Tarea1

April 22, 2024

#

Tarea 1 Agustín Astete.

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 principio no parecen haber datos faltantes (Anexo 1.1). La variable a explicar "Disease" toma valores enteros entre 0 y 4 (Anexo 1.2), tiene una cantidad de observaciones similares cuando toma valores entre 0 y 3, mientras que cuando toma el valor 4, tiene alrededor de 4 veces menos observaciones (Anexo 1.3). Por otro lado el resto de las variables son del tipo "Float" y toman valores entre 0 y 1, lo que indica que están normalizadas (Anexo 1.1).

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:
Selección de variables:
A partir del mapa de calor (Anexo 2.1) de correlaciones se observó que Cholesterol tiene una muy baja correlación con la variable a explicar, sin embargo se retiró "HDL Cholesterol", el llamado "colesterol bueno" puesto que debería estar altamente relacionado con "Cholesterol" y en el contexto de la medicina no es un buen indicador de presencia de enfermedades, luego se retiró la variable "Systolic Blood Pressure" por no ser significativa según el valor p obtenido en los modelos probit y logit, si bien es una variable que sería mejor mantener, aun se cuenta con "Diastolic Blood Pressure" variable con la cual debería estar relacionada.

Analisis Regresión Lineal
La regresión Lineal demostró ser poco apta para el set de datos (Anexo 2.2), sin embargo se obtuvo que casi todas las variables son significativas, a excepción de "Creatinine" que tiene un valor p bastante superior a 0.05 y además el coeficiente obtenido es el mas bajo, por lo que se puede interpretar como "ruido" en esta regresión. Según los coeficientes obtenidos, el volumen corposcular medio por ejemplo, es el que más incide en la probabilidad de tener enfermedades, en este caso el coeficiente es 0.64, es decir que, en promedio, la probabildiad de tener enfermedades aumentará 0.64% por cada cambio porcentual del volumen corposcular medio.

Según el modelo las variables que mas aumentan la probabildiad de tener enfermedades son: Volumen Corpuscular Medio, Colesterol, IMC y Ritmo Cardiaco; mientras que las variables que mas disminuyen esta probabilidad son: Globulos blancos, Plaquetas, Globulos Rojos y Glucosa (Anexo 2.3).

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: Para este modelo, observamos los coeficientes como cambios marginales, el cambio más grande en la probabilidad de tener enfermedades lo explica la variable "Glucose", ya que el coeficiente es -1.58, esto quiere decir que, en promedio, la probabildiad de tener enfermedades disminuirá 1.58% por cada cambio porcentual de "Glucose" (Anexo 5.1).

Según este modelo las variables que mas inciden en la probabilidad de tener enfermedades son: Colesterol, Volumen Corposcular Medio, Ritmo Cardiaco, IMC, Proteina C-reactiva y Hemoglobina Corposcular Media; mientras que las variables que mas inciden de forma negativa son: Glucosa, Plaquetas, Globulos Blancos, Trigliceridos, Globulos Rojos y Hematocrito

```
Optimization terminated successfully.
         Current function value: 0.153596
         Iterations 13
```

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: al igual que en el modelo anterior, observamos los coeficientes como cambios marginales, los cambios mas grandes en la probabilidad de tener enfermedades sigue siendo explicado por las mismas variables, y se puede interpretar de la misma forma, cabe mencionar que las variables "HbA1c" y "LDL Cholesterol" fueron las que mantuvieron los efectos marginales mas bajos, es decir la probabilidad de tener menos enfermedades es menos sensible a estas variables (Anexo 5.1).

```
Optimization terminated successfully.
         Current function value: 0.151444
         Iterations 13
```

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 primer lugar, ya que la variable de interés se está tratando como una variable dicotomica, una regresión lineal no es adecuada para explicar su comportamiento, esto se ve reflejeado en su bajo valor R^2, es decir explica poco de la varianza de "Disease"(Anexo 2.2).

Los modelos Probit y Logit entregaron resultados sumamente similares tanto en valores de coeficientes como en significancia de las variables, además ambos obtuvieron un pseudo R^2 de alrededor de 0.72 (Anexos 3.1 y 4.1).

En el modelo de probabilidad lineal, en general, los efectos de las variables en la probabilidad de tener enfermedades tomaron la misma dirección que en los modelos Probit y Logit, en particular las variables con mayor magnitud en sus coeficientes no cambiaron. Solo en este modelo se obtuvo que la variable "Creatinine" no es significativa, además la regresión obtuvo un bajo R^2 de 0.48 (Anexo 5.1).

Respecto a la robustez de las variables "Creatinine" se muestra como la menos robusta debido a tener un valor p de 0.54 en el modelo de probabilidad lineal, por otro lado en los modelos Probit y Logit la mayoria de las variables mostró valores p menores a 0.001 a excepción de "HbA1c" y "LDL Cholesterol" que en ambos modelos obtuvieron valores mayores, aunque se mantuvieron

como significativas, sus coeficientes fueron pequeños comparados con el resto de variables, por lo que podrían interpretarse como "ruido" en la explicación de la varianza.

