Tarea_1-LMAA

April 21, 2024

1 Variables de Estudio

- **Glucose (*Glu*)**: This is the level of glucose in the blood, measured in milligrams per deciliter (mg/dL)
- **Cholesterol (*Cho*)**: This is the level of cholesterol in the blood, measured in milligrams per deciliter (mg/dL)
- **Hemoglobin (*Hemo*)**: This is the protein in red blood cells that carries oxygen from the lungs to the rest of the body
- **Platelets (*Pla*)**: 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 (*Hema*)**: 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 (*Ins*)**: This is a hormone that helps regulate blood sugar levels
- **Body Mass Index (*BMI*))**: 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 (*Tri*)**: These are a type of fat found in the blood, measured in milligrams per deciliter (mg/dL)
- **Glycated Hemoglobin (*HbA1c*)**: This is a measure of average blood sugar levels over the past two to three months
- **Low-Density Lipoprotein Cholesterol (*LDL_Ch*)**: This is the "bad" cholesterol that can build up in the arteries
- **High-Density Lipoprotein Cholesterol (*HDL_Ch*)**: This is the "good" cholesterol that helps remove LDL cholesterol from the arteries
- **Heart Rate (*HR*)**: This is the number of heartbeats per minute (bpm)
- **Creatinine (*Cre*)**: 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)

** Nombres entre paréntesis indican como se les referirá durante el procedimiento de la tarea.*

2 Procedimiento y Respuestas

2.1 Pregunta 1:

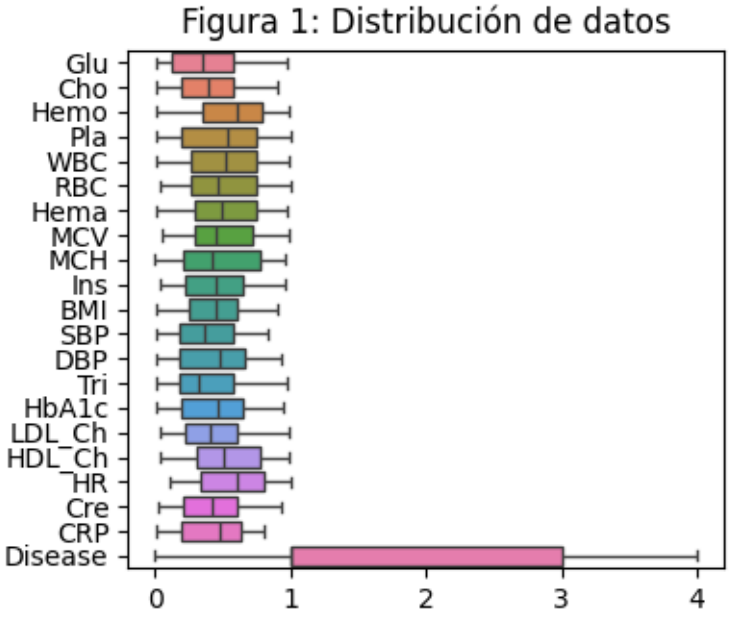
Cargar la base de datos* disease.csv *en el ambiente. Identifique los tipos de datos que se encuentran en la base, realice estadisticas descriptivas sobre las variables importantes (Hint: Revisar la distribuciones, datos faltantes, outliers, etc.) y limpie las variables cuando sea necesario.

- Se cargan los datos desde el dataset para observarlos inicialmente y se cambian los nombres de las columnas para una mayor facilidad de manejo (*Anexo A.1*).

Tamaño del DataFrame:

```
(2351, 21)
```

Distribuciones de datos:



→ Los datos se observan correctamente formateados, así como no se observan valores nulos para ninguna de las columnas (*Anexo A.2*).

→ Además, los datos se encuentran, aparentemente, correctamente distribuidos, sin presencia de *outliers* ni datos sin lógica dentro del estudio (*Figura 1*).

R1: Se ha importado el dataset 'disease.csv' exitosamente. Este no presenta datos faltantes en ninguna columna y, aparentemente, no presenta outliers ni distribuciones inadecuadas respecto a las variables de estudio. Por tanto, no será necesario un proceso de limpieza para el dataset y se trabajará con los datos tal cual se importaron (cambiando solo los nombres de las columnas). Cabe destacar que los datos se

encuentran estandarizados, tal que la unidad de medida de cada uno es una "unidad de distribución". Esto será utilizado para el análisis de los modelos posteriores.

2.2 Pregunta 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.

- Para realizar esto, se crea una columna "Disease_b" que generará un valor binario: 1 si la variable "Disease" es mayor que 0, 0 si no. En el *Anexo B.1* se comprueba la coherencia en la cantidad de datos.
- Además, antes de comenzar, se hace un análisis de la matriz de correlación (*Anexo B.2*) para revisar el comportamiento entre variables y la posibilidad de autocorrelación. Resulta no observarse mayores relaciones entre variables (no es relevante).
- Finalmente, se realiza un el modelo de MCO (*Anexo B.3*) obteniendo inicialmente las siguientes variables no significativas:

```
----> ['Systolic Blood Pressure', 'Creatinine', 'C-reactive Protein']
```

