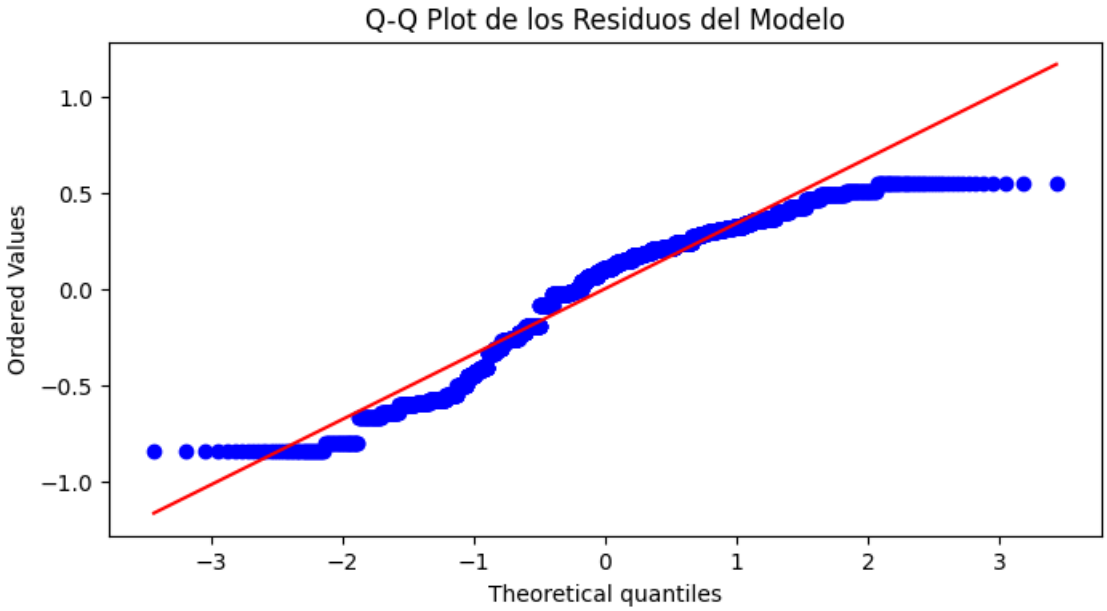
```
0.608
Triglycerides              -0.3152      0.031    -10.113      0.000      -0.376
-0.254
LDL Cholesterol             0.0819      0.027      2.981      0.003       0.028
0.136
Heart Rate                  0.4133      0.032     12.735      0.000       0.350
0.477
Creatinine                 -0.2079      0.033     -6.303      0.000      -0.273
-0.143
C-reactive Protein          0.0989      0.034      2.941      0.003       0.033
0.165
==============================================================================
Omnibus:                      198.459   Durbin-Watson:                   0.612
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              199.624
Skew:                          -0.661   Prob(JB):                     4.49e-44
Kurtosis:                       2.462   Cond. No.                         13.5
==============================================================================

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

Modelo Lineal Final

Al igual que en el modelo lineal inicial, el modelo sigue siendo significativo dado que no cambiamos ninguna de las variables explicativas. En ambos modelos, existen señales de autorrelacion positiva en los residuos de la regresion. En cuanto a la interpretacion de algunas de las variables explicativas: -La probabilidad de que la persona tenga al menos una enfermedad disminuye en un 38.9% cuando aumenta en una unidad el nivel de glucosa en la sangre. -La probabilidad de que la persona tenga al menos una enfermedad aumenta en un 31.6% cuando aumenta en una unidad el nivel de Colesterol en la sangre. -La probabilidad de que la persona tenga al menos una enfermedad disminuye en un 29.3% cuando aumenta en una unidad el nivel de Plaquetas en la sangre. -La probabilidad de que la persona tenga al menos una enfermedad disminuye en un 41.1% cuando aumenta en una unidad el nivel de Globulos rojos en la sangre. -La probabilidad de que la persona tenga al menos una enfermedad aumentan en un 56% cuando aumenta en una unidad el Volumen promedio de globulos rojos o Volumen corpuscular promedio. -La probabilidad de que la persona tenga al menos una enfermedad disminuye en un 31.5% cuando aumenta en una unidad el nivel de trigliceridos en la sangre. -La probabilidad de que la persona tenga al menos una enfermedad aumenta en un 8.2% cuando aumenta en una unidad el nivel de Colesterol LDL en la sangre. -La probabilidad de que la persona tenga al menos una enfermedad aumenta en un 41.3% cuando aumenta en una unidad la Frecuencia cardiaca. -La probabilidad de que la persona tenga al menos una enfermedad disminuye en un 20.8% cuando aumenta en una unidad el nivel de Creatinina en la sangre. -La probabilidad de que la persona tenga al menos una enfermedad aumenta en un 9.9% cuando aumenta en una unidad el nivel de Proteina C reactiva (PCR) en la sangre.