Debido a lo comentado anteriormente acerca de la significancia de las variables y el valor R^2 obtenido, tanto el modelo Probit como Logit son adecuados para responder la pregunta de investigación.

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:
Selección de variables:
Se retiró la variable Insulin, C-reactive Protein y Creatinine por tener Correlacion muy baja con "Disease" (Anexo 6.1 y 6.2) además, se podria argumentar que estas variables no son tan importantes para medir la cantidad de enfermedades como si lo puede ser la Presión o la Glucosa, las cuales si bien tambien presentan baja correlación podrían ser importantes en la explicación de la varianza. Posteriormente, dada su baja significancia para el modelo de Poisson, se retiró Red Blood Cells y White Blood Cells.

Analisis Modelo Poisson
Para este modelo los coeficientes estimados representan el cambio porcentual en el numero de eventos, dado un cambio en la variable. Para el caso estudiado la variable que mas efecto tiene sobre la cantidad de enfermedades alrededor de la media (es decir 1.58 enfermedades) es "Platelets", ya que su coeficiente es -1.33 y la variable está normalizada, en este caso, si las plaquetas en un punto porcentual, la cantidad de enfermedades debería disminuir en 0.133 (Anexo 6.3).

En resumen las variables que mas indicen en el aumento del numero de enfermedades son: Hemoglobina, Volumen Corposcular Medio, IMC, Colesterol, Hemoglobina Corpuscular media y el test HbA1c, por otro lado las variables mas que reducen el numero de enfermedades son: Plaquetas, Colesterol LDL, Trigliceridos, Glucosa y Colesterol HDL.

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

R: Para determinar si existe sobre dispersión se realiza un test con los resultados del modelo Poisson, el cuociente del Test de Pearson Chi2 dividido por los grados de libertad de los residuos debe ser mayor a 1 para que exista sobre dispersión, dado que el couciente obtenido es 0.6 no se indica sobre dispersión.

```
0.594841177216956
```

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

De esta forma se obtuvo un posible valor optimo de alpha para un modelo Binomial negativo de 0.79 (Anexo 7.1).

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 coeficientes de este modelo se interpretan de forma similar al modelo anterior, es decir, al divir los coeficientes por 100, estos representan el cambio en el numero de eventos de la variable estimada por cambio porcentual en las variables independientes (Anexo 8.1). En resumen este modelo sugiere que las variables que mas aumentan la cantidad de enfermedades son: Volumen Corposcular Medio, Hemoglobina, IMC, Colesterol y Hemoglobina Corposcular Media; mientras que las variables que mas disminuten la cantidad de enfermedades son: Plaquetas, Colesterol LDL, Trigliceridos, Colesterol HDL y Glucosa.

0.0.2 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: Dado que el resultado del test de sobre dispersión fue negativo, la utilización del Modelo Binomial Negativa no parece ser de utilidad, los datos no parecen ajustarse de mejor forma.
Ambos modelos produjeron resultados similares, los variables con coeficientes de mayor magnitud se mantuvieron, tanto las que aumentan la cantidad de enfermedades como las que las disminuyen, así como la significancia de las variables (Anexo 9.1).
Todas las variables escogidas resultaron ser robustas, sin embargo las variables con mayor incidencia en ambos modelos fueron: Hemoglobina, Volumen Corposcular Medio, IMC, Colesterol, Hemoglobina Corpuscular media, test HbA1c, Plaquetas, Colesterol LDL, Trigliceridos, Glucosa y Colesterol HDL.
Finalmente debido a que no se demostró sobre dispersión el Modelo de Poisson es mas adecuado para resolver la pregunta de investigación.

##

Anexo

0.0.3 Anexo 1.1: Cantidad y tipo de dato por Columna del dataframe

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2351 entries, 0 to 2350
Data columns (total 21 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   Glucose                      2351 non-null   float64
 1   Cholesterol                  2351 non-null   float64
 2   Hemoglobin                   2351 non-null   float64
 3   Platelets                    2351 non-null   float64
 4   White Blood Cells            2351 non-null   float64
 5   Red Blood Cells              2351 non-null   float64
 6   Hematocrit                   2351 non-null   float64
 7   Mean Corpuscular Volume      2351 non-null   float64
 8   Mean Corpuscular Hemoglobin  2351 non-null   float64
 9   Insulin                      2351 non-null   float64
 10  BMI                          2351 non-null   float64
 11  Systolic Blood Pressure      2351 non-null   float64
 12  Diastolic Blood Pressure     2351 non-null   float64
```

