# nonlinear

April 3, 2025

Section 2: non linear models

### 0.0.1 Housekeeping

**import libraries** Pandas for data management, statsmodels, numpy and sklearn for analysis, matplotlib for visualization. Other libraries as needed for specific tasks (e.g. semopy for SEM). Remember to use the bash terminal or the enviroment manager to add libraries.

**read data** Read data files using pandas as noted below. We can clean and organize data in many ways (for example, using the **dropna** command over a dataset).

**describe data** There are many was to analyze data and do descriptive statistics. A good command to start is ***head***, to describe a section of the data.

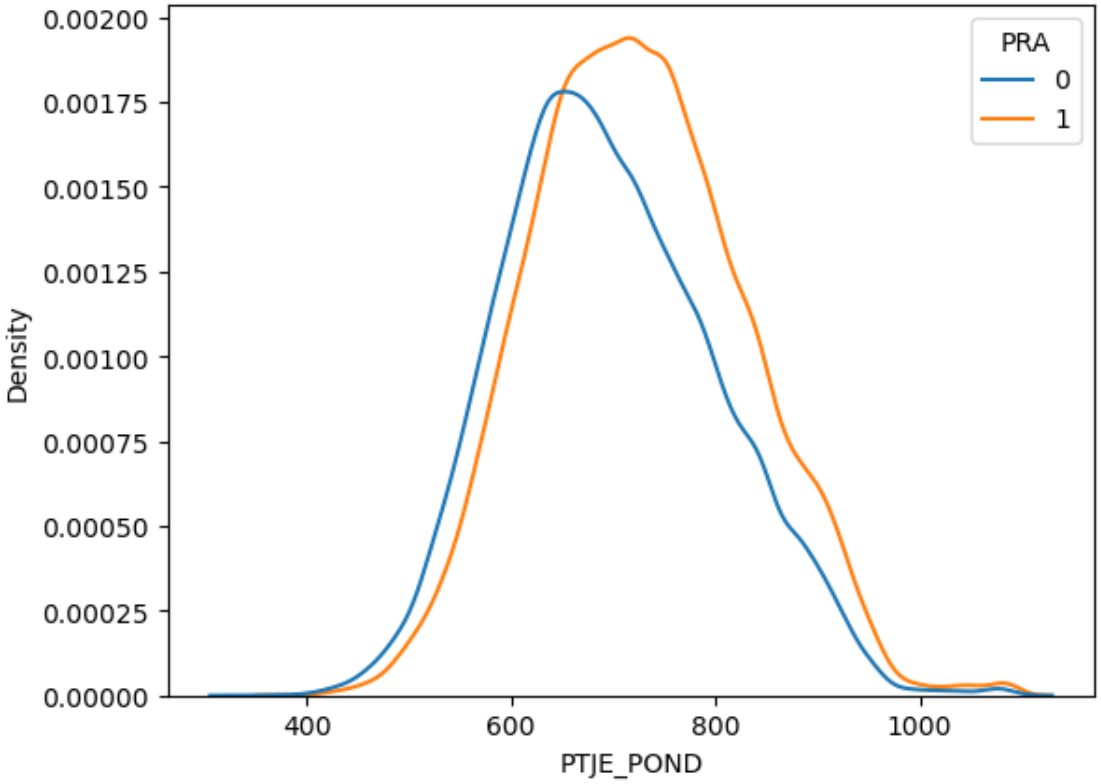
```
ID_aux             object
CODIGO_UNIV         int64
CODIGO              int64
VIA                 int64
PREFERENCIA         int64
PTJE_POND         float64
TIPO_MATRICULA      int64
TIPO               object
PRA                 int32
dtype: object
```

```
         CODIGO_UNIV         CODIGO            VIA    PREFERENCIA  \
count  115353.000000  115353.000000  115353.000000  115353.000000
mean       33.082963   33207.054580       1.057770       2.296169
std        15.086797   15139.994206       0.354068       2.328242
min        11.000000   11001.000000       1.000000       1.000000
25%        17.000000   17077.000000       1.000000       1.000000
50%        36.000000   36031.000000       1.000000       1.000000
75%        47.000000   47006.000000       1.000000       3.000000
max        58.000000   58202.000000       4.000000      20.000000

           PTJE_POND  TIPO_MATRICULA            PRA
count  115353.000000   115353.000000  115353.000000
mean      711.946851        1.200584       0.533354
```

```
std       105.284423       1.162438       0.498888
min       340.700000       1.000000       0.000000
25%       635.850000       1.000000       0.000000
50%       705.500000       1.000000       1.000000
75%       783.400000       1.000000       1.000000
max      1095.000000      11.000000       1.000000
```

```
c:\Users\juanc\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
c:\Users\juanc\anaconda3\lib\site-packages\seaborn\_oldcore.py:1075:
FutureWarning: When grouping with a length-1 list-like, you will need to pass a
length-1 tuple to get_group in a future version of pandas. Pass `(name,)`
instead of `name` to silence this warning.
  data_subset = grouped_data.get_group(pd_key)
```

```
<Axes: xlabel='PTJE_POND', ylabel='Density'>
```




### 0.0.2 OLS

We can use statsmodels to estimate a simple OLS regression (linear probability model).

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | PRA | R-squared: | 0.019 |
| Model: | OLS | Adj. R-squared: | 0.019 |
| Method: | Least Squares | F-statistic: | 312.1 |
| Date: | Sat, 22 Mar 2025 | Prob (F-statistic): | 0.00 |
| Time: | 17:40:50 | Log-Likelihood: | -82383. |
| No. Observations: | 115353 | AIC: | 1.648e+05 |
| Df Residuals: | 115345 | BIC: | 1.649e+05 |
| Df Model: | 7 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 0.0607 | 0.010 | 5.823 | 0.000 | 0.040 | 0.081 |
| C(TIPO, Treatment(reference='REGULAR'))[T.+MC] | 0.0639 | 0.114 | 0.561 | 0.575 | -0.159 | 0.287 |
| C(TIPO, Treatment(reference='REGULAR'))[T.BEA] | 0.4862 | 0.152 | 3.189 | 0.001 | 0.187 | 0.785 |
| C(TIPO, Treatment(reference='REGULAR'))[T.PACE] | 0.3113 | 0.142 | 2.193 | 0.028 | 0.033 | 0.590 |
| PTJE_POND | 0.0007 | 1.46e-05 | 45.652 | 0.000 | 0.001 | 0.001 |
| PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.+MC] | -0.0001 | 0.000 | -0.721 | 0.471 | -0.000 | 0.000 |
| PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.BEA] | -0.0007 | 0.000 | -3.312 | 0.001 | -0.001 | -0.000 |
| PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.PACE] | -0.0004 | 0.000 | -2.608 | 0.009 | -0.001 | -9.56e-05 |

| | | | |
|---|---|---|---|
| Omnibus: | 415919.828 | Durbin-Watson: | 1.981 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 17834.320 |
| Skew: | -0.131 | Prob(JB): | 0.00 |
| Kurtosis: | 1.092 | Cond. No. | 7.54e+04 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.54e+04. This might indicate that there are strong multicollinearity or other numerical problems.