[280]:
```
residuos = model1_2.resid;
bp_test = het_breuschpagan(residuos, X);
etiquetas = ['Valor de LM', 'valor-p LM', 'Valor de F', 'valor-p F'];
plt.figure(figsize=(8, 4))
stats.probplot(model1_2.resid, dist="norm", plot=plt);
```

```
plt.title('Q-Q Plot de los Residuos del Modelo');
plt.show()
shapiro_test = stats.shapiro(model1_2.resid)
print(dict(zip(etiquetas, bp_test)));
print(f"Estadístico de Shapiro-Wilk: {shapiro_test[0]}")
print(f"Valor-p: {shapiro_test[1]}")
```




```
{'Valor de LM': 546.1029040515763, 'valor-p LM': 6.115235142350107e-111, 'Valor
de F': 70.80075636163608, 'valor-p F': 1.1195140691210342e-126}
Estadístico de Shapiro-Wilk: 0.9337704845562138
Valor-p: 4.089630405944323e-31
```

-Dado que ambos valores-p son significativamente menores que 0.05, hay evidencia fuerte contra la hipotesis nula de homocedasticidad (que sugiere que los errores tienen varianza constante). Es una mala señal para el modelo.
-Nuevamente valores-p menores a 0.05, nos sugiere que tendremos problemas. Nuestros residuos no siguen una distribucion normal, demostrando nuevamente que el modelo de regresion lineal no seria la mejor opcion para este dataset.

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.

[281]:
```
model2 = sm.Probit(y, X)
probit_model = model2.fit()
mfx = probit_model.get_margeff()
print(probit_model.summary())
print(mfx.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.336181
         Iterations 7
                          Probit Regression Results
==============================================================================
Dep. Variable:              Disease_2   No. Observations:                 2351
Model:                         Probit   Df Residuals:                     2340
Method:                           MLE   Df Model:                           10
Date:                Tue, 23 Apr 2024   Pseudo R-squ.:                  0.3854
Time:                        21:11:15   Log-Likelihood:                -790.36
converged:                       True   LL-Null:                       -1286.0
Covariance Type:            nonrobust   LLR p-value:                1.399e-206
==============================================================================
===========
                               coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-----------
const                        1.9235      0.251      7.674      0.000       1.432
2.415
Glucose                     -3.0307      0.242    -12.540      0.000      -3.504
-2.557
Cholesterol                  1.8433      0.203      9.087      0.000       1.446
2.241
Platelets                   -1.9154      0.155    -12.391      0.000      -2.218
-1.612
Red Blood Cells             -2.2899      0.161    -14.192      0.000      -2.606
-1.974
Mean Corpuscular Volume      3.2074      0.189     16.952      0.000       2.837
3.578
Triglycerides               -2.3805      0.185    -12.861      0.000      -2.743
-2.018
LDL Cholesterol              0.5791      0.174      3.335      0.001       0.239
0.919
Heart Rate                   2.0367      0.164     12.430      0.000       1.716
2.358
Creatinine                  -1.7553      0.179     -9.783      0.000      -2.107
-1.404
C-reactive Protein           1.0763      0.174      6.172      0.000       0.735
1.418
==============================================================================
===========
       Probit Marginal Effects
=====================================
Dep. Variable:              Disease_2
Method:                          dydx
At:                           overall
==============================================================================