```
 13  Triglycerides                2351 non-null   float64
 14  HbA1c                        2351 non-null   float64
 15  LDL Cholesterol              2351 non-null   float64
 16  HDL Cholesterol              2351 non-null   float64
 17  Heart Rate                   2351 non-null   float64
 18  Creatinine                   2351 non-null   float64
 19  C-reactive Protein           2351 non-null   float64
 20  Disease                      2351 non-null   int64
dtypes: float64(20), int64(1)
memory usage: 385.8 KB
```

0.0.4 Anexo 1.2: Estadisticas por variable

	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	

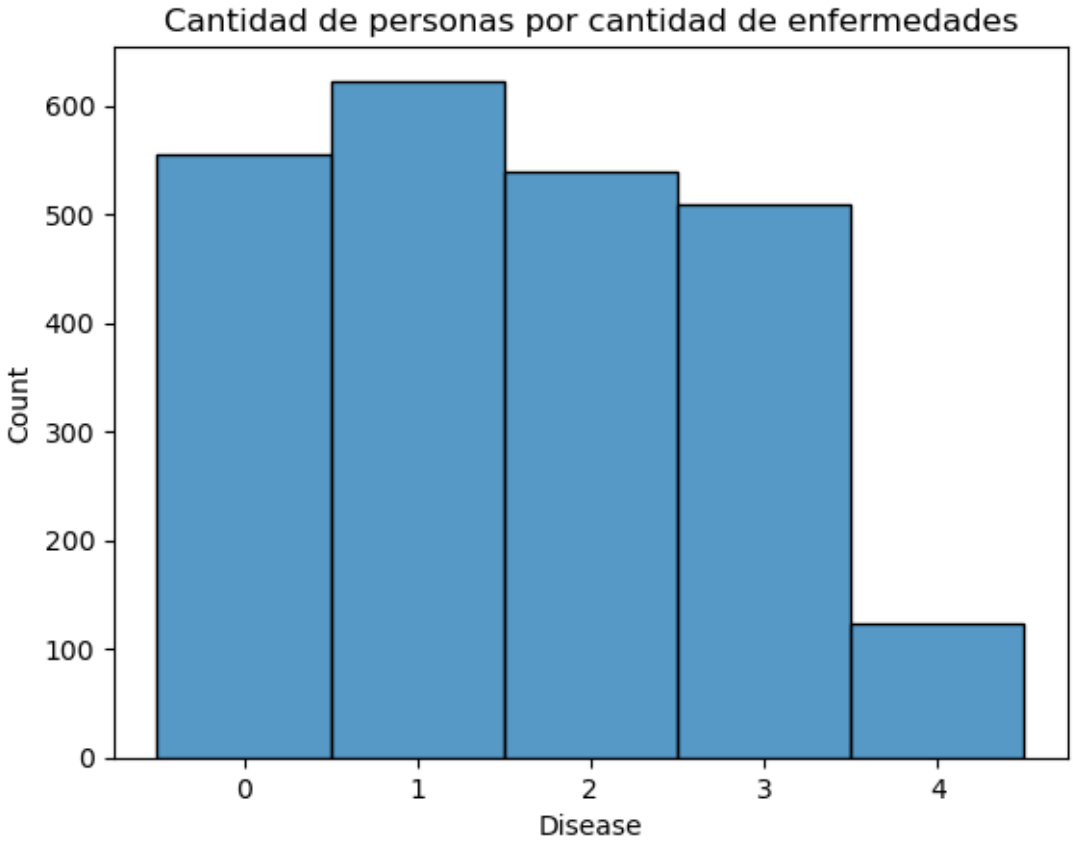
	Red Blood Cells	Hematocrit	Mean Corpuscular Volume	\
count	2351.000000	2351.000000	2351.000000	
mean	0.506590	0.507152	0.492200	
std	0.266565	0.285537	0.275735	
min	0.044565	0.011772	0.046942	
25%	0.263589	0.288132	0.287532	
50%	0.467431	0.493428	0.453052	
75%	0.743670	0.753657	0.722293	
max	1.000000	0.977520	0.995263	