* ***Nota****: Dentro de esta tarea, se considerará como no significativas todas aquellas variables cuyo valor-p inidique que no son significativos al 95% (valor* $-p > 0.05$*) a menos de que se indique lo contrario.*

- Remodelando la regresión eliminando las primeras dos variables no significativas, la tercera se torna significativa, por lo que se realizará este proceso para la estimación por OLS (resultados en *Anexo B.4*)

* ***Nota Importante****: Como se menciona en R1, las unidades se encuentran estandarizadas a una unidad de distribución total. Para ello, la lógica utilizada para los análisis indicará que* ***cada movimiento en un decil de la distribución de la variable aumentará en (coef/10) por ciento la probabilidad de Y.***

R2: Se excluyen las variables 'Systolic Blood Pressure' y 'Creatinine' al ser no significativas para el modelo. En cuanto a la estimación, el modelo indica que la variable de mayor peso es el *Volumen Corpuscular Medio (MCV)*, indicando que un aumento en 1 decil dentro de la distribución de esta variable implicaría un aumento de un 6.25% de tener enfermedades, seguido del indicador de *colesterol (Cho)* con un 5.23% y un aumento del *ritmo cardiaco (HR)* con un 4.05%, todos bajo la misma lógica de un aumento en el decil. Además, destacar que, por otro lado, el aumento en 1 decil de distribución en la medición de *plaquetas (Pla)* y *glóbulos blancos (WBC)* implicarían una disminución en la probabilidad de estar enfermo en un 4.41% y 4.29% respectivamente.

2.3 Pregunta 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.

- Se aplica el modelo probit inicial con todas las variables (*Anexo C.1*) obteniendo las siguientes variables no significativas:

```
----> ['Systolic Blood Pressure']
```

- Dado que solo hay una variable no significativa, la estimación se hará restando solamente esta variable del modelo (resultados en *Anexo C.2*).

R3: Se excluye la variable de Systolic Blood Pressure por ser no significativa para el modelo. En cuanto a la estimación, el modelo entrega que las variables que más aumentan a la probabilidad de estar enfermo son el *colesterol (Cho)*, el *volumen corpuscular medio (MCV)* y el *ritmo cardíaco (HR)* con un 16%, 15.9% y 15.5% de aumento de probabilidad de estar enfermo ante el aumento en un decil dentro de la distribución de cada variable, respectivamente. Seguidos de estos se encuentra el *índice de masa corporal (BMI)* y la *proteína C-Reactiva (CRP)*, con un 12.8% y 10.9% respectivamente. Además, las variables que más disminuyen las probabilidades de estar enfermo respecto al aumento en un decil de la distribución son la *glucosa (Glu)*, las *plaquetas (Pla)*, los *triglicéridos (Tri)* y los *glóbulos rojos (RBC) y blancos (WBC)*, con un considerable 18.9%, 16.4%, 15.5%, 14.4% y 13.4% de disminución de probabilidades de tener enfermedades, respectivamente. Cabe destacar que no se considera muy correcto decir que al aumentar de decil en los niveles de glucosa y/o triglicéridos impliquen una disminución de posibilidades de estar enfermo, por lo que estas variables requerirían un mayor estudio.

2.4 Pregunta 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.

- Se aplica el modelo logit inicial con todas las variables (*Anexo D.1*) obteniendo las siguientes variables no significativas:

```
----> ['Systolic Blood Pressure']
```

- Nuevamente, para este modelo solo se descartará esta única variable (Resultados en *Anexo D.2*).

R4: Los resultados son virtualmente idénticos a los obtenidos en el modelo Probit, con mismas variables no significativas, coeficientes similares y dispersión similares, con variaciones entre menores e insignificantes.

2.5 Pregunta 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?

Tabla 1: Tabla comparativa entre los resultados de los modelos de OLS, Probit y Logit.

	OLS Coef	Prob Coef	Log Coef	OLS Val-p	Prob Val-p	Log Val-p
BMI	0.3429	1.2777	1.2590	0.000	0.000	0.000
CRP	0.0660	1.0925	1.0942	0.041	0.000	0.000
Cho	0.5226	1.5975	1.5552	0.000	0.000	0.000
Cre	NaN	-0.7216	-0.7125	NaN	0.000	0.000
DBP	-0.1286	-0.4824	-0.4424	0.000	0.000	0.000
Glu	-0.3636	-1.8943	-1.8939	0.000	0.000	0.000
HDL_Ch	-0.1389	0.3289	0.2857	0.000	0.000	0.001
HR	0.4053	1.5520	1.5276	0.000	0.000	0.000
HbA1c	0.1241	-0.2912	-0.3097	0.000	0.000	0.000
Hema	-0.2269	-0.9822	-0.9335	0.000	0.000	0.000
Hemo	0.2003	0.3264	0.3164	0.000	0.000	0.000
Ins	-0.1040	0.5157	0.4976	0.001	0.001	0.000
LDL_Ch	-0.1964	-0.2028	-0.1988	0.000	0.039	0.028
MCH	0.2926	0.4613	0.4452	0.000	0.000	0.000
MCV	0.6254	1.5858	1.5505	0.000	0.000	0.000
Pla	-0.4415	-1.6376	-1.5886	0.000	0.000	0.000
RBC	-0.3744	-1.4369	-1.3883	0.000	0.000	0.000
Tri	-0.2603	-1.5496	-1.4966	0.000	0.000	0.000
WBC	-0.4286	-1.3409	-1.2474	0.000	0.000	0.000
const	0.7573	NaN	NaN	0.000	NaN	NaN

* **Glosario**: - coef: coeficientes obtenidos (β_i) - val-p: p-value obtenido. - OLS: correspondiente a modelo por MCO. - Prob: correspondiente a modelo Probit. - Log: correspondiente a modelo Logit.