```

```
==========
                             dy/dx    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
----------
Glucose                    -0.5659      0.041    -13.844      0.000      -0.646
-0.486
Cholesterol                 0.3442      0.036      9.559      0.000       0.274
0.415
Platelets                  -0.3577      0.025    -14.325      0.000      -0.407
-0.309
Red Blood Cells            -0.4276      0.026    -16.358      0.000      -0.479
-0.376
Mean Corpuscular Volume     0.5989      0.027     21.984      0.000       0.546
0.652
Triglycerides              -0.4445      0.030    -14.754      0.000      -0.504
-0.385
LDL Cholesterol             0.1081      0.032      3.353      0.001       0.045
0.171
Heart Rate                  0.3803      0.028     13.601      0.000       0.326
0.435
Creatinine                 -0.3278      0.031    -10.424      0.000      -0.389
-0.266
C-reactive Protein          0.2010      0.032      6.358      0.000       0.139
0.263
==============================================================================
==========
```

R: Se utilizaron las mismas variables que en la estiamacion por MCO. En este modelo, los valores de los coeficientes estimados son irrelevantes para el analisis, lo que si es importante conocer son las significancias individuales y los cambios marginales. Se puede apreciar que todas las variables son significativas para el modelo. En cuanto a los cambios marginales, el nivel de Colesterol y la Frecuencia cardiaca explican positivamente el hecho de que la persona tenga al menos una enfermedad. Por ejemplo, si la persona aumentara en una unidad su nivel de colesterol, aumenta en un 35.4% la probabilidad que te tenga al menos una enfermedad. En contraste, si la persona aumentara en una unidad su nivel de Trigliceridos, disminuye en un 38.8% la probabilidad de que tenga al menos una enfermedad. Por ultimo, si aumenta en una unidad su nivel de Volumen corpuscular medio, aumenta en un 67.5% la probabilidad de que tenga al menos una enfermedad. Se observa que el Pseudo Rsquared es superior al R squared de la estimacion por MCO, en concreto, de un 4.67%, por lo que es posible afirmar que el Modelo Probit es superior al modelo de Regresion por estimacion MCO.

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.

[282]:
```
#Probit - Modelo probit;
model3 = sm.Logit(y, X)
logit_model = model3.fit()
```

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

```
Optimization terminated successfully.
         Current function value: 0.337279
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:              Disease_2   No. Observations:                 2351
Model:                          Logit   Df Residuals:                     2340
Method:                           MLE   Df Model:                           10
Date:                Tue, 23 Apr 2024   Pseudo R-squ.:                 0.3834
Time:                        21:11:15   Log-Likelihood:                -792.94
converged:                       True   LL-Null:                       -1286.0
Covariance Type:            nonrobust   LLR p-value:                1.811e-205
==============================================================================
===========
                               coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-----------
const                        3.1667      0.446      7.102      0.000       2.293
4.041
Glucose                     -5.5587      0.456    -12.181      0.000      -6.453
-4.664
Cholesterol                  3.3656      0.361      9.319      0.000       2.658
4.073