### 0.0.3 Probit

```
Optimization terminated successfully.
         Current function value: 0.681557
         Iterations 4
                          Probit Regression Results
==============================================================================
Dep. Variable:                    PRA   No. Observations:               115353
Model:                         Probit   Df Residuals:                   115345
Method:                           MLE   Df Model:                            7
Date:                Sat, 22 Mar 2025   Pseudo R-squ.:                 0.01355
Time:                        17:42:48   Log-Likelihood:                -78620.
converged:                       True   LL-Null:                       -79700.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
==========================================
                                                                    coef    std err
z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------------
Intercept                                                        -1.1180      0.027
-41.710      0.000      -1.170      -1.065
C(TIPO, Treatment(reference='REGULAR'))[T.+MC]                    0.1691      0.291
0.581      0.562      -0.402       0.740
C(TIPO, Treatment(reference='REGULAR'))[T.BEA]                    1.2359      0.387
3.192      0.001       0.477       1.995
C(TIPO, Treatment(reference='REGULAR'))[T.PACE]                   0.7600      0.368
2.065      0.039       0.039       1.482
PTJE_POND                                                        0.0017   3.75e-05
45.154      0.000       0.002       0.002
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.+MC]        -0.0003      0.000
-0.739      0.460      -0.001       0.001
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.BEA]        -0.0017      0.001
-3.318      0.001      -0.003      -0.001
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.PACE]       -0.0009      0.000
-2.459      0.014      -0.002      -0.000
==============================================================================
==========================================
       Probit Marginal Effects
=====================================
Dep. Variable:                    PRA
Method:                          dydx
At:                           overall
==============================================================================
==========================================
                                                                  dy/dx    std err
z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
```

```
-------------------------------------------
C(TIPO, Treatment(reference='REGULAR'))[T.+MC]                0.0663      0.114
0.581      0.562      -0.157       0.290
C(TIPO, Treatment(reference='REGULAR'))[T.BEA]                0.4841      0.152
3.192      0.001       0.187       0.781
C(TIPO, Treatment(reference='REGULAR'))[T.PACE]               0.2977      0.144
2.065      0.039       0.015       0.580
PTJE_POND                                                     0.0007   1.43e-05
46.441      0.000       0.001       0.001
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.+MC]     -0.0001      0.000
-0.739      0.460      -0.000       0.000
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.BEA]     -0.0007      0.000
-3.319      0.001      -0.001      -0.000
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.PACE]    -0.0004      0.000
-2.460      0.014      -0.001    -7.5e-05
==============================================================================
==============================================
```

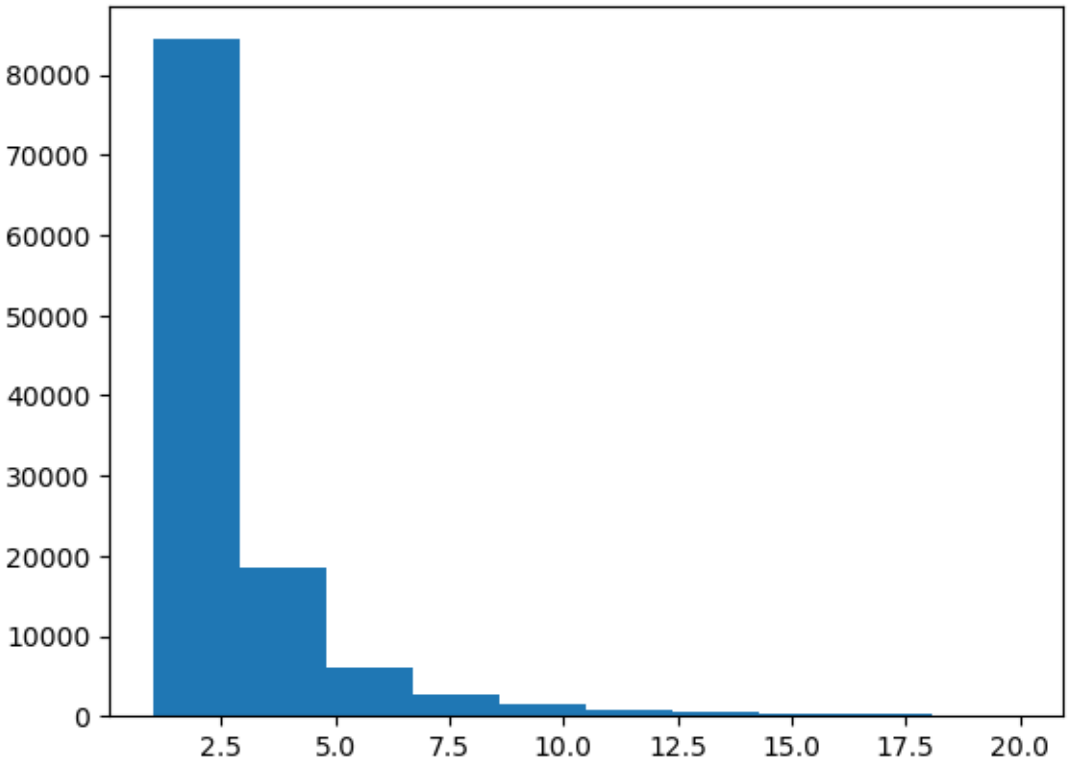
### 0.0.4 Logit

```
Optimization terminated successfully.
         Current function value: 0.681547
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:                    PRA   No. Observations:               115353
Model:                          Logit   Df Residuals:                   115345
Method:                           MLE   Df Model:                            7
Date:                Sat, 22 Mar 2025   Pseudo R-squ.:                 0.01357
Time:                        17:43:45   Log-Likelihood:                -78618.
converged:                       True   LL-Null:                       -79700.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
==============================================
                                                                 coef    std err
z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
-------------------------------------------
Intercept                                                    -1.7940      0.043
-41.510      0.000      -1.879      -1.709
C(TIPO, Treatment(reference='REGULAR'))[T.+MC]                0.2786      0.467
0.596      0.551      -0.637       1.194
C(TIPO, Treatment(reference='REGULAR'))[T.BEA]                1.9823      0.619
3.205      0.001       0.770       3.195
C(TIPO, Treatment(reference='REGULAR'))[T.PACE]               1.2104      0.596
2.031      0.042       0.042       2.378
PTJE_POND                                                     0.0027   6.06e-05
44.870      0.000       0.003       0.003
```

```
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.+MC]     -0.0005      0.001
-0.755      0.450      -0.002       0.001
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.BEA]     -0.0028      0.001
-3.334      0.001      -0.004      -0.001
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.PACE]    -0.0015      0.001
-2.420      0.016      -0.003      -0.000
==============================================================================
==========================================
        Logit Marginal Effects
=====================================
Dep. Variable:                    PRA
Method:                          dydx
At:                           overall
==============================================================================
==========================================
                                                                  dy/dx    std err
z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------------
C(TIPO, Treatment(reference='REGULAR'))[T.+MC]                0.0680      0.114
0.596      0.551      -0.156       0.292
C(TIPO, Treatment(reference='REGULAR'))[T.BEA]                0.4842      0.151
3.205      0.001       0.188       0.780
C(TIPO, Treatment(reference='REGULAR'))[T.PACE]               0.2957      0.146
2.031      0.042       0.010       0.581
PTJE_POND                                                     0.0007   1.43e-05
46.489      0.000       0.001       0.001
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.+MC]     -0.0001      0.000
-0.755      0.450      -0.000       0.000
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.BEA]     -0.0007      0.000
-3.334      0.001      -0.001      -0.000
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.PACE]    -0.0004      0.000
-2.420      0.016      -0.001   -6.98e-05
==============================================================================
==========================================
```

### 0.0.5 Poisson

```
(array([84375., 18629.,  6147.,  2695.,  1394.,   869.,   531.,   341.,
          230.,   142.]),
 array([ 1. ,  2.9,  4.8,  6.7,  8.6, 10.5, 12.4, 14.3, 16.2, 18.1, 20. ]),
 <BarContainer object of 10 artists>)
```




```
Optimization terminated successfully.
         Current function value: 2.007541
         Iterations 8
                          Poisson Regression Results
==============================================================================
Dep. Variable:            PREFERENCIA   No. Observations:               115353
Model:                        Poisson   Df Residuals:                   115345
Method:                           MLE   Df Model:                            7
Date:                Sat, 22 Mar 2025   Pseudo R-squ.:                0.008066
Time:                        17:48:50   Log-Likelihood:            -2.3158e+05
converged:                       True   LL-Null:                   -2.3346e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
================================================
                                                                  coef    std err
z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------------
Intercept                                                       1.6584      0.014
120.126      0.000       1.631       1.685
C(TIPO, Treatment(reference='REGULAR'))[T.+MC]                 -0.0161      0.151
-0.107      0.915      -0.311       0.279
```

```
C(TIPO, Treatment(reference='REGULAR'))[T.BEA]                -0.6286      0.207
-3.031      0.002      -1.035      -0.222
C(TIPO, Treatment(reference='REGULAR'))[T.PACE]               0.0597      0.201
0.297      0.767      -0.335       0.454
PTJE_POND                                                   -0.0012   1.96e-05
-59.831      0.000      -0.001      -0.001
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.+MC]  -7.534e-05      0.000
-0.326      0.744      -0.001       0.000
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.BEA]      0.0009      0.000
3.037      0.002       0.000       0.001
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.PACE]  6.448e-05      0.000
0.306      0.760      -0.000       0.000
==============================================================================
==============================================
```

```
0         2.282521
1         2.273032
2         1.878068
3         2.468069
4         2.481428
            …
118354    2.065103
118355    2.543190
118356    2.189514
118357    2.359874
118358    2.683776
Length: 115353, dtype: float64
```