* Se rellena con NaN aquellos valores que no son significativos para la especificación.

R5: A partir de los resultados se puede notar como entre el modelo OLS y los modelos probit y logit hay una diferencia importante, desde las variables que son significativas para la especificación, hasta los mismo coeficientes de cambio marginal en la variable estimada, los cuales para el modelo OLS son sustancialmente menores, aunque más precisos (errores estándar menores). Esto era esperado, puesto que el estudio no cumple con el supuesto de linealidad para que el modelo OLS entregue los mejores resultados, por lo que se consideran más confiables los resultados obtenidos por los modelos Probit y Logit, los cuales son similares entre ambos y podría considerarse cualquiera de los resultados entre ambos. En cuanto al estudio de variables, se puede notar que las variables que no fueron robustas (fueron no significativas para alguno de los modelos) fueron la creatinina (Cre), la proteína C-reactiva y, con aún menor robustez, la presión sanguínea sistólica (la cual no fue significativa para ningún

modelo), por tanto, todo el resto de variables podrían considerarse robustas a la especificación.

2.6 Pregunta 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.

- Se aplicará el modelo de poisson sobre las variables, pero esta vez se cambiará la variable dependiente "y" a la variable categórica original, obteniendo los resultados del *Anexo E.1.* Se obtienen las siguientes variables no significativas.

```
----> ['White Blood Cells', 'Red Blood Cells']
```

→ Se eliminarán ambas variables, pues ambas resultan ser no significativas aunque se eliminen por separado.

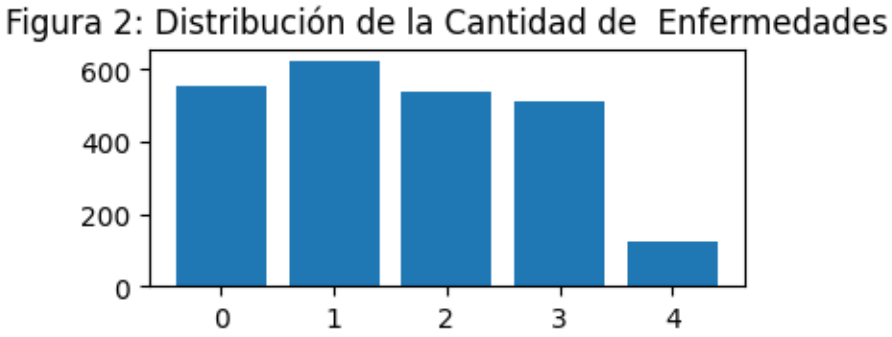
- Por tanto, se re modelará la regresión esta vez descartando lasa variables WBC y RBC, obteniendo los resultados observables en el *Anexo E.2.*

R6: Se elimina las variables de White Blood Cells y Red Blood Cells por ser no significativas para la especificación. Respecto a los resultados de la estimación, en este caso los coeficientes indican cuanto aumentaría la cantidad de enfermedades ante el aumento en una unidad de distribución, por tanto, las variables más importantes serían el *índice de masa corporal (BMI)*, implicando un aumento en 0.14 la cantidad de enfermedades por cada incremento en el decil de la distribución, luego la *hemoglobina (Hem)*, con un 0.12, y luego el *colesterol (Cho)* y el *volumen corpuscular medio*, con un 0.11 y 0.10 respectivamente. Además, el incremento de un decil en la cantidad de plaquetas implicaría una disminución de 0.12 la cantidad de enfermedades.

2.7 Pregunta 7

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

- Primero, se puede observar la distribución de la variable categórica "Disease":



→ Dada la forma de la distribución, no se puede observar una evidente dispersión, pero si una forma asimétrica de la densidad de los datos.

- Se realiza un test de sobredispersión a partir de nuestro modelo de poisson obteniendo los resultados observables en el *Anexo F.1*.

```
----> Resumen del test: Coeficiente: -0.24699
                        Valor-p:      0.00000 - Sign.
                        Valor alpha:  0.78115
```

R7: Dado que el coeficiente obtenido en el test de sobredispersión es significativo y distinto de 0, esto implica que existe sobre dispersión, aunque no muy grande. Por tanto, conviene utilizar el modelo binomial negativo con el valor alpha obtenido.

2.8 Pregunta 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.

- Se aplica un modelo binomial negativo especificando el valor de alpha obtenido por medio del test de sobredispersión (*Anexo G.1*).

```
----> ['White Blood Cells', 'Red Blood Cells', 'Insulin']
```

- Se realiza el modelo eliminando las primeras dos variables no significativas, pues la variable insulina resulta ser significativa al eliminar las otras 2 (resultados en *Anexo G.2*)

R8: Este modelo entrega resultados muy similares a los del modelo Poisson, con las mismas variables significativas y valores similares (aunque más intenso en casi todos los casos). Además, este modelo presenta valores más dispersos (error estándar mayor en tabla *Anexo G.2*).