Platelets                   -3.0715      0.266    -11.532      0.000      -3.594
-2.549
Red Blood Cells             -4.2035      0.312    -13.480      0.000      -4.815
-3.592
Mean Corpuscular Volume      5.5130      0.343     16.077      0.000       4.841
6.185
Triglycerides               -4.0518      0.325    -12.477      0.000      -4.688
-3.415
LDL Cholesterol              1.1484      0.312      3.675      0.000       0.536
1.761
Heart Rate                   3.9138      0.320     12.223      0.000       3.286
4.541
Creatinine                  -3.1253      0.319     -9.786      0.000      -3.751
-2.499
C-reactive Protein           1.7914      0.313      5.718      0.000       1.177
2.405
==============================================================================
===========
        Logit Marginal Effects
=====================================
```

```
Dep. Variable:              Disease_2
Method:                         dydx
At:                          overall
================================================================================
===========
                              dy/dx    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
-----------
Glucose                     -0.5967      0.044    -13.606      0.000      -0.683
-0.511
Cholesterol                  0.3613      0.036     10.036      0.000       0.291
0.432
Platelets                   -0.3297      0.025    -12.959      0.000      -0.380
-0.280
Red Blood Cells             -0.4512      0.028    -15.834      0.000      -0.507
-0.395
Mean Corpuscular Volume      0.5918      0.028     20.943      0.000       0.536
0.647
Triglycerides               -0.4349      0.030    -14.383      0.000      -0.494
-0.376
LDL Cholesterol              0.1233      0.033      3.709      0.000       0.058
0.188
Heart Rate                   0.4201      0.030     13.915      0.000       0.361
0.479
Creatinine                  -0.3355      0.032    -10.518      0.000      -0.398
-0.273
C-reactive Protein           0.1923      0.033      5.863      0.000       0.128
0.257
================================================================================
===========
```

R: Se utilizaron las mismas variables que en la estiamacion por MCO. En este modelo, los valores de los coeficientes estimados son irrelevantes para el analisis, lo que si es importante conocer son las significancias individuales y los cambios marginales. Se puede apreciar que la variable Insulina no es significativa para el modelo. En cuanto a los cambios marginales, el nivel de Colesterol y la Frecuencia cardiaca explican positivamente el hecho de que la persona tenga al menos una enfermedad. Por ejemplo, si la persona aumentara en una unidad su nivel de colesterol, aumenta en un 37.2% la probabilidad que te tenga al menos una enfermedad. En contraste, si la persona aumentara en una unidad su nivel de Trigliceridos, disminuye en un 37.4% la probabilidad de que tenga al menos una enfermedad. Por ultimo, si aumenta en una unidad su nivel de Volumen corpuscular medio, aumenta en un 67.1% la probabilidad de que tenga al menos una enfermedad. Se observa que el Pseudo Rsquared es superior al R squared de la estimacion por MCO, en concreto, de un 4.73%, por lo que es posible afirmar que el Modelo Logit es superior al modelo de Regresion por estimacion MCO y superior al Modelo Probit.

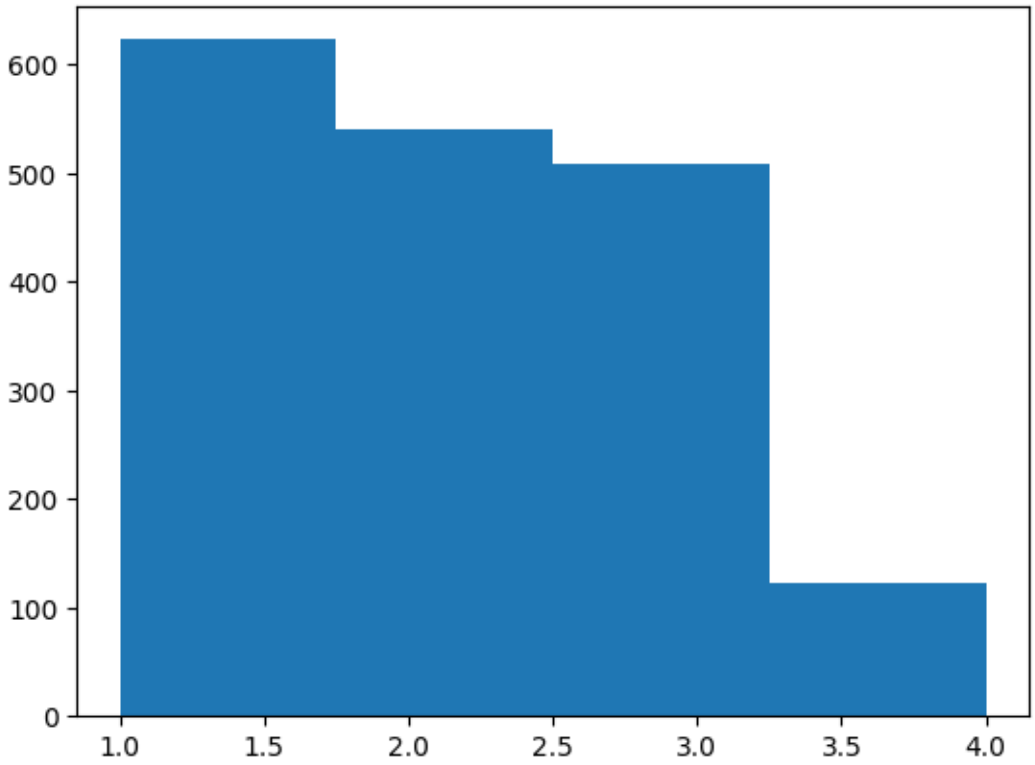
5. Comente los resultados obtenidos en 2, 3 y 4. ¿Cuáles y por qué existen las diferencias entre los resultados?. En su opinión, ¿Cuál sería el más adecuado para responder la pregunta de investgación y por qué? ¿Qué variables resultaron ser robustas a la especificación?