### 0.0.6 Negative Binomial

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 1.887567
         Iterations: 35
         Function evaluations: 48
         Gradient evaluations: 48
```

```
c:\Users\juanc\anaconda3\lib\site-packages\statsmodels\base\model.py:607:
ConvergenceWarning: Maximum Likelihood optimization failed to converge. Check
mle_retvals
  warnings.warn("Maximum Likelihood optimization failed to "
```

```
                     NegativeBinomial Regression Results
==============================================================================
Dep. Variable:            PREFERENCIA   No. Observations:               115353
Model:               NegativeBinomial   Df Residuals:                   115345
Method:                           MLE   Df Model:                            7
Date:                Sat, 22 Mar 2025   Pseudo R-squ.:                0.005100
Time:                        17:55:37   Log-Likelihood:           -2.1774e+05
converged:                      False   LL-Null:                  -2.1885e+05
```

```
Covariance Type:             nonrobust   LLR p-value:                      0.000
================================================================================
================================================
                                                                    coef    std err
z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
------------------------------------------------
Intercept                                                          1.6322      0.017
93.685      0.000       1.598       1.666
C(TIPO, Treatment(reference='REGULAR'))[T.+MC]                     0.0223      0.194
0.115      0.909      -0.358       0.402
C(TIPO, Treatment(reference='REGULAR'))[T.BEA]                    -0.6145      0.260
-2.360      0.018      -1.125      -0.104
C(TIPO, Treatment(reference='REGULAR'))[T.PACE]                    0.0757      0.251
0.301      0.763      -0.416       0.567
PTJE_POND                                                         -0.0011   2.46e-05
-46.218      0.000      -0.001      -0.001
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.+MC]          -0.0001      0.000
-0.444      0.657      -0.001       0.000
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.BEA]           0.0008      0.000
2.362      0.018       0.000       0.002
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.PACE]      3.787e-05      0.000
0.144      0.885      -0.000       0.001
alpha                                                              0.2722      0.003
107.984      0.000       0.267       0.277
================================================================================
================================================

0         2.283147
1         2.273955
2         1.890228
3         2.462640
4         2.475546
           …
118354    2.072231
118355    2.535186
118356    2.193003
118357    2.358029
118358    2.670772
Length: 115353, dtype: float64

 5.340407628269691
```

### 0.0.7 Test overdispersion

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.

The Negative Binomial model estimated above is using a value of $\theta$ (or $\alpha = 1/\theta$) equal to 1. In

order to determine the appropiate value of $\alpha$, you can estimate a simple regression using the output of the Poisson model:

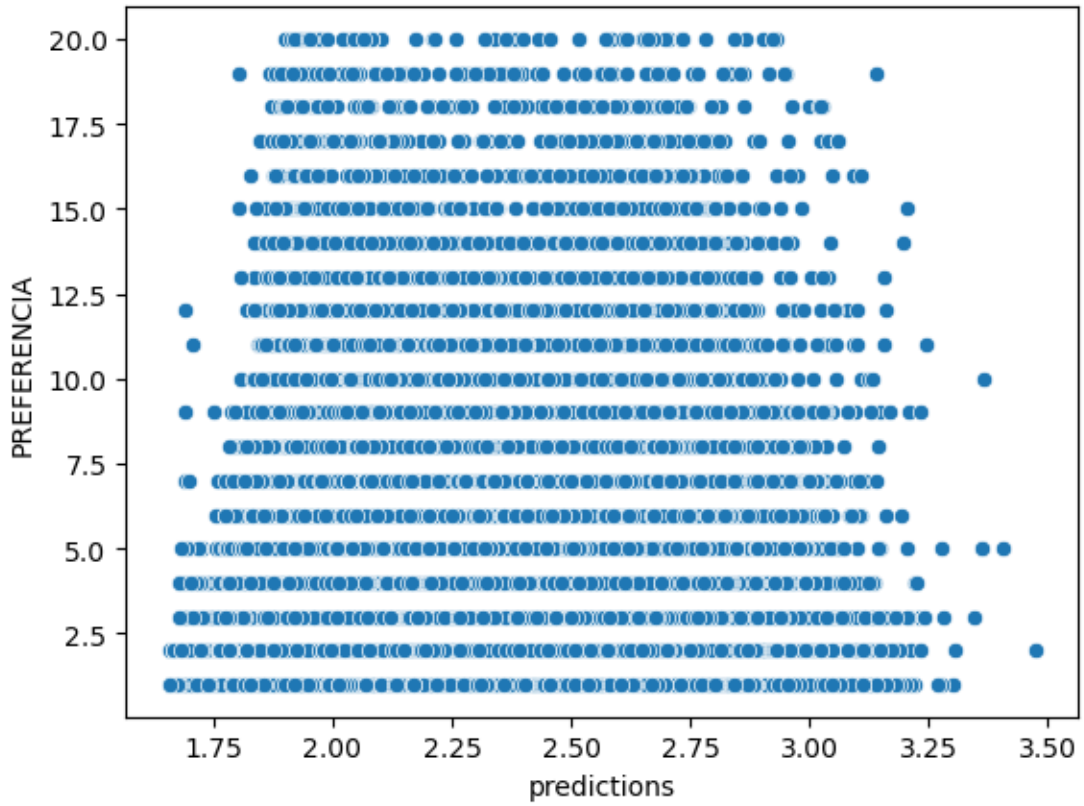
1. Construct the following variable aux=$[(y-\lambda)^2-\lambda]/\lambda$
2. Regress the variable aux with $\lambda$ as the only explanatory variable (no constant)
3. The estimated value is an appropiate guess for $\ln \alpha$

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.

```
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:                      y   R-squared (uncentered):
0.273
Model:                            OLS   Adj. R-squared (uncentered):
0.273
Method:                 Least Squares   F-statistic:
4.329e+04
Date:                Sat, 22 Mar 2025   Prob (F-statistic):
0.00
Time:                        18:09:57   Log-Likelihood:
-1.4129e+05
No. Observations:              115353   AIC:
2.826e+05
Df Residuals:                  115352   BIC:
2.826e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.2183      0.001    208.055      0.000       0.216       0.220
==============================================================================
Omnibus:                   456346.254   Durbin-Watson:                   1.963
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            15901.008
Skew:                           0.172   Prob(JB):                         0.00
Kurtosis:                       1.214   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.

c:\Users\juanc\anaconda3\lib\site-packages\statsmodels\base\optimizer.py:18:
FutureWarning: Keyword arguments have been passed to the optimizer that have no
effect. The list of allowed keyword arguments for method bfgs is: gtol, norm,
epsilon. The list of unsupported keyword arguments passed include: alpha. After
```

```
release 0.14, this will raise.
  warnings.warn(

Warning: Maximum number of iterations has been exceeded.
         Current function value: 1.887567
         Iterations: 35
         Function evaluations: 48
         Gradient evaluations: 48

c:\Users\juanc\anaconda3\lib\site-packages\statsmodels\base\model.py:607:
ConvergenceWarning: Maximum Likelihood optimization failed to converge. Check
mle_retvals
  warnings.warn("Maximum Likelihood optimization failed to "