2.9 Pregunta 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?

Tabla 2: Tabla comparativa entre los resultados de los modelos de Poisson y Binomial Negativo.

	Pois Coef	NegB Coef	Pois Std	NegB Std	Pois Val-p	NegB Val-p
const	-1.0753	-1.1536	0.163398	0.241546	0.000	0.000

Glu	-0.7476	-0.7014	0.085513	0.132321	0.000	0.000
Cho	1.0081	1.1080	0.092404	0.136813	0.000	0.000
Hemo	1.2169	1.1769	0.084613	0.123765	0.000	0.000
Pla	-1.1994	-1.3523	0.064520	0.099962	0.000	0.000
Hema	-0.2341	-0.3334	0.071718	0.105778	0.001	0.002
MCV	1.0664	1.2706	0.072506	0.107569	0.000	0.000
MCH	0.8478	0.9293	0.064078	0.096814	0.000	0.000
Ins	0.2456	0.2838	0.085696	0.128666	0.004	0.027
BMI	1.4396	1.4345	0.115061	0.161237	0.000	0.000
SBP	-0.2663	-0.2677	0.080741	0.122822	0.001	0.029
DBP	-0.7338	-0.7643	0.090969	0.133404	0.000	0.000
Tri	-0.5043	-0.6218	0.080921	0.119550	0.000	0.000
HbA1c	0.2859	0.3830	0.068696	0.106732	0.000	0.000
LDL_Ch	-0.8976	-0.8789	0.081415	0.122627	0.000	0.000
HDL_Ch	-0.2620	-0.3904	0.071334	0.109981	0.000	0.000
HR	0.2054	0.3105	0.075535	0.115056	0.007	0.007
Cre	0.5236	0.5727	0.087574	0.131906	0.000	0.000
CRP	0.5400	0.5354	0.082422	0.127854	0.000	0.000

* **Glosario**: - coef: coeficientes obtenidos (β_i) - val-p: p-value obtenido. - Std: error estándar obtenido. - Pois: correspondiente a modelo Poisson. - NegB: correspondiente a modelo Binomial Negativo.

R9: Entre ambos modelos resulta haber pequeñas diferencias, resultando en mismas variables significativas (todas robustas exceptuando los glóbulos blancos y rojos) y valores de coeficientes similares, teniendo algunas diferencias que varían entre lo insignificante y lo moderadamente importante. Dada la presencia de sobredispersión, debería ser más confiable el modelo entregado por la regresión binomial negativa, pero puesto que la sobredispersión no es muy alta, y además de que los resultados entregados por este modelo resultan ser, en general, más dispersos, dependería meramente del problema cual modelo llegase a ser más útil, pues ambos entregan información relevante.

---

3 Anexos

3.1 Anexo A - Anexos para pregunta 1:

3.1.1 A.1) Glosario de Columnas

Por motivos de mejorar el formato de despliegue de las variables en función de la tarea, se cambiaron los nombres de las columnas a nombres mas contenidos.

A continuación se presenta el glosario de nombres:

	Nombre Original	Nombre Nuevo
0	Glucose	Glu
1	Cholesterol	Cho
2	Hemoglobin	Hemo
3	Platelets	Pla
4	White Blood Cells	WBC
5	Red Blood Cells	RBC
6	Hematocrit	Hema
7	Mean Corpuscular Volume	MCV
8	Mean Corpuscular Hemoglobin	MCH
9	Insulin	Ins
10	BMI	BMI
11	Systolic Blood Pressure	SBP
12	Diastolic Blood Pressure	DBP
13	Triglycerides	Tri
14	HbA1c	HbA1c
15	LDL Cholesterol	LDL_Ch
16	HDL Cholesterol	HDL_Ch
17	Heart Rate	HR
18	Creatinine	Cre
19	C-reactive Protein	CRP
20	Disease	Disease

3.1.2 A.2) Información sobre el DataFrame

Utilizado para descubrir la presencia de valores nulos o malos formatos.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2351 entries, 0 to 2350
Data columns (total 21 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   Glu      2351 non-null   float64
 1   Cho      2351 non-null   float64
 2   Hemo     2351 non-null   float64
 3   Pla      2351 non-null   float64
 4   WBC      2351 non-null   float64
 5   RBC      2351 non-null   float64
 6   Hema     2351 non-null   float64
```