R: Se sabe que la estimacion MCO es debil al querer estimar variables binarias, ya que graficamente en gran parte de la recta de regresion no estan incluidos los valores de 0 y 1 y ademas se asume que los errores se distribuyen noramlmente y tienen una varianza constante. Esta suposicion no se cumple cuando se trabaja con variables binarias, ya que los erorres tienen una distribucion Bernoulli o Binomial. Por estas razones, la estimacion por MCO no es la mejor tecnica para modelar variables binarias. En su lugar, utilizamos otras tecnicas de modelado estadistico especificas para variables binarias, como la regresion Logit y la regresion Probit. En este caso, al comparar las bondades de ajuste de los modelos se llega a la conclusion de que el mejor modelo para este caso particular es el Logit, lo cual tiene sentido, ya que la regresion Probit se utiliza mas en campos relacionados con las ciencias sociales, en cambio, Logit se utiliza mas en campos de ciencias medicas y biologicas. En todos los casos, las variables que resultaron significativas son: 'Glucose', 'Cholesterol', 'Platelets', 'Red Blood Cells', 'Mean Corpuscular Volume', 'Triglycerides', 'LDL Cholesterol', 'Heart Rate', 'Creatinine' y 'C-reactive Protein'.

6. Ejecute un modelo Poisson para explicar el numero de enfermedades que tiene una persona. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

[283]:
```
df.reset_index(drop=True, inplace=True)
subset=df.loc[df['Disease']>0]
plt.hist(subset.Disease, bins=4)
```

[283]:
```
(array([623., 540., 509., 123.]),
 array([1.  , 1.75, 2.5 , 3.25, 4.  ]),
 <BarContainer object of 4 artists>)
```



[284]:
```
y = df['Disease']; #Variable dependiente
X = df[['Glucose', 'Cholesterol', 'Platelets', 'Red Blood Cells', 'Mean␣
 ↪Corpuscular Volume', 'Triglycerides', 'LDL Cholesterol', 'Heart Rate',␣
 ↪'Creatinine', 'C-reactive Protein']]
poisson=sm.GLM(y,X,family=sm.families.Poisson()).fit()
print(poisson.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                Disease   No. Observations:                 2351
Model:                            GLM   Df Residuals:                     2341
Model Family:                 Poisson   Df Model:                            9
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3458.7
Date:                Tue, 23 Apr 2024   Deviance:                       2335.5
Time:                        21:11:15   Pearson chi2:                 1.96e+03
No. Iterations:                     5   Pseudo R-squ. (CS):             0.1781
Covariance Type:            nonrobust
==============================================================================
===========
                                 coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-----------
Glucose                     -0.2084      0.071     -2.935      0.003      -0.348
-0.069
Cholesterol                  0.7220      0.074      9.798      0.000       0.578
0.866
Platelets                   -0.8911      0.054    -16.383      0.000      -0.998
-0.785
Red Blood Cells              0.0385      0.062      0.624      0.533      -0.082
0.159
Mean Corpuscular Volume      0.8114      0.054     14.998      0.000       0.705
0.917
Triglycerides               -0.4243      0.070     -6.104      0.000      -0.561
-0.288
LDL Cholesterol             -0.1363      0.067     -2.038      0.042      -0.267
-0.005
Heart Rate                   0.4475      0.062      7.243      0.000       0.326
0.569
Creatinine                   0.0337      0.070      0.482      0.629      -0.103
0.171
C-reactive Protein           0.3513      0.069      5.061      0.000       0.215
0.487
==============================================================================
===========
```

-Dado que las variables Creatinina y Globulos Rojos no son significativas en este modelo inicial, se procederá a quitarlas y a comparar con el modelo final de Poisson. El pseudo R-squared indica que el numero de enfermedades que tiene una persona es explicada en un 17.8% por las variables explicativas del modelo.