	Mean Corpuscular Hemoglobin	Insulin	…	Systolic Blood Pressure	\
count	2351.000000	2351.000000	…	2351.000000	
mean	0.484459	0.447062	…	0.381211	
std	0.315618	0.242861	…	0.232785	
min	0.000554	0.034129	…	0.005988	
25%	0.207938	0.219111	…	0.179951	
50%	0.420723	0.444806	…	0.359064	
75%	0.778160	0.654441	…	0.580903	
max	0.963235	0.966784	…	0.829100	

	Diastolic Blood Pressure	Triglycerides	HbA1c	LDL Cholesterol	\
count	2351.000000	2351.000000	2351.000000	2351.000000	
mean	0.421708	0.374373	0.439112	0.421777	
std	0.248768	0.256981	0.263779	0.252124	

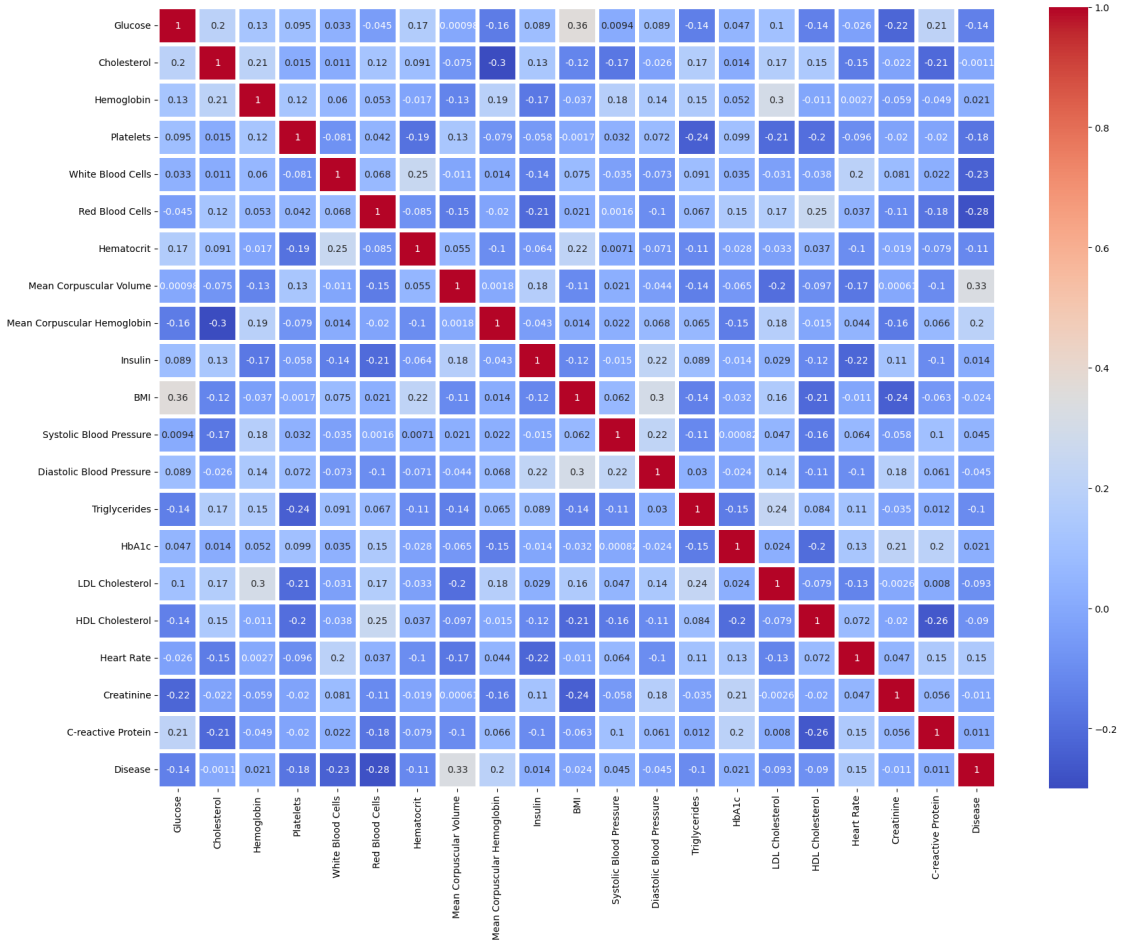
```
min                   0.005579       0.005217     0.016256         0.033037
25%                   0.175469       0.184604     0.188750         0.217757
50%                   0.474378       0.317857     0.466375         0.413071
75%                   0.663382       0.572330     0.652514         0.604753
max                   0.934617       0.973679     0.950218         0.983826