```
 7   MCV      2351 non-null   float64
 8   MCH      2351 non-null   float64
 9   Ins      2351 non-null   float64
 10  BMI      2351 non-null   float64
 11  SBP      2351 non-null   float64
 12  DBP      2351 non-null   float64
 13  Tri      2351 non-null   float64
 14  HbA1c    2351 non-null   float64
 15  LDL_Ch   2351 non-null   float64
 16  HDL_Ch   2351 non-null   float64
 17  HR       2351 non-null   float64
 18  Cre      2351 non-null   float64
 19  CRP      2351 non-null   float64
 20  Disease  2351 non-null   int64
dtypes: float64(20), int64(1)
memory usage: 385.8 KB
```

→ Todas las variables tienen una cantidad de valores no nulos coincidentes con la cantidad de filas, por lo que no hay valores nulos en el dataframe.

→ Además, todas las variables están en formato "float" para valores decimales e "int" para valores enteros, por lo que están correctamente formateadas las variables.

3.2 Anexo B - Anexos para pregunta 2:

3.2.1 B.1) Comprobando asignación de valores binarios en el dataframe

Se crea una tabla con la cantidad de valores asignados en las columnas de disease categórica y binaria para poder comprobar si la asignación de datos fue correcta (de ser así, la cantidad de valores 1 en la columna binaria debería coincidir con la cantidad de valores distintos de 0 en la categórica).

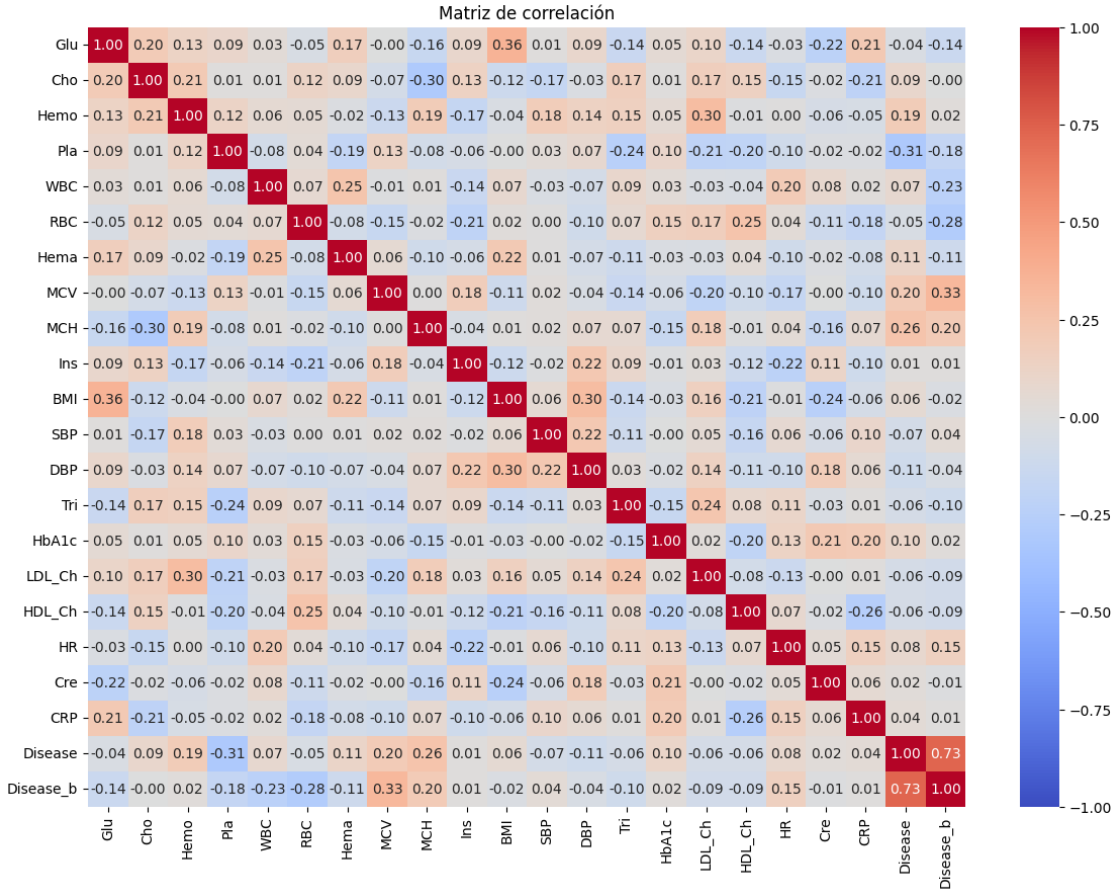
- Tabla comparativa:

	Disease Categórico	Disease Binario
0	556	556
1	623	1795
2	540	0
3	509	0
4	123	0

→ La cantidad de datos de valor 1 en la columna binaria coincide con la suma de las variables distintas de 0 en la columna categórica, por lo que se asume una correcta asignación a la columna binaria.

3.2.2 B.2) Observando la corrrelación entre las variables de estudio

Se hace uso de un mapa de calor a partir de la matriz de correlación de los datos para poder observar la presencia de autocorrelación entre variables.



→ Se observa, en general, baja autocorrelación entre las variables de estudio.

3.2.3 B.3) Regresión OLS inicial con todas las variables.

Dep. Variable:	Disease_b	**R-squared:**	0.488
Model:	OLS	**Adj. R-squared:**	0.483
Method:	Least Squares	**F-statistic:**	110.9
Date:	Sun, 21 Apr 2024	**Prob (F-statistic):**	8.97e-320
Time:	22:55:44	**Log-Likelihood:**	-537.89
No. Observations:	2351	**AIC:**	1118.
Df Residuals:	2330	**BIC:**	1239.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
const	0.7275	0.060	12.119	0.000	0.610	0.845
Glu	-0.3551	0.033	-10.865	0.000	-0.419	-0.291
Cho	0.5331	0.033	16.149	0.000	0.468	0.598
Hemo	0.1909	0.029	6.548	0.000	0.134	0.248
Pla	-0.4391	0.024	-18.067	0.000	-0.487	-0.391
WBC	-0.4294	0.025	-17.187	0.000	-0.478	-0.380
RBC	-0.3768	0.028	-13.617	0.000	-0.431	-0.323
Hema	-0.2309	0.025	-9.083	0.000	-0.281	-0.181
MCV	0.6241	0.025	24.873	0.000	0.575	0.673
MCH	0.3015	0.024	12.725	0.000	0.255	0.348
Ins	-0.1071	0.032	-3.356	0.001	-0.170	-0.045
BMI	0.3538	0.037	9.692	0.000	0.282	0.425
SBP	0.0488	0.030	1.611	0.107	-0.011	0.108
DBP	-0.1461	0.032	-4.622	0.000	-0.208	-0.084
Tri	-0.2501	0.029	-8.611	0.000	-0.307	-0.193
HbA1c	0.1247	0.028	4.528	0.000	0.071	0.179
LDL_Ch	-0.2013	0.031	-6.550	0.000	-0.262	-0.141