                     NegativeBinomial Regression Results
==============================================================================
Dep. Variable:            PREFERENCIA   No. Observations:               115353
Model:               NegativeBinomial   Df Residuals:                   115345
Method:                          MLE   Df Model:                            7
Date:                Sat, 22 Mar 2025   Pseudo R-squ.:                0.005100
Time:                        18:11:04   Log-Likelihood:            -2.1774e+05
converged:                      False   LL-Null:                   -2.1885e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
==========================================
                                                                  coef    std err
z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
------------------------------------------
Intercept                                                       1.6322      0.017
93.685      0.000       1.598       1.666
C(TIPO, Treatment(reference='REGULAR'))[T.+MC]                  0.0223      0.194
0.115      0.909      -0.358       0.402
C(TIPO, Treatment(reference='REGULAR'))[T.BEA]                 -0.6145      0.260
-2.360      0.018      -1.125      -0.104
C(TIPO, Treatment(reference='REGULAR'))[T.PACE]                 0.0757      0.251
0.301      0.763      -0.416       0.567
PTJE_POND                                                      -0.0011   2.46e-05
-46.218      0.000      -0.001      -0.001
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.+MC]       -0.0001      0.000
-0.444      0.657      -0.001       0.000
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.BEA]        0.0008      0.000
2.362      0.018       0.000       0.002
PTJE_POND:C(TIPO, Treatment(reference='REGULAR'))[T.PACE]   3.787e-05      0.000
0.144      0.885      -0.000       0.001
alpha                                                           0.2722      0.003
107.984      0.000       0.267       0.277
==============================================================================
==========================================
```

```
<Axes: xlabel='predictions', ylabel='PREFERENCIA'>
```




**Tarea 1**

*Instrucciones*

Su notebook con las respuestas a la tarea se deben entregar a mas tardar el dia 21/04/25 hasta las 21:00, subiendolo al repositorio en la carpeta tareas/2025.

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 **.csv**.

Las variables tienen la siguiente descripcion:

Preguntas (todas tienen el mismo puntaje):

1.

**Tarea 1 2024 (Pauta)**

*Instrucciones*

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 **disease.csv**.

Las variables tienen la siguiente descripcion:

Glucose: This is the level of glucose in the blood, measured in milligrams per deciliter (mg/dL)

Cholesterol: This is the level of cholesterol in the blood, measured in milligrams per deciliter (mg/dL)

Hemoglobin: This is the protein in red blood cells that carries oxygen from the lungs to the rest of the body

Platelets: Platelets are blood cells that help with clotting

White Blood Cells (WBC): These are cells of the immune system that help fight infections

Red Blood Cells (RBC): These are the cells that carry oxygen from the lungs to the rest of the body

Hematocrit: This is the percentage of blood volume that is occupied by red blood cells

Mean Corpuscular Volume (MCV): This is the average volume of red blood cells

Mean Corpuscular Hemoglobin (MCH): This is the average amount of hemoglobin in a red blood cell

Insulin: This is a hormone that helps regulate blood sugar levels

BMI (Body Mass Index): This is a measure of body fat based on height and weight

Systolic Blood Pressure (SBP): This is the pressure in the arteries when the heart beats

Diastolic Blood Pressure (DBP): This is the pressure in the arteries when the heart is at rest between beats

Triglycerides: These are a type of fat found in the blood, measured in milligrams per deciliter (mg/dL)

HbA1c (Glycated Hemoglobin): This is a measure of average blood sugar levels over the past two to three months

LDL (Low-Density Lipoprotein) Cholesterol: This is the "bad" cholesterol that can build up in the arteries

HDL (High-Density Lipoprotein) Cholesterol: This is the "good" cholesterol that helps remove LDL cholesterol from the arteries

Heart Rate: This is the number of heartbeats per minute (bpm)

Creatinine: This is a waste product produced by muscles and filtered out of the blood by the kidneys

C-reactive Protein (CRP): This is a marker of inflammation in the body

Disease: This indicates the number of diseases (0 indicates healthy)

Preguntas:

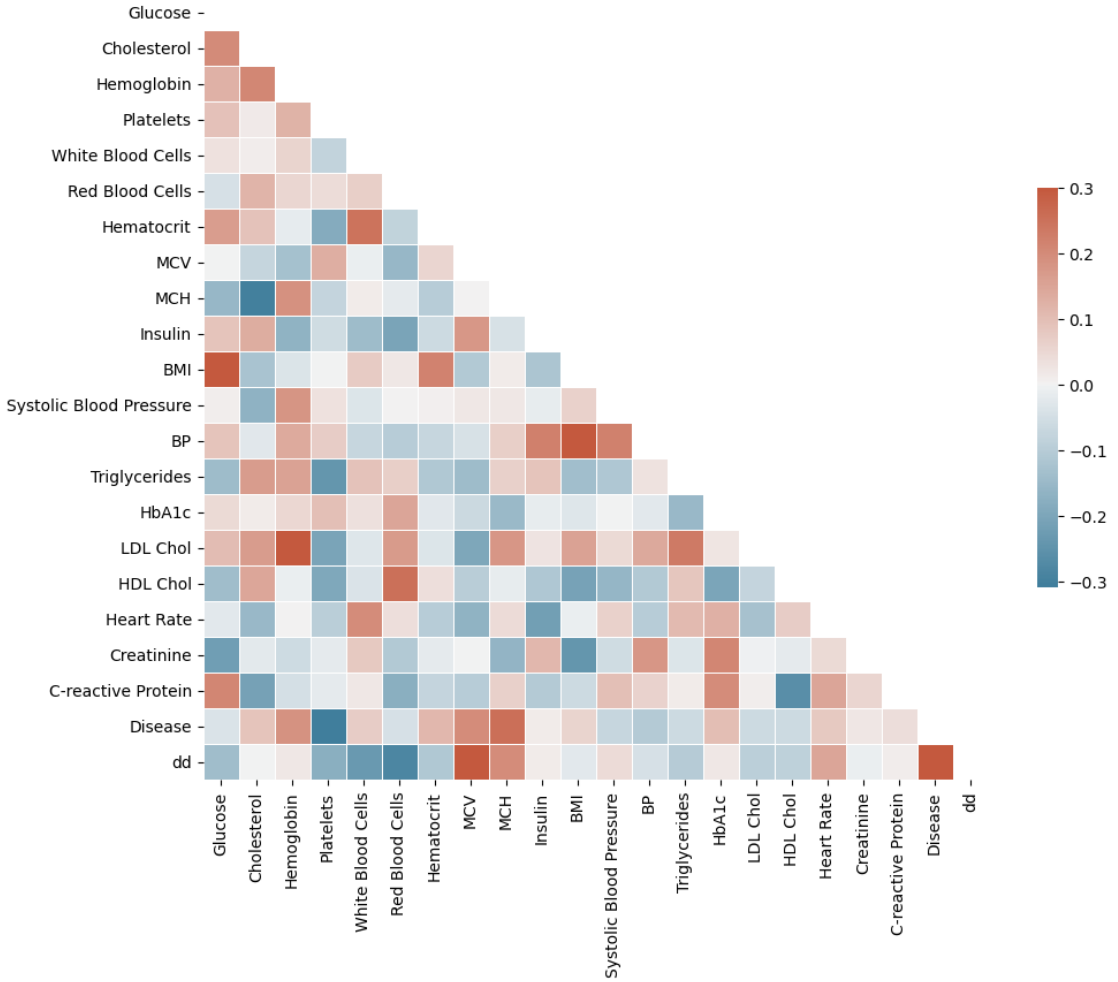
1. Cargar la base de datos *disease.csv* en el ambiente. Identifique los tipos de datos que se encuentran en la base, realice estadisticas descriptivas sobre las variables importantes (Hint:

Revisar la distribuciones, datos faltantes, outliers, etc.) y limpie las variables cuando sea necesario.

**R**: En este caso particular solo era necesario generar una variable binaria a partir de *Disease*, mientras que la matriz de correlaciones permite analizar que variables podrian excluirse del modelo.