       HDL Cholesterol   Heart Rate   Creatinine  C-reactive Protein  \
count      2351.000000  2351.000000  2351.000000         2351.000000
mean          0.546079     0.582255     0.425075            0.430308
std           0.269511     0.250915     0.229298            0.243034
min           0.039505     0.114550     0.021239            0.004867
25%           0.307132     0.339125     0.213026            0.196192
50%           0.512941     0.610860     0.417295            0.481601
75%           0.779378     0.800666     0.606719            0.631426
max           0.989411     0.996873     0.925924            0.797906

           Disease
count  2351.000000
mean      1.583156
std       1.209799
min       0.000000
25%       1.000000
50%       1.000000
75%       3.000000
max       4.000000

[8 rows x 21 columns]
```

0.0.5 Anexo 1.3: Cantidad de personas por cantidad de enfermedades



0.0.6 Anexo 2.1: Matriz de correlaciones dataframe con "Desease" como variable dicotómica

```
<AxesSubplot:>
```



0.0.7 Anexo 2.2: Resultados Regresión lineal

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                Disease   R-squared:                       0.482
Model:                            OLS   Adj. R-squared:                  0.478
Method:                 Least Squares   F-statistic:                     120.4
Date:                Mon, 22 Apr 2024   Prob (F-statistic):          4.58e-316
Time:                        20:53:09   Log-Likelihood:                -551.45
No. Observations:                2351   AIC:                             1141.
Df Residuals:                    2332   BIC:                             1250.
Df Model:                          18
Covariance Type:            nonrobust
==============================================================================
===============
```

```
                                 coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
---------------
const                          0.6209      0.053     11.646      0.000
0.516       0.725
Glucose                       -0.3685      0.033    -11.269      0.000
-0.433      -0.304
Cholesterol                    0.5095      0.033     15.651      0.000
0.446       0.573
Hemoglobin                     0.2069      0.029      7.258      0.000
0.151       0.263
Platelets                     -0.4146      0.024    -17.438      0.000
-0.461      -0.368
White Blood Cells             -0.4184      0.025    -16.764      0.000
-0.467      -0.369
Red Blood Cells               -0.4068      0.027    -15.218      0.000
-0.459      -0.354
Hematocrit                    -0.2364      0.025     -9.300      0.000
-0.286      -0.187
Mean Corpuscular Volume        0.6377      0.025     25.419      0.000
0.588       0.687
Mean Corpuscular Hemoglobin    0.2925      0.024     12.421      0.000
0.246       0.339
Insulin                       -0.0815      0.032     -2.570      0.010
-0.144      -0.019
BMI                            0.3934      0.035     11.089      0.000
0.324       0.463
Diastolic Blood Pressure      -0.1469      0.031     -4.760      0.000
-0.207      -0.086
Triglycerides                 -0.2561      0.029     -8.910      0.000
-0.312      -0.200
HbA1c                          0.1477      0.027      5.468      0.000
0.095       0.201
LDL Cholesterol               -0.1783      0.031     -5.832      0.000
-0.238      -0.118
Heart Rate                     0.3919      0.028     13.789      0.000
0.336       0.448
Creatinine                     0.0199      0.033      0.609      0.542
-0.044       0.084
C-reactive Protein             0.1022      0.032      3.244      0.001
0.040       0.164
==============================================================================
Omnibus:                      129.544   Durbin-Watson:                   0.945
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              144.161
Skew:                          -0.588   Prob(JB):                     4.96e-32
Kurtosis:                       2.701   Cond. No.                         23.0
```

```
==============================================================================

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

0.0.8 Anexo 2.3: Coeficientes de la regresión lineal ordenados de mayor a menor

	coef	P>|t|
Mean Corpuscular Volume	0.6377	0.000
const	0.6209	0.000
Cholesterol	0.5095	0.000
BMI	0.3934	0.000
Heart Rate	0.3919	0.000
Mean Corpuscular Hemoglobin	0.2925	0.000
Hemoglobin	0.2069	0.000
HbA1c	0.1477	0.000
C-reactive Protein	0.1022	0.001
Creatinine	0.0199	0.542
Insulin	-0.0815	0.010
Diastolic Blood Pressure	-0.1469	0.000
LDL Cholesterol	-0.1783	0.000
Hematocrit	-0.2364	0.000
Triglycerides	-0.2561	0.000
Glucose	-0.3685	0.000
Red Blood Cells	-0.4068	0.000
Platelets	-0.4146	0.000
White Blood Cells	-0.4184	0.000

0.0.9 Anexo 3.1: Resultados Modelo Probit

```
Optimization terminated successfully.
         Current function value: 0.153596
         Iterations 13

<class 'statsmodels.iolib.summary.Summary'>
"""
                           Probit Regression Results
==============================================================================
Dep. Variable:                Disease   No. Observations:                 2351
Model:                         Probit   Df Residuals:                     2332
Method:                           MLE   Df Model:                           18
Date:                Mon, 22 Apr 2024   Pseudo R-squ.:                 0.7192
Time:                        20:53:10   Log-Likelihood:                -361.10
converged:                       True   LL-Null:                       -1286.0
Covariance Type:            nonrobust   LLR p-value:                     0.000
```

```
==============================================================================
==============
                                 coef    std err          z      P>|z|
[0.025      0.975]
------------------------------------------------------------------------------
--------------
const                          13.8698      1.905      7.279      0.000
10.135      17.604
Glucose                       -18.2319      2.079     -8.767      0.000
-22.308     -14.156
Cholesterol                    17.6284      2.266      7.780      0.000
13.187      22.069
Hemoglobin                      1.9768      0.436      4.535      0.000
1.122       2.831
Platelets                     -16.8538      1.790     -9.415      0.000
-20.362     -13.345
White Blood Cells             -14.9102      1.986     -7.508      0.000
-18.802     -11.018
Red Blood Cells               -13.6328      1.371     -9.947      0.000
-16.319     -10.947
Hematocrit                    -10.6724      1.303     -8.189      0.000
-13.227      -8.118
Mean Corpuscular Volume        16.2139      1.593     10.178      0.000
13.092      19.336
Mean Corpuscular Hemoglobin     6.6431      0.732      9.076      0.000
5.209       8.078
Insulin                         2.5996      0.947      2.746      0.006
0.744       4.455
BMI                            11.8987      1.504      7.910      0.000
8.950      14.847
Diastolic Blood Pressure       -5.5229      0.842     -6.556      0.000
-7.174      -3.872
Triglycerides                 -14.0645      1.829     -7.691      0.000
-17.648     -10.480
HbA1c                          -1.4800      0.600     -2.469      0.014
-2.655      -0.305
LDL Cholesterol                -2.2990      0.857     -2.683      0.007
-3.979      -0.619
Heart Rate                     15.7178      1.832      8.578      0.000
12.126      19.309
Creatinine                     -4.7360      0.744     -6.364      0.000
-6.195      -3.277
C-reactive Protein              8.8002      1.011      8.701      0.000
6.818      10.782
==============================================================================
==============
```

```
Possibly complete quasi-separation: A fraction 0.65 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
"""
```

0.0.10 Anexo 4.1: Resultados Modelo Probit