HDL_Ch	-0.1323	0.028	-4.649	0.000	-0.188	-0.076
HR	0.3981	0.029	13.915	0.000	0.342	0.454
Cre	0.0293	0.033	0.893	0.372	-0.035	0.094
CRP	0.0621	0.032	1.925	0.054	-0.001	0.125

Omnibus:	119.055	**Durbin-Watson:**	0.963
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	131.761
Skew:	-0.563	**Prob(JB):**	2.45e-29
Kurtosis:	2.722	**Cond. No.**	26.5

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

3.2.4 B.4) Regresión OLS Corregida

Dep. Variable:	Disease_b	**R-squared:**	0.487
Model:	OLS	**Adj. R-squared:**	0.483
Method:	Least Squares	**F-statistic:**	123.0
Date:	Sun, 21 Apr 2024	**Prob (F-statistic):**	3.30e-321
Time:	22:55:44	**Log-Likelihood:**	-539.42
No. Observations:	2351	**AIC:**	1117.
Df Residuals:	2332	**BIC:**	1226.
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
const	0.7573	0.057	13.351	0.000	0.646	0.869
Glu	-0.3636	0.032	-11.352	0.000	-0.426	-0.301
Cho	0.5226	0.032	16.098	0.000	0.459	0.586
Hemo	0.2003	0.028	7.077	0.000	0.145	0.256
Pla	-0.4415	0.024	-18.192	0.000	-0.489	-0.394
WBC	-0.4286	0.025	-17.359	0.000	-0.477	-0.380
RBC	-0.3744	0.027	-13.722	0.000	-0.428	-0.321
Hema	-0.2269	0.025	-8.962	0.000	-0.277	-0.177
MCV	0.6254	0.025	24.952	0.000	0.576	0.675
MCH	0.2926	0.023	12.762	0.000	0.248	0.338
Ins	-0.1040	0.032	-3.262	0.001	-0.166	-0.041
BMI	0.3429	0.035	9.761	0.000	0.274	0.412
DBP	-0.1286	0.029	-4.366	0.000	-0.186	-0.071
Tri	-0.2603	0.028	-9.150	0.000	-0.316	-0.205
HbA1c	0.1241	0.027	4.608	0.000	0.071	0.177
LDL_Ch	-0.1964	0.030	-6.457	0.000	-0.256	-0.137
HDL_Ch	-0.1389	0.028	-4.935	0.000	-0.194	-0.084
HR	0.4053	0.028	14.307	0.000	0.350	0.461
CRP	0.0660	0.032	2.049	0.041	0.003	0.129

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.**	24.5

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

3.3 Anexo C - Anexos para pregunta 3:

3.3.1 C.1) Regresión Probit inicial con todas las variables

Dep. Variable:	Disease_b	**No. Observations:**	2351
Model:	Probit	**Df Residuals:**	2330
Method:	MLE	**Df Model:**	20
Date:	Sun, 21 Apr 2024	**Pseudo R-squ.:**	0.7259
Time:	22:55:44	**Log-Likelihood:**	-352.45
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	14.0190	2.858	4.905	0.000	8.417	19.621
Glu	-23.9767	4.132	-5.803	0.000	-32.075	-15.878
Cho	19.2898	3.374	5.717	0.000	12.677	25.903
Hemo	4.6403	1.057	4.391	0.000	2.569	6.712
Pla	-19.9810	3.123	-6.398	0.000	-26.102	-13.860
WBC	-15.8611	3.014	-5.262	0.000	-21.769	-9.953
RBC	-17.5714	2.671	-6.580	0.000	-22.806	-12.337
Hema	-11.6574	1.935	-6.024	0.000	-15.450	-7.864
MCV	19.4605	2.916	6.673	0.000	13.744	25.177
MCH	5.3667	0.668	8.039	0.000	4.058	6.675
Ins	6.5925	1.898	3.473	0.001	2.872	10.312
BMI	15.9443	2.995	5.324	0.000	10.075	21.814
SBP	-0.8221	0.525	-1.565	0.118	-1.852	0.208
DBP	-5.7089	1.213	-4.707	0.000	-8.086	-3.332
Tri	-19.2673	3.614	-5.331	0.000	-26.351	-12.184
HbA1c	-3.8744	1.025	-3.781	0.000	-5.883	-1.866
LDL_Ch	-2.4780	1.194	-2.076	0.038	-4.818	-0.138
HDL_Ch	4.3352	1.081	4.009	0.000	2.216	6.455
HR	19.0415	3.218	5.917	0.000	12.734	25.349
Cre	-9.3072	1.950	-4.773	0.000	-13.129	-5.485
CRP	14.1758	2.530	5.604	0.000	9.218	19.134

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.