[291]:
```
y = df['Disease']; #Variable dependiente
X_3 = df[['Glucose', 'Cholesterol', 'Platelets','Mean Corpuscular Volume',␣
 ↪'Triglycerides', 'LDL Cholesterol', 'Heart Rate', 'C-reactive Protein']]
poisson=sm.GLM(y,X_3,family=sm.families.Poisson()).fit()
print(poisson.summary())
print("fitted lambda")
print(poisson.mu)
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                Disease   No. Observations:                 2351
Model:                            GLM   Df Residuals:                     2343
Model Family:                 Poisson   Df Model:                            7
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3459.0
Date:                Tue, 23 Apr 2024   Deviance:                       2336.1
Time:                        21:33:52   Pearson chi2:                 1.97e+03
No. Iterations:                     5   Pseudo R-squ. (CS):             0.1779
Covariance Type:            nonrobust
==============================================================================
===========
                                 coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-----------
Glucose                     -0.2158      0.070     -3.105      0.002      -0.352
-0.080
Cholesterol                  0.7406      0.070     10.582      0.000       0.603
0.878
Platelets                   -0.8838      0.054    -16.455      0.000      -0.989
-0.779
Mean Corpuscular Volume      0.8201      0.053     15.482      0.000       0.716
0.924
Triglycerides               -0.4218      0.069     -6.100      0.000      -0.557
-0.286
LDL Cholesterol             -0.1210      0.064     -1.884      0.060      -0.247
0.005
Heart Rate                   0.4643      0.058      7.986      0.000       0.350
0.578
C-reactive Protein           0.3569      0.068      5.283      0.000       0.224
0.489
==============================================================================
===========
```

```
fitted lambda
[1.61755219 0.9099733  1.2740728  ... 2.92168165 2.57687652 2.92168165]
```

Modelo Final Poisson

Se puede visualizar que el valor del Pseudo R-Squared es practicamente el mismo que en el modelo de Poisson inicial, por lo cual es posible afirmar que el numero de enfermedades que tiene una persona es explicada de igual manera por las variables que ya quedaron seleccionadas. -El aumento en el nivel de glucosa en la sangre implica en una disminucion en 21.6% en el numero de enfermedades que tiene una persona. -El aumento en el nivel de colesterol en la sangre implica en un aumento en 74.1% en el numero de enfermedades que tiene una persona. -El aumento en el nivel de plaquetas en la sangre implica en una disminucion en 88.4% en el numero de enfermedades que tiene una persona. -El aumento en el nivel de Volumen corpuscular medio implica en un aumento en 82% en el numero de enfermedades que tiene una persona. -El aumento en el nivel de Trigliceridos en la sangre implica en una disminucion en 42.2% en el numero de enfermedades que tiene una persona. -El aumento en el nivel de Colesterol LDL en la sangre implica en una disminucion en 12.1% en el numero de enfermedades que tiene una persona. -El aumento en el nivel de Frecuencia Cardiaca implica en un aumento en 46.4% en el numero de enfermedades que tiene una persona. -El aumento en el nivel de Proteina C-reactiva en la sangre implica en un aumento en 35.7% en el numero de enfermedades que tiene una persona.