```
           Glucose  Cholesterol   Hemoglobin    Platelets  White Blood Cells  \
count  2351.000000  2351.000000  2351.000000  2351.000000        2351.000000
mean      0.362828     0.393648     0.586190     0.504027           0.511086
std       0.251889     0.239449     0.271498     0.303347           0.277270
min       0.010994     0.012139     0.003021     0.012594           0.010139
25%       0.129198     0.195818     0.346092     0.200865           0.259467
50%       0.351722     0.397083     0.609836     0.533962           0.527381
75%       0.582278     0.582178     0.791215     0.754841           0.743164
max       0.968460     0.905026     0.983306     0.999393           0.990786

       Red Blood Cells   Hematocrit          MCV          MCH      Insulin  \
count      2351.000000  2351.000000  2351.000000  2351.000000  2351.000000
mean          0.506590     0.507152     0.492200     0.484459     0.447062
std           0.266565     0.285537     0.275735     0.315618     0.242861
min           0.044565     0.011772     0.046942     0.000554     0.034129
25%           0.263589     0.288132     0.287532     0.207938     0.219111
50%           0.467431     0.493428     0.453052     0.420723     0.444806
75%           0.743670     0.753657     0.722293     0.778160     0.654441
max           1.000000     0.977520     0.995263     0.963235     0.966784

       …           BP  Triglycerides        HbA1c     LDL Chol     HDL Chol  \
count  …  2351.000000    2351.000000  2351.000000  2351.000000  2351.000000
mean   …     0.421708       0.374373     0.439112     0.421777     0.546079
std    …     0.248768       0.256981     0.263779     0.252124     0.269511
min    …     0.005579       0.005217     0.016256     0.033037     0.039505
25%    …     0.175469       0.184604     0.188750     0.217757     0.307132
50%    …     0.474378       0.317857     0.466375     0.413071     0.512941
75%    …     0.663382       0.572330     0.652514     0.604753     0.779378
max    …     0.934617       0.973679     0.950218     0.983826     0.989411

        Heart Rate   Creatinine  C-reactive Protein      Disease           dd
count  2351.000000  2351.000000         2351.000000  2351.000000  2351.000000
mean      0.582255     0.425075            0.430308     1.583156     0.763505
std       0.250915     0.229298            0.243034     1.209799     0.425020
min       0.114550     0.021239            0.004867     0.000000     0.000000
25%       0.339125     0.213026            0.196192     1.000000     1.000000
50%       0.610860     0.417295            0.481601     1.000000     1.000000
75%       0.800666     0.606719            0.631426     3.000000     1.000000
max       0.996873     0.925924            0.797906     4.000000     1.000000

[8 rows x 22 columns]
```

```
<Axes: >
```




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.

**R**: Segun el modelo, excluyendo las variables que no contribuyen dada alta correlacion, el BMI y ritmo cardiaco se asocian positivamente de forma importante a la probabilidad de tener alguna morbilidad, asi como el colesterol total y ciertas caracteristicas de la sangre (MCV, MCH). De la misma manera, otros marcadores bioquimicos indican una asociacion negativa, tales como la glucosa, los trigliceridos y la insulina. Los resultados se interpretan como cambio en puntos porcentuales por cambio de una unidad en la variable (escala de 0 a 100). Por ejemplo, un cambio de una unidad en colesterol total indica un cambio de 0.5 puntos porcentuales en la probabilidad de tener al menos una enfermedad.

```
                            OLS Regression Results
==============================================================================
```

| | | | |
|---|---|---|---|
| Dep. Variable: | dd | R-squared: | 0.487 |
| Model: | OLS | Adj. R-squared: | 0.483 |
| Method: | Least Squares | F-statistic: | 236.2 |
| Date: | Thu, 02 May 2024 | Prob (F-statistic): | 0.00 |
| Time: | 14:59:51 | Log-Likelihood: | -539.42 |
| No. Observations: | 2351 | AIC: | 1117. |
| Df Residuals: | 2332 | BIC: | 1226. |
| Df Model: | 18 | | |
| Covariance Type: | HC0 | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.7573 | 0.044 | 17.136 | 0.000 | 0.671 | 0.844 |
| Glucose | -0.0036 | 0.000 | -13.196 | 0.000 | -0.004 | -0.003 |
| Cholesterol | 0.0052 | 0.000 | 17.836 | 0.000 | 0.005 | 0.006 |
| Hemoglobin | 0.0020 | 0.000 | 7.071 | 0.000 | 0.001 | 0.003 |
| Platelets | -0.0044 | 0.000 | -20.314 | 0.000 | -0.005 | -0.004 |
| White Blood Cells | -0.0043 | 0.000 | -20.011 | 0.000 | -0.005 | -0.004 |
| Red Blood Cells | -0.0037 | 0.000 | -15.924 | 0.000 | -0.004 | -0.003 |
| Hematocrit | -0.0023 | 0.000 | -8.921 | 0.000 | -0.003 | -0.002 |
| MCV | 0.0063 | 0.000 | 24.897 | 0.000 | 0.006 | 0.007 |
| MCH | 0.0029 | 0.000 | 13.634 | 0.000 | 0.003 | 0.003 |
| Insulin | -0.0010 | 0.000 | -3.317 | 0.001 | -0.002 | -0.000 |
| BMI | 0.0034 | 0.000 | 10.476 | 0.000 | 0.003 | 0.004 |
| BP | -0.0013 | 0.000 | -4.715 | 0.000 | -0.002 | -0.001 |
| Triglycerides | -0.0026 | 0.000 | -8.228 | 0.000 | -0.003 | -0.002 |
| HbA1c | 0.0012 | 0.000 | 5.479 | 0.000 | 0.001 | 0.002 |
| LDL Chol | -0.0020 | 0.000 | -6.363 | 0.000 | -0.003 | -0.001 |
| HDL Chol | -0.0014 | 0.000 | -6.112 | 0.000 | -0.002 | |

```
-0.001
Heart Rate              0.0041      0.000     12.544      0.000       0.003
0.005
C-reactive Protein      0.0007      0.000      2.141      0.032    5.58e-05
0.001
==============================================================================
Omnibus:                      121.515   Durbin-Watson:                   0.963
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              133.124
Skew:                          -0.562   Prob(JB):                     1.24e-29
Kurtosis:                       2.692   Cond. No.                     1.82e+03
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
[2] The condition number is large, 1.82e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