3.3.2 C.1) Regresión Probit corregida.

Dep. Variable:	Disease_b	**No. Observations:**	2351
Model:	Probit	**Df Residuals:**	2331
Method:	MLE	**Df Model:**	19
Date:	Sun, 21 Apr 2024	**Pseudo R-squ.:**	0.7251
Time:	22:55:44	**Log-Likelihood:**	-353.55
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	13.9419	2.756	5.059	0.000	8.540	19.343
Glu	-22.1894	3.646	-6.087	0.000	-29.335	-15.044
Cho	18.7121	3.186	5.873	0.000	12.468	24.956
Hemo	3.8230	0.899	4.251	0.000	2.060	5.586
Pla	-19.1819	2.864	-6.697	0.000	-24.796	-13.568
WBC	-15.7065	2.883	-5.447	0.000	-21.358	-10.055
RBC	-16.8318	2.416	-6.967	0.000	-21.567	-12.097
Hema	-11.5053	1.844	-6.241	0.000	-15.119	-7.892
MCV	18.5752	2.649	7.013	0.000	13.384	23.766
MCH	5.4038	0.753	7.177	0.000	3.928	6.880
Ins	6.0405	1.786	3.383	0.001	2.541	9.540
BMI	14.9659	2.706	5.530	0.000	9.661	20.270
DBP	-5.6502	1.172	-4.823	0.000	-7.946	-3.354
Tri	-18.1512	3.277	-5.539	0.000	-24.574	-11.728
HbA1c	-3.4115	0.983	-3.470	0.001	-5.339	-1.484
LDL_Ch	-2.3755	1.160	-2.047	0.041	-4.650	-0.101
HDL_Ch	3.8528	1.123	3.430	0.001	1.651	6.054
HR	18.1793	2.949	6.164	0.000	12.399	23.960
Cre	-8.4522	1.776	-4.760	0.000	-11.932	-4.972
CRP	12.7969	2.224	5.753	0.000	8.437	17.157

Possibly complete quasi-separation: A fraction 0.64 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

- **Efectos marginales**:

Dep. Variable:	Disease_b
Method:	dydx
At:	overall

dy/dx	std err	z	P> |z|	[0.025	0.975]

Glu	-1.8943	0.292	-6.497	0.000	-2.466	-1.323
Cho	1.5975	0.256	6.249	0.000	1.096	2.099
Hemo	0.3264	0.074	4.401	0.000	0.181	0.472
Pla	-1.6376	0.224	-7.306	0.000	-2.077	-1.198
WBC	-1.3409	0.232	-5.784	0.000	-1.795	-0.886
RBC	-1.4369	0.187	-7.675	0.000	-1.804	-1.070
Hema	-0.9822	0.146	-6.727	0.000	-1.268	-0.696
MCV	1.5858	0.206	7.709	0.000	1.183	1.989
MCH	0.4613	0.058	7.949	0.000	0.348	0.575
Ins	0.5157	0.149	3.455	0.001	0.223	0.808
BMI	1.2777	0.219	5.835	0.000	0.849	1.707
DBP	-0.4824	0.096	-5.046	0.000	-0.670	-0.295
Tri	-1.5496	0.264	-5.869	0.000	-2.067	-1.032
HbA1c	-0.2912	0.082	-3.540	0.000	-0.453	-0.130
LDL_Ch	-0.2028	0.098	-2.060	0.039	-0.396	-0.010
HDL_Ch	0.3289	0.093	3.523	0.000	0.146	0.512
HR	1.5520	0.235	6.616	0.000	1.092	2.012
Cre	-0.7216	0.146	-4.956	0.000	-1.007	-0.436
CRP	1.0925	0.179	6.098	0.000	0.741	1.444

3.4 Anexo D - Anexos para pregunta 4:

3.4.1 D.1) Regresión Logit inicial con todas las variables

Dep. Variable:	Disease_b	**No. Observations:**	2351
Model:	Logit	**Df Residuals:**	2330
Method:	MLE	**Df Model:**	20
Date:	Sun, 21 Apr 2024	**Pseudo R-squ.:**	0.7289
Time:	22:55:44	**Log-Likelihood:**	-348.67
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	22.9505	4.382	5.237	0.000	14.362	31.539
Glu	-41.9756	6.959	-6.032	0.000	-55.615	-28.336
Cho	33.0255	5.643	5.853	0.000	21.966	44.085
Hemo	7.7796	1.733	4.488	0.000	4.382	11.177
Pla	-34.0071	5.092	-6.679	0.000	-43.986	-24.028
WBC	-25.8608	4.672	-5.535	0.000	-35.019	-16.703
RBC	-29.7860	4.347	-6.852	0.000	-38.306	-21.266
Hema	-19.5104	3.206	-6.086	0.000	-25.794	-13.227
MCV	33.4290	4.872	6.862	0.000	23.880	42.978
MCH	9.1377	1.222	7.478	0.000	6.743	11.533
Ins	11.1286	3.075	3.619	0.000	5.102	17.155
BMI	27.6083	4.983	5.540	0.000	17.842	37.375
SBP	-1.2499	0.935	-1.337	0.181	-3.082	0.583
DBP	-9.2059	1.996	-4.612	0.000	-13.118	-5.294
Tri	-32.6178	5.822	-5.602	0.000	-44.030	-21.206
HbA1c	-7.1262	1.740	-4.096	0.000	-10.536	-3.716
LDL_Ch	-4.2427	1.902	-2.230	0.026	-7.971	-0.514
HDL_Ch	6.6977	1.745	3.838	0.000	3.277	10.118
HR	32.8277	5.209	6.302	0.000	22.619	43.037
Cre	-16.1228	3.239	-4.978	0.000	-22.471	-9.775
CRP	24.8677	4.318	5.759	0.000	16.405	33.330

Possibly complete quasi-separation: A fraction 0.64 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

3.4.2 D.2) Regresión Logit corregida.

Dep. Variable:	Disease_b	**No. Observations:**	2351
Model:	Logit	**Df Residuals:**	2331
Method:	MLE	**Df Model:**	19
Date:	Sun, 21 Apr 2024	**Pseudo R-squ.:**	0.7282
Time:	22:55:44	**Log-Likelihood:**	-349.48
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	23.0126	4.322	5.324	0.000	14.541	31.485
Glu	-39.1052	6.097	-6.414	0.000	-51.055	-27.156
Cho	32.1121	5.351	6.001	0.000	21.624	42.601
Hemo	6.5327	1.425	4.583	0.000	3.739	9.326
Pla	-32.8016	4.710	-6.965	0.000	-42.033	-23.571
WBC	-25.7556	4.577	-5.627	0.000	-34.726	-16.785
RBC	-28.6654	3.954	-7.250	0.000	-36.414	-20.916
Hema	-19.2748	3.081	-6.255	0.000	-25.314	-13.235
MCV	32.0141	4.435	7.218	0.000	23.321	40.707
MCH	9.1933	1.306	7.041	0.000	6.634	11.753
Ins	10.2743	2.890	3.555	0.000	4.609	15.939
BMI	25.9970	4.494	5.785	0.000	17.189	34.805
DBP	-9.1351	1.950	-4.685	0.000	-12.956	-5.314
Tri	-30.9016	5.295	-5.836	0.000	-41.280	-20.523
HbA1c	-6.3946	1.636	-3.908	0.000	-9.602	-3.188
LDL_Ch	-4.1051	1.878	-2.186	0.029	-7.786	-0.424
HDL_Ch	5.9001	1.740	3.392	0.001	2.491	9.310
HR	31.5420	4.817	6.549	0.000	22.102	40.982
Cre	-14.7123	2.881	-5.107	0.000	-20.359	-9.066
CRP	22.5926	3.702	6.103	0.000	15.337	29.848

Possibly complete quasi-separation: A fraction 0.64 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

- **Efectos marginales**:

Dep. Variable:	Disease_b
Method:	dydx
At:	overall

dy/dx	std err	z	P> |z|	[0.025	0.975]

Glu	-1.8939	0.276	-6.872	0.000	-2.434	-1.354
Cho	1.5552	0.246	6.325	0.000	1.073	2.037
Hemo	0.3164	0.067	4.727	0.000	0.185	0.448
Pla	-1.5886	0.211	-7.545	0.000	-2.001	-1.176
WBC	-1.2474	0.212	-5.872	0.000	-1.664	-0.831
RBC	-1.3883	0.176	-7.871	0.000	-1.734	-1.043
Hema	-0.9335	0.141	-6.622	0.000	-1.210	-0.657
MCV	1.5505	0.196	7.921	0.000	1.167	1.934
MCH	0.4452	0.058	7.691	0.000	0.332	0.559
Ins	0.4976	0.137	3.630	0.000	0.229	0.766
BMI	1.2590	0.206	6.102	0.000	0.855	1.663
DBP	-0.4424	0.092	-4.834	0.000	-0.622	-0.263
Tri	-1.4966	0.244	-6.140	0.000	-1.974	-1.019
HbA1c	-0.3097	0.077	-4.023	0.000	-0.461	-0.159
LDL_Ch	-0.1988	0.091	-2.191	0.028	-0.377	-0.021
HDL_Ch	0.2857	0.083	3.450	0.001	0.123	0.448
HR	1.5276	0.218	7.008	0.000	1.100	1.955
Cre	-0.7125	0.134	-5.324	0.000	-0.975	-0.450
CRP	1.0942	0.168	6.529	0.000	0.766	1.423

3.5 Anexo E - Anexos para pregunta 6:

3.5.1 E.1) Modelo Poisson inicial con todas las variables

Dep. Variable:	Disease	**No. Observations:**	2351
Model:	GLM	**Df Residuals:**	2330
Model Family:	Poisson	**Df Model:**	20
Link Function:	Log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3128.9
Date:	Sun, 