```
Optimization terminated successfully.
         Current function value: 0.151444
         Iterations 13

<class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:                Disease   No. Observations:                 2351
Model:                          Logit   Df Residuals:                     2332
Method:                           MLE   Df Model:                           18
Date:                Mon, 22 Apr 2024   Pseudo R-squ.:                  0.7231
Time:                        20:53:10   Log-Likelihood:                -356.05
converged:                       True   LL-Null:                       -1286.0
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
===============
                                 coef    std err          z      P>|z|
[0.025      0.975]
------------------------------------------------------------------------------
---------------
const                         22.1280      3.483      6.352      0.000
15.301      28.955
Glucose                      -32.1377      3.989     -8.056      0.000
-39.956     -24.319
Cholesterol                   29.2802      4.216      6.945      0.000
21.017      37.543
Hemoglobin                     3.6226      0.774      4.679      0.000
2.105       5.140
Platelets                    -28.4869      3.346     -8.513      0.000
-35.046     -21.928
White Blood Cells            -23.7035      3.610     -6.566      0.000
-30.779     -16.628
Red Blood Cells              -23.2825      2.583     -9.015      0.000
-28.344     -18.221
Hematocrit                   -17.3355      2.426     -7.146      0.000
-22.090     -12.581
Mean Corpuscular Volume       27.6954      3.037      9.121      0.000
21.744      33.647
Mean Corpuscular Hemoglobin   10.8716      1.350      8.055      0.000
```

```
8.226      13.517
Insulin                          4.6287      1.689      2.741      0.006
1.319       7.938
BMI                             20.5255      2.850      7.201      0.000
14.939      26.112
Diastolic Blood Pressure        -8.3599      1.570     -5.324      0.000
-11.437      -5.282
Triglycerides                  -23.8323      3.368     -7.076      0.000
-30.433     -17.231
HbA1c                           -3.4489      1.139     -3.027      0.002
-5.682      -1.216
LDL Cholesterol                 -3.6219      1.508     -2.402      0.016
-6.577      -0.666
Heart Rate                      27.0865      3.385      8.002      0.000
20.452      33.721
Creatinine                      -8.7663      1.421     -6.171      0.000
-11.551      -5.982
C-reactive Protein              16.0661      1.989      8.076      0.000
12.167      19.965
================================================================================
==============

Possibly complete quasi-separation: A fraction 0.62 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
"""
```

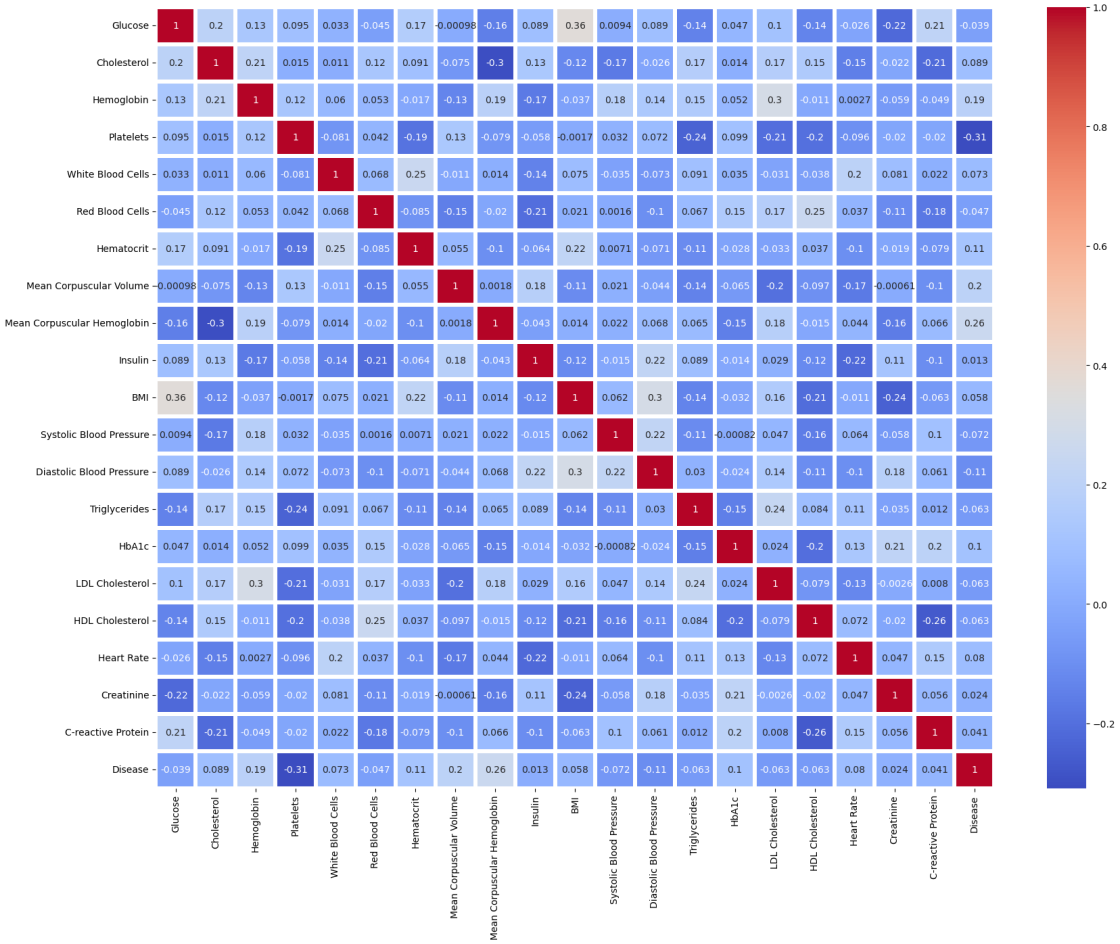
0.0.11 Anexo 5.1: Comparación Coeficientes Regresión lineal y Cambios marginales Modelos Probit y Logit

	Probit: dy/dx	Logit: dy/dx	OLS: coef
Cholesterol	1.5365	1.4381	0.5095
Mean Corpuscular Volume	1.4132	1.3603	0.6377
Heart Rate	1.3700	1.3304	0.3919
BMI	1.0371	1.0081	0.3934
C-reactive Protein	0.7670	0.7891	0.1022
Mean Corpuscular Hemoglobin	0.5790	0.5340	0.2925
Insulin	0.2266	0.2273	-0.0815
Hemoglobin	0.1723	0.1779	0.2069
HbA1c	-0.1290	-0.1694	0.1477
LDL Cholesterol	-0.2004	-0.1779	-0.1783
Creatinine	-0.4128	-0.4306	0.0199
Diastolic Blood Pressure	-0.4814	-0.4106	-0.1469
Hematocrit	-0.9302	-0.8515	-0.2364
Red Blood Cells	-1.1882	-1.1436	-0.4068
Triglycerides	-1.2259	-1.1706	-0.2561
White Blood Cells	-1.2996	-1.1642	-0.4184
Platelets	-1.4690	-1.3992	-0.4146