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

[290]:
```
aux=((y-poisson.mu)**2-poisson.mu)/poisson.mu
auxr=sm.OLS(aux,poisson.mu).fit()
print(auxr.summary())
```

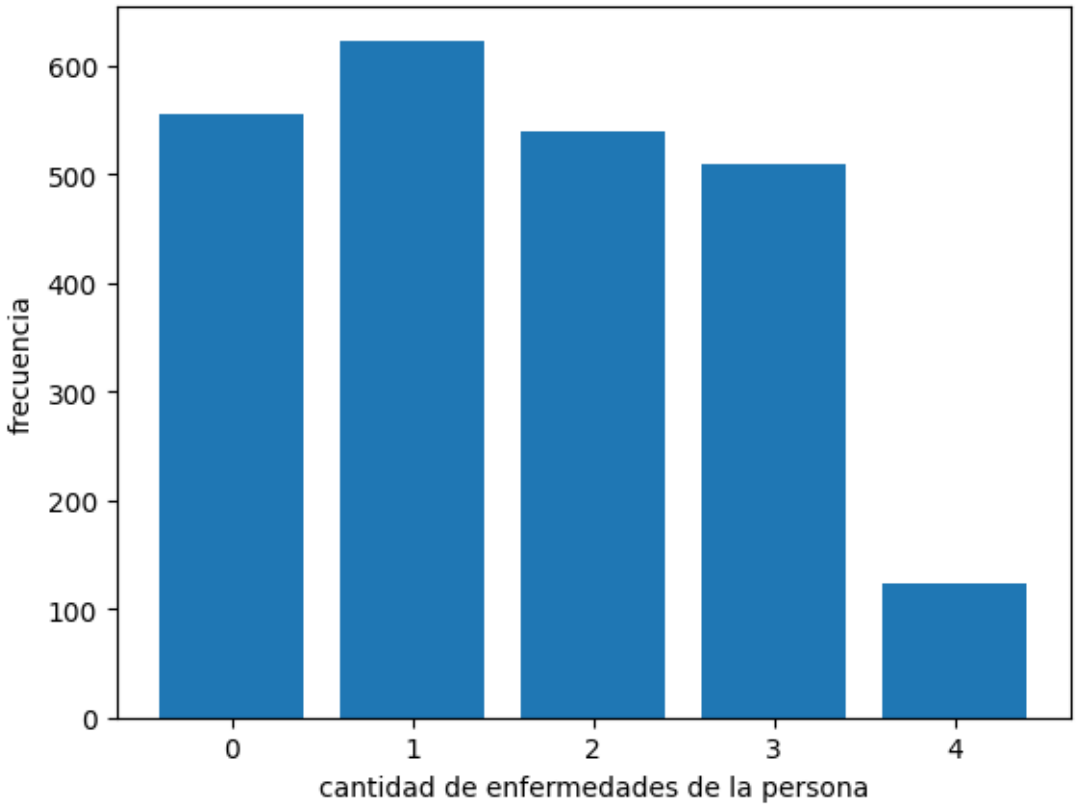
```
                                 OLS Regression Results
==============================================================================
=======
Dep. Variable:                 Disease   R-squared (uncentered):
0.065
Model:                             OLS   Adj. R-squared (uncentered):
0.064
Method:                  Least Squares   F-statistic:
162.8
Date:                Tue, 23 Apr 2024   Prob (F-statistic):
4.28e-36
Time:                        21:15:49   Log-Likelihood:
-3176.7
No. Observations:                2351   AIC:
6355.
Df Residuals:                    2350   BIC:
6361.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
```

```
x1            -0.1468      0.012    -12.759      0.000      -0.169      -0.124
==============================================================================
Omnibus:                     1047.193   Durbin-Watson:                   1.193
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5450.076
Skew:                           2.095   Prob(JB):                         0.00
Kurtosis:                       9.172   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.
```

R: En virtud de los resultados, podemos ver que existe cierta evidencia de sobredispersion (Pearson Chi2 sobre los DF residuos da un valor de 1.97). Ademas, al correr el test de sobredispersion vemos que el valor es estadisticamente distinto de 1, confirmando lo anterior. Esto significa que la media es diferente a la varianza, en este caso la media es mayor que la varianza.

8. Usando la informacion anterior, ejecute un modelo Binomial Negativa para responder la pregunta 6. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

[287]:
```
hola= df['Disease']
freq = hola.value_counts()
freq= freq.sort_index()
media= df['Disease'].mean()
varianza = df['Disease'].var()
plt.bar(freq.index, freq.values)
plt.xlabel('cantidad de enfermedades de la persona')
plt.ylabel('frecuencia')
plt.show()
print(media, varianza)
```




```
1.5831561037856232 1.463614034770175
```

[293]:
```
y=df['Disease']
X = df[['Glucose', 'Cholesterol', 'Platelets', 'Red Blood Cells', 'Mean␣
 ↪Corpuscular Volume', 'Triglycerides', 'LDL Cholesterol', 'Heart Rate',␣
 ↪'Creatinine', 'C-reactive Protein']]
negbin=sm.GLM(y,X,family=sm.families.NegativeBinomial()).fit()
print(negbin.summary())
print("fitted lambda")
print(negbin.mu)
```

```
/Users/igarridourra/Library/Python/3.9/lib/python/site-
packages/statsmodels/genmod/families/family.py:1367: ValueWarning: Negative
binomial dispersion parameter alpha not set. Using default value alpha=1.0.
  warnings.warn("Negative binomial dispersion parameter alpha not "
```

Generalized Linear Model Regression Results

Dep. Variable:	Disease	No. Observations:	2351
Model:	GLM	Df Residuals:	2341
Model Family:	NegativeBinomial	Df Model:	9
Link Function:	Log	Scale:	1.0000

```
Method:                           IRLS   Log-Likelihood:                -3957.3
Date:                Tue, 23 Apr 2024   Deviance:                       1219.6
Time:                        21:35:13   Pearson chi2:                     849.