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**: Hay cambios importantes de magnitud en los efectos marginales (considerando a OLS), las asociaciones aumentan (en valor absoluto), en particular para variables como glucosa, colesterol y plaquetas. Un cambio de una unidad, en torno al promedio (en otras palabras un percentil en la distribucion), se traduce en cambios de una un punto porcentual, en promedio. Esta interpretacion permite entender de mejor manera un incremento relativo (en torno a la media) de un percentil sobre el riesgo (medido en porcentaje).

```
Optimization terminated successfully.
         Current function value: 0.169501
         Iterations 12
                          Probit Regression Results
==============================================================================
Dep. Variable:                     dd   No. Observations:                 2351
Model:                         Probit   Df Residuals:                     2334
Method:                           MLE   Df Model:                           16
Date:                Thu, 02 May 2024   Pseudo R-squ.:                  0.6901
Time:                        14:56:46   Log-Likelihood:                -398.50
converged:                       True   LL-Null:                       -1286.0
Covariance Type:                  HC0   LLR p-value:                     0.000
==============================================================================
======
                           coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
------
const                    7.2068      0.671     10.745      0.000       5.892
8.521
Glucose                 -0.1102      0.006    -17.013      0.000      -0.123
-0.097
```

```
Cholesterol            0.1077      0.006     16.576      0.000       0.095
0.120
Hemoglobin             0.0089      0.002      4.782      0.000       0.005
0.013
Platelets             -0.1061      0.006    -16.481      0.000      -0.119
-0.093
White Blood Cells     -0.0844      0.006    -14.745      0.000      -0.096
-0.073
Red Blood Cells       -0.0886      0.004    -19.938      0.000      -0.097
-0.080
Hematocrit            -0.0704      0.005    -15.222      0.000      -0.079
-0.061
MCV                    0.1123      0.006     19.122      0.000       0.101
0.124
MCH                    0.0564      0.004     15.722      0.000       0.049
0.063
BMI                    0.0750      0.004     18.611      0.000       0.067
0.083
BP                    -0.0478      0.003    -14.658      0.000      -0.054
-0.041
Triglycerides         -0.0697      0.004    -18.543      0.000      -0.077
-0.062
HbA1c                 -0.0124      0.003     -4.036      0.000      -0.018
-0.006
HDL Chol              -0.0100      0.003     -3.517      0.000      -0.016
-0.004
Heart Rate             0.0923      0.006     16.763      0.000       0.082
0.103
C-reactive Protein     0.0538      0.004     14.054      0.000       0.046
0.061
==============================================================================
======

Possibly complete quasi-separation: A fraction 0.56 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
       Probit Marginal Effects
=====================================
Dep. Variable:                      dd
Method:                           dydx
At:                            overall
==============================================================================
======
                        dy/dx    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
------
Glucose               -0.0105      0.000    -24.741      0.000      -0.011
```

```
-0.010
Cholesterol            0.0103      0.001     20.090      0.000       0.009
0.011
Hemoglobin             0.0008      0.000      4.732      0.000       0.000
0.001
Platelets             -0.0101      0.000    -22.300      0.000      -0.011
-0.009
White Blood Cells     -0.0080      0.000    -16.467      0.000      -0.009
-0.007
Red Blood Cells       -0.0084      0.000    -26.038      0.000      -0.009
-0.008
Hematocrit            -0.0067      0.000    -16.600      0.000      -0.008
-0.006
MCV                    0.0107      0.000     30.718      0.000       0.010
0.011
MCH                    0.0054      0.000     19.323      0.000       0.005
0.006
BMI                    0.0071      0.000     27.174      0.000       0.007
0.008
BP                    -0.0046      0.000    -14.164      0.000      -0.005
-0.004
Triglycerides         -0.0066      0.000    -21.844      0.000      -0.007
-0.006
HbA1c                 -0.0012      0.000     -4.114      0.000      -0.002
-0.001
HDL Chol              -0.0010      0.000     -3.732      0.000      -0.001
-0.000
Heart Rate             0.0088      0.000     23.990      0.000       0.008
0.010
C-reactive Protein     0.0051      0.000     20.053      0.000       0.005
0.006
==============================================================================
======
```

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**: Los cambios en los coeficientes estimados con Probit (en terminos de efectos marginales) no son estadisticamente significativos, lo cual es esperado. Sin embargo, logit permite entender tambien los cambios en riesgo relativo (odds ratio) de tener al menos una enfermadad (reportado en la ultima tabla).

```
Optimization terminated successfully.
         Current function value: 0.168140
         Iterations 12
                           Logit Regression Results
==============================================================================
Dep. Variable:                     dd   No. Observations:                 2351
Model:                          Logit   Df Residuals:                     2334
```

| | | | |
|---|---|---|---|
| Method: | MLE | Df Model: | 16 |
| Date: | Thu, 02 May 2024 | Pseudo R-squ.: | 0.6926 |
| Time: | 16:15:57 | Log-Likelihood: | -395.30 |
| converged: | True | LL-Null: | -1286.0 |
| Covariance Type: | HC0 | LLR p-value: | 0.000 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 12.6531 | 1.048 | 12.069 | 0.000 | 10.598 | 14.708 |
| Glucose | -0.2044 | 0.022 | -9.182 | 0.000 | -0.248 | -0.161 |
| Cholesterol | 0.1965 | 0.015 | 12.994 | 0.000 | 0.167 | 0.226 |
| Hemoglobin | 0.0152 | 0.004 | 4.238 | 0.000 | 0.008 | 0.022 |
| Platelets | -0.1921 | 0.015 | -12.903 | 0.000 | -0.221 | -0.163 |
| White Blood Cells | -0.1475 | 0.010 | -14.827 | 0.000 | -0.167 | -0.128 |
| Red Blood Cells | -0.1573 | 0.011 | -14.192 | 0.000 | -0.179 | -0.136 |
| Hematocrit | -0.1241 | 0.008 | -15.174 | 0.000 | -0.140 | -0.108 |
| MCV | 0.2015 | 0.015 | 13.228 | 0.000 | 0.172 | 0.231 |
| MCH | 0.1036 | 0.009 | 11.039 | 0.000 | 0.085 | 0.122 |
| BMI | 0.1325 | 0.009 | 13.975 | 0.000 | 0.114 | 0.151 |
| BP | -0.0811 | 0.005 | -15.891 | 0.000 | -0.091 | -0.071 |
| Triglycerides | -0.1248 | 0.009 | -14.258 | 0.000 | -0.142 | -0.108 |
| HbA1c | -0.0247 | 0.011 | -2.352 | 0.019 | -0.045 | -0.004 |
| HDL Chol | -0.0252 | 0.009 | -2.738 | 0.006 | -0.043 | -0.007 |
| Heart Rate | 0.1717 | 0.018 | 9.739 | 0.000 | 0.137 | 0.206 |
| C-reactive Protein | 0.0988 | 0.011 | 8.659 | 0.000 | 0.076 | 0.121 |

```
Possibly complete quasi-separation: A fraction 0.53 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
        Logit Marginal Effects
=====================================
Dep. Variable:                      dd
Method:                           dydx
At:                            overall
==============================================================================
======
                        dy/dx    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
------
Glucose               -0.0110      0.001    -15.740      0.000      -0.012
-0.010
Cholesterol            0.0106      0.000     23.977      0.000       0.010
0.011
Hemoglobin             0.0008      0.000      4.445      0.000       0.000
0.001
Platelets             -0.0104      0.000    -27.334      0.000      -0.011
-0.010
White Blood Cells     -0.0080      0.001    -14.799      0.000      -0.009
-0.007
Red Blood Cells       -0.0085      0.000    -28.729      0.000      -0.009
-0.008
Hematocrit            -0.0067      0.000    -17.813      0.000      -0.007
-0.006
MCV                    0.0109      0.000     34.225      0.000       0.010
0.012
MCH                    0.0056      0.000     19.680      0.000       0.005
0.006
BMI                    0.0072      0.000     31.697      0.000       0.007
0.008
BP                    -0.0044      0.000    -12.872      0.000      -0.005
-0.004
Triglycerides         -0.0067      0.000    -24.245      0.000      -0.007
-0.006
HbA1c                 -0.0013      0.001     -2.534      0.011      -0.002
-0.000
HDL Chol              -0.0014      0.000     -3.133      0.002      -0.002
-0.001
Heart Rate             0.0093      0.001     18.168      0.000       0.008
0.010
C-reactive Protein     0.0053      0.000     14.792      0.000       0.005
0.006
==============================================================================
======
```

Odds Ratios

| | Odds Ratio | 5% | 95% |
|---|---|---|---|
| Glucose | 0.780333 | 0.851485 | 0.815133 |
| Cholesterol | 1.181553 | 1.253701 | 1.217092 |
| Hemoglobin | 1.008204 | 1.022479 | 1.015317 |
| Platelets | 0.801461 | 0.849636 | 0.825197 |
| White Blood Cells | 0.846248 | 0.879887 | 0.862904 |
| Red Blood Cells | 0.836061 | 0.873193 | 0.854425 |
| Hematocrit | 0.869205 | 0.897531 | 0.883255 |
| MCV | 1.187241 | 1.260289 | 1.223220 |
| MCH | 1.088910 | 1.129701 | 1.109118 |
| BMI | 1.120630 | 1.163054 | 1.141645 |
| BP | 0.912877 | 0.931334 | 0.922060 |
| Triglycerides | 0.867703 | 0.897980 | 0.882711 |
| HbA1c | 0.955712 | 0.995886 | 0.975593 |
| HDL Chol | 0.957723 | 0.992873 | 0.975139 |
| Heart Rate | 1.147008 | 1.229087 | 1.187339 |
| C-reactive Protein | 1.079460 | 1.128850 | 1.103879 |

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**: En base a los resultados, el modelo MCO produce resultados incorrectos, por tanto Probit o Logit podrian ser apropiados. Dado el contexto, el modelo Logit provee mas informacion, por tanto es mas conveniente. Hay solo algunas variables como Insulina, LDL Chol, Creatinina y BP Sistolica no afectan directamente (teniendo en cuenta que el modelo puede estar mal especificado al no permitir interacciones entre variables).