```
Glucose                          -1.5891        -1.5785     -0.3685
```

0.0.12 Anexo 6.1: Matriz de correlaciones dataframe

```
<AxesSubplot:>
```



0.0.13 Anexo 6.2: Resultados Modelo Poisson

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                Disease   No. Observations:                 2351
Model:                            GLM   Df Residuals:                     2335
Model Family:                 Poisson   Df Model:                           15
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3194.1
Date:                Mon, 22 Apr 2024   Deviance:                       1806.2
```

```
Time:                          20:53:12   Pearson chi2:                 1.40e+03
No. Iterations:                       5   Pseudo R-squ. (CS):             0.3438
Covariance Type:            nonrobust
==============================================================================
===============
                                 coef    std err          z      P>|z|
[0.025      0.975]
------------------------------------------------------------------------------
---------------
Cholesterol                     0.7357      0.085      8.665      0.000
0.569       0.902
Hemoglobin                      0.9056      0.076     11.885      0.000
0.756       1.055
Platelets                      -1.3133      0.061    -21.532      0.000
-1.433      -1.194
White Blood Cells               0.0439      0.066      0.670      0.503
-0.085       0.172
Red Blood Cells                 0.0249      0.073      0.342      0.733
-0.118       0.168
Hematocrit                     -0.3364      0.072     -4.687      0.000
-0.477      -0.196
Mean Corpuscular Volume         0.8589      0.062     13.945      0.000
0.738       0.980
Mean Corpuscular Hemoglobin     0.7834      0.058     13.550      0.000
0.670       0.897
BMI                             0.6271      0.086      7.317      0.000
0.459       0.795
Systolic Blood Pressure        -0.3391      0.074     -4.557      0.000
-0.485      -0.193
Diastolic Blood Pressure       -0.3696      0.078     -4.733      0.000
-0.523      -0.217
Triglycerides                  -0.5126      0.078     -6.565      0.000
-0.666      -0.360
HbA1c                           0.3797      0.063      6.000      0.000
0.256       0.504
LDL Cholesterol                -0.8605      0.080    -10.816      0.000
-1.016      -0.705
HDL Cholesterol                -0.4537      0.064     -7.040      0.000
-0.580      -0.327
Heart Rate                      0.2521      0.072      3.524      0.000
0.112       0.392
==============================================================================
===============
"""
```