No. Iterations:                     9   Pseudo R-squ. (CS):            0.07849
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Glucose                    -0.3024      0.117     -2.575      0.010      -0.533
-0.072
Cholesterol                 0.5879      0.120      4.889      0.000       0.352
0.824
Platelets                  -0.9264      0.089    -10.420      0.000      -1.101
-0.752
Red Blood Cells            -0.1141      0.101     -1.134      0.257      -0.311
0.083
Mean Corpuscular Volume     0.9558      0.089     10.790      0.000       0.782
1.129
Triglycerides              -0.5289      0.113     -4.680      0.000      -0.750
-0.307
LDL Cholesterol             0.0337      0.111      0.303      0.762      -0.184
0.251
Heart Rate                  0.5524      0.101      5.452      0.000       0.354
0.751
Creatinine                  0.0060      0.114      0.052      0.958      -0.218
0.230
C-reactive Protein          0.4166      0.114      3.647      0.000       0.193
0.640
==============================================================================
fitted lambda
[1.48938259 0.86271081 1.20719023 ... 2.99878401 2.36643945 2.99878401]
```

R: El modelo de Binomial Negativa entrega resultados en general muy similares a Poisson, sin embardo, hay diferencias significativas en algunas variables como (log) Colesterol LDL. Por ejemplo, en el caso de una relacion positiva, un aumento en una unidad en el nivel de Colesterol en la sangre implicara en un aumento en la cantidad de enfermedades que tenga la persona en un 58.8%.

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?

[295]:
```
#BN final
y=df['Disease']
```

```
X = df[['Glucose', 'Cholesterol', 'Platelets','Mean Corpuscular Volume',␣
 ↪'Triglycerides', 'Heart Rate', 'C-reactive Protein']]
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:                Disease   No. Observations:                 2351
Model:                            GLM   Df Residuals:                     2344
Model Family:        NegativeBinomial   Df Model:                            6
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3957.9
Date:                Tue, 23 Apr 2024   Deviance:                       1220.8
Time:                        21:54:39   Pearson chi2:                     841.
No. Iterations:                     8   Pseudo R-squ. (CS):            0.07803
Covariance Type:            nonrobust
==============================================================================
==========
                             coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
----------
Glucose                    -0.2873      0.113     -2.544      0.011      -0.509
-0.066
Cholesterol                 0.5626      0.112      5.044      0.000       0.344
0.781
Platelets                  -0.9383      0.086    -10.857      0.000      -1.108
-0.769
Mean Corpuscular Volume     0.9480      0.086     10.980      0.000       0.779
1.117
Triglycerides              -0.5283      0.108     -4.902      0.000      -0.740
-0.317
Heart Rate                  0.5088      0.094      5.397      0.000       0.324
0.694
C-reactive Protein          0.4178      0.110      3.788      0.000       0.202
0.634
==============================================================================
==========
fitted lambda
[1.47805983 0.86890449 1.18712082 ... 2.93372543 2.43123637 2.93372543]

/Users/igarridourra/Library/Python/3.9/lib/python/site-
packages/statsmodels/genmod/families/family.py:1367: ValueWarning: Negative
binomial dispersion parameter alpha not set. Using default value alpha=1.0.
  warnings.warn("Negative binomial dispersion parameter alpha not "
```

R: Claramente, debido a que hay sobre dispersion y una gran proporcion de ceros, es posible

afirmar que es mejor ocupar la distribucion normal negativa, la cuyal dice que las variables 'Red Blood Cells' (numero de globulos rojos en la sangre), 'LDL Cholesterol' (nivel de colesterol LDL en la sangre) y 'Creatinine' (Nivel de creatinina en la sangre) son no significativas a diferencia del modelo Poisson que solo consideraba no significativa a las variables 'Creatinine' y 'Red Blood Cells'. Tras lo dicho anteriormente, solo se consideraran los resultados de la Distribucion binomial negativa, los cuales dicen que las variables 'Glucose', 'Cholesterol', 'Platelets', 'Mean Corpuscular Volume', 'Tryglicerides', 'Heart Rate' y 'C-Reactive Protein'; son las que explican la cantidad de enfermedades que pueden tener una persona. El pseudo R-squared sugiere que la cantidad de enfermedades es verdaderamente explicado (ya que Binomial negativa se considera mas apropiado) en un 7.8% por las variables antes mencionadas.