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | dd | R-squared: | 0.475 |
| Model: | OLS | Adj. R-squared: | 0.472 |
| Method: | Least Squares | F-statistic: | 214.8 |
| Date: | Thu, 02 May 2024 | Prob (F-statistic): | 0.00 |
| Time: | 14:57:20 | Log-Likelihood: | -566.18 |
| No. Observations: | 2351 | AIC: | 1166. |
| Df Residuals: | 2334 | BIC: | 1264. |
| Df Model: | 16 | | |
| Covariance Type: | HC0 | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.6200 | 0.041 | 15.220 | 0.000 | 0.540 | 0.700 |
| Glucose | -0.0040 | 0.000 | -15.968 | 0.000 | -0.004 | |

```
-0.004
Cholesterol            0.0049      0.000     16.095      0.000       0.004
0.005
Hemoglobin             0.0019      0.000      7.540      0.000       0.001
0.002
Platelets             -0.0039      0.000    -17.566      0.000      -0.004
-0.003
White Blood Cells     -0.0041      0.000    -18.960      0.000      -0.005
-0.004
Red Blood Cells       -0.0039      0.000    -15.390      0.000      -0.004
-0.003
Hematocrit            -0.0021      0.000     -8.447      0.000      -0.003
-0.002
MCV                    0.0063      0.000     25.507      0.000       0.006
0.007
MCH                    0.0026      0.000     11.804      0.000       0.002
0.003
BMI                    0.0034      0.000     10.348      0.000       0.003
0.004
BP                    -0.0017      0.000     -6.115      0.000      -0.002
-0.001
Triglycerides         -0.0030      0.000    -10.436      0.000      -0.004
-0.002
HbA1c                  0.0011      0.000      4.555      0.000       0.001
0.002
HDL Chol              -0.0010      0.000     -4.022      0.000      -0.001
-0.001
Heart Rate             0.0045      0.000     15.590      0.000       0.004
0.005
C-reactive Protein     0.0009      0.000      2.579      0.010       0.000
0.002
==============================================================================
Omnibus:                      144.439   Durbin-Watson:                   0.945
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              151.256
Skew:                          -0.584   Prob(JB):                     1.43e-33
Kurtosis:                       2.579   Cond. No.                     1.55e+03
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
[2] The condition number is large, 1.55e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

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.

**R**: Una vez excluidas las variables que no contribuian a la varianza total del modelo (tampoco directamente), hay ciertas variables que contribuyen de forma significativa, aunque de magnitud menor, a la variacion en el numero de enfermedades contabilizadas. Por ejemplo, un incremento en

10 percentiles en el Indice de Masa Corporal (BMI) se traduce en un incremento (sobre el promedio) de 0.14 en el numero de enfermedades.

Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Disease | No. Observations: | 2351 |
| Model: | GLM | Df Residuals: | 2333 |
| Model Family: | Poisson | Df Model: | 17 |
| Link Function: | Log | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -3133.6 |
| Date: | Thu, 02 May 2024 | Deviance: | 1685.3 |
| Time: | 16:29:47 | Pearson chi2: | 1.31e+03 |
| No. Iterations: | 5 | Pseudo R-squ. (CS): | 0.3767 |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -0.8870 | 0.150 | -5.929 | 0.000 | -1.180 | -0.594 |
| Glucose | -0.0069 | 0.001 | -8.282 | 0.000 | -0.009 | -0.005 |
| Cholesterol | 0.0103 | 0.001 | 11.078 | 0.000 | 0.008 | 0.012 |
| Hemoglobin | 0.0114 | 0.001 | 14.270 | 0.000 | 0.010 | 0.013 |
| Platelets | -0.0124 | 0.001 | -19.502 | 0.000 | -0.014 | -0.011 |
| Hematocrit | -0.0027 | 0.001 | -3.806 | 0.000 | -0.004 | -0.001 |
| MCV | 0.0110 | 0.001 | 15.420 | 0.000 | 0.010 | 0.012 |
| MCH | 0.0085 | 0.001 | 13.236 | 0.000 | 0.007 | 0.010 |
| BMI | 0.0137 | 0.001 | 12.223 | 0.000 | 0.011 | 0.016 |
| Systolic Blood Pressure | -0.0029 | 0.001 | -3.582 | 0.000 | -0.004 | -0.001 |
| BP | -0.0067 | 0.001 | -7.633 | 0.000 | -0.008 | -0.005 |
| Triglycerides | -0.0047 | 0.001 | -5.918 | 0.000 | -0.006 | -0.003 |
| HbA1c | 0.0032 | 0.001 | 4.614 | 0.000 | 0.002 | 0.004 |
| LDL Chol | -0.0088 | 0.001 | -10.848 | 0.000 | -0.010 | -0.007 |

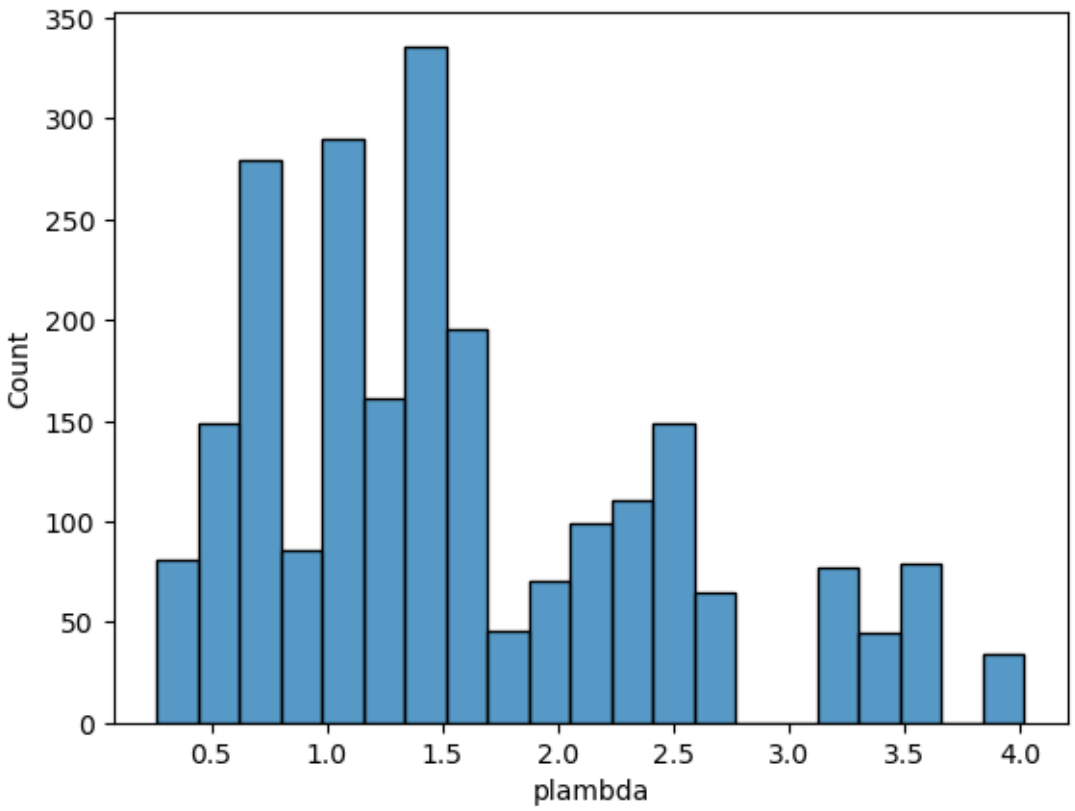
```
HDL Chol                -0.0032      0.001     -4.736      0.000      -0.005
-0.002
Heart Rate               0.0019      0.001      2.486      0.013       0.000
0.003
Creatinine               0.0052      0.001      5.936      0.000       0.003
0.007
C-reactive Protein       0.0051      0.001      6.186      0.000       0.003
0.007
==============================================================================
==========
```

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

**R**: El analisis indica que la sobredispersion es baja, y el test muestra un valor de alpha estadisticamente significativo, pero sugiriendo que existe subdispersion (menor varianza que la media).

```
c:\Users\juanc\anaconda3\lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

```
<Axes: xlabel='plambda', ylabel='Count'>
```



```
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:                Disease   R-squared (uncentered):
0.381
Model:                            OLS   Adj. R-squared (uncentered):
0.381
Method:                 Least Squares   F-statistic:
1449.
Date:                Thu, 02 May 2024   Prob (F-statistic):
1.97e-247
Time:                        16:31:11   Log-Likelihood:
-2008.7
No. Observations:                2351   AIC:
4019.
Df Residuals:                    2350   BIC:
4025.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.2466      0.006    -38.068      0.000      -0.259      -0.234
==============================================================================
Omnibus:                      581.822   Durbin-Watson:                   1.455
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1196.992
Skew:                           1.450   Prob(JB):                    1.19e-260
Kurtosis:                       4.951   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 responder la pregunta 6. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

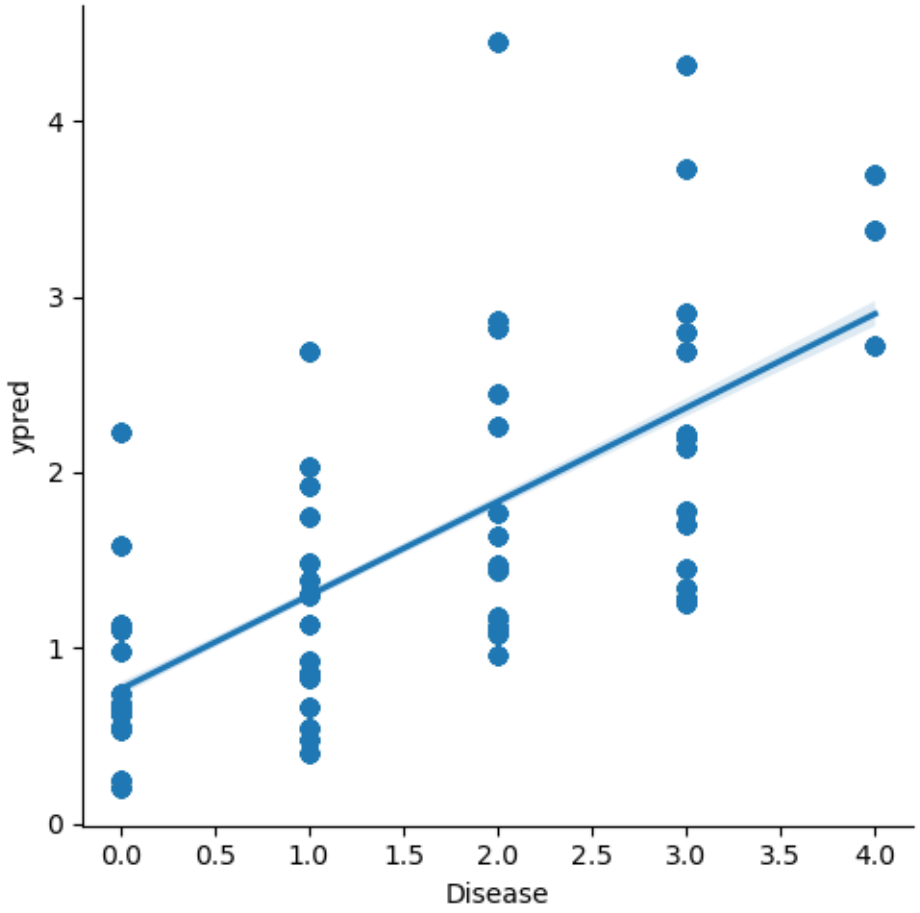
**R**: Los resultados del modelo Binonial Negativa, con el valor de alpha sugerido en la regresion auxiliar, entrega un peor ajuste respecto del modelo Poisson (basado en el valor de la Log-Likelihood). Sin embargo, los valores estimados son arbitrariamente similares entre ambos modelos (y su interpretacion es exactamente la misma).

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                Disease   No. Observations:                2351
Model:                            GLM   Df Residuals:                    2333
```

| | | | |
|---|---|---|---|
| Model Family: | NegativeBinomial | Df Model: | 17 |
| Link Function: | Log | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -3703.3 |
| Date: | Thu, 02 May 2024 | Deviance: | 1005.4 |
| Time: | 16:36:58 | Pearson chi2: | 672. |
| No. Iterations: | 10 | Pseudo R-squ. (CS): | 0.2058 |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -0.9291 | 0.225 | -4.134 | 0.000 | -1.370 | -0.489 |
| Glucose | -0.0063 | 0.001 | -4.867 | 0.000 | -0.009 | -0.004 |
| Cholesterol | 0.0110 | 0.001 | 8.006 | 0.000 | 0.008 | 0.014 |
| Hemoglobin | 0.0111 | 0.001 | 9.342 | 0.000 | 0.009 | 0.013 |
| Platelets | -0.0140 | 0.001 | -14.102 | 0.000 | -0.016 | -0.012 |
| Hematocrit | -0.0037 | 0.001 | -3.531 | 0.000 | -0.006 | -0.002 |
| MCV | 0.0132 | 0.001 | 12.359 | 0.000 | 0.011 | 0.015 |
| MCH | 0.0093 | 0.001 | 9.563 | 0.000 | 0.007 | 0.011 |
| BMI | 0.0136 | 0.002 | 8.632 | 0.000 | 0.010 | 0.017 |
| Systolic Blood Pressure | -0.0029 | 0.001 | -2.344 | 0.019 | -0.005 | -0.000 |
| BP | -0.0069 | 0.001 | -5.327 | 0.000 | -0.009 | -0.004 |
| Triglycerides | -0.0058 | 0.001 | -4.881 | 0.000 | -0.008 | -0.003 |
| HbA1c | 0.0041 | 0.001 | 3.835 | 0.000 | 0.002 | 0.006 |
| LDL Chol | -0.0087 | 0.001 | -7.063 | 0.000 | -0.011 | -0.006 |
| HDL Chol | -0.0044 | 0.001 | -4.082 | 0.000 | -0.006 | -0.002 |
| Heart Rate | 0.0026 | 0.001 | 2.258 | 0.024 | 0.000 | 0.005 |
| Creatinine | 0.0056 | 0.001 | 4.204 | 0.000 | 0.003 | 0.008 |
| C-reactive Protein | 0.0048 | 0.001 | 3.809 | 0.000 | 0.002 | |

```
0.007
==============================================================================
==========

 <seaborn.axisgrid.FacetGrid at 0x1d33cd8a9d0>
```




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**: En virtud de los resultados, ambos modelos entregan resultados similares, sin embargo el modelo Poisson es mas parsimonioso y explica una mayor fraccion de la varianza. La significancia de las variables en cada modelo son arbitrariamente similares (desde una perspectiva estadistica). Todas las variables que quedan en el modelo son robustas en su asociacion con el numero de enfermedades.