0.0.14 Anexo 6.2: Resultados Modelo Poisson Corregido

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                Disease   No. Observations:                 2351
Model:                            GLM   Df Residuals:                     2336
Model Family:                 Poisson   Df Model:                           14
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3170.2
Date:                Mon, 22 Apr 2024   Deviance:                       1758.4
Time:                        20:53:12   Pearson chi2:                 1.39e+03
No. Iterations:                     5   Pseudo R-squ. (CS):             0.3570
Covariance Type:            nonrobust
==============================================================================
===============
                                 coef    std err          z      P>|z|
[0.025      0.975]
------------------------------------------------------------------------------
---------------
Glucose                       -0.5239      0.076     -6.930      0.000
-0.672      -0.376
Cholesterol                    0.7596      0.083      9.161      0.000
0.597       0.922
Hemoglobin                     0.9951      0.076     13.064      0.000
0.846       1.144
Platelets                     -1.3279      0.060    -21.966      0.000
-1.446      -1.209
Hematocrit                    -0.2774      0.067     -4.115      0.000
-0.410      -0.145
Mean Corpuscular Volume        0.9607      0.064     15.058      0.000
0.836       1.086
Mean Corpuscular Hemoglobin    0.7415      0.057     12.974      0.000
0.629       0.854
BMI                            0.8614      0.091      9.513      0.000
0.684       1.039
Systolic Blood Pressure       -0.3982      0.076     -5.268      0.000
-0.546      -0.250
Diastolic Blood Pressure      -0.3830      0.077     -4.950      0.000
-0.535      -0.231
Triglycerides                 -0.5262      0.077     -6.864      0.000
-0.676      -0.376
HbA1c                          0.4130      0.061      6.788      0.000
0.294       0.532
LDL Cholesterol               -0.8227      0.078    -10.601      0.000
-0.975      -0.671
```

```
HDL Cholesterol                -0.4957      0.060     -8.202      0.000
-0.614      -0.377
Heart Rate                      0.2940      0.070      4.187      0.000
0.156       0.432
==============================================================================
===============
"""
```

0.0.15 Anexo 7.1: Resultados Modelo Poisson para estimación de alpha.

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:                Disease   R-squared (uncentered):
0.345
Model:                            OLS   Adj. R-squared (uncentered):
0.345
Method:                 Least Squares   F-statistic:
1237.
Date:                Mon, 22 Apr 2024   Prob (F-statistic):
4.85e-218
Time:                        20:53:12   Log-Likelihood:
-2020.0
No. Observations:                2351   AIC:
4042.
Df Residuals:                    2350   BIC:
4048.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.2297      0.007    -35.167      0.000      -0.242      -0.217
==============================================================================
Omnibus:                      600.140   Durbin-Watson:                   1.317
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1352.614
Skew:                           1.434   Prob(JB):                    1.92e-294
Kurtosis:                       5.363   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.
```

"""

0.0.16 Anexo 8.1: Resultados Modelo Binomial Negativa

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 Disease   No. Observations:                 2351
Model:                             GLM   Df Residuals:                     2336
Model Family:        NegativeBinomial   Df Model:                           14
Link Function:                     Log   Scale:                          1.0000
Method:                           IRLS   Log-Likelihood:                -3714.3
Date:                Mon, 22 Apr 2024   Deviance:                       1043.2
Time:                        20:53:12   Pearson chi2:                     715.
No. Iterations:                     9   Pseudo R-squ. (CS):             0.1955
Covariance Type:            nonrobust
==============================================================================
===============
                                 coef    std err          z      P>|z|
[0.025      0.975]
------------------------------------------------------------------------------
---------------
Glucose                       -0.5336      0.117     -4.565      0.000
-0.763      -0.304
Cholesterol                    0.8422      0.126      6.687      0.000
0.595       1.089
Hemoglobin                     1.0373      0.115      9.008      0.000
0.812       1.263
Platelets                     -1.4827      0.092    -16.030      0.000
-1.664      -1.301
Hematocrit                    -0.4053      0.100     -4.040      0.000
-0.602      -0.209
Mean Corpuscular Volume        1.1668      0.096     12.148      0.000
0.979       1.355
Mean Corpuscular Hemoglobin    0.8386      0.089      9.459      0.000
0.665       1.012
BMI                            0.9082      0.133      6.833      0.000
0.648       1.169
Systolic Blood Pressure       -0.4173      0.117     -3.580      0.000
-0.646      -0.189
Diastolic Blood Pressure      -0.4239      0.116     -3.643      0.000
-0.652      -0.196
Triglycerides                 -0.6434      0.114     -5.664      0.000
-0.866      -0.421
HbA1c                          0.4960      0.097      5.120      0.000
0.306       0.686
```

```
LDL Cholesterol                -0.8609      0.119     -7.241      0.000
-1.094      -0.628
HDL Cholesterol                -0.6095      0.094     -6.506      0.000
-0.793      -0.426
Heart Rate                      0.3162      0.107      2.959      0.003
0.107       0.526
==============================================================================
==============
"""
```

0.0.17 Anexo 9.1: Comparación Coeficientes Modelos Poisson y Binomial Negativa

	Poisson: coef	NegBin: coef
Hemoglobin	0.9951	1.0373
Mean Corpuscular Volume	0.9607	1.1668
BMI	0.8614	0.9082
Cholesterol	0.7596	0.8422
Mean Corpuscular Hemoglobin	0.7415	0.8386
HbA1c	0.4130	0.4960
Heart Rate	0.2940	0.3162
Hematocrit	-0.2774	-0.4053
Diastolic Blood Pressure	-0.3830	-0.4239
Systolic Blood Pressure	-0.3982	-0.4173
HDL Cholesterol	-0.4957	-0.6095
Glucose	-0.5239	-0.5336
Triglycerides	-0.5262	-0.6434
LDL Cholesterol	-0.8227	-0.8609
Platelets	-1.3279	-1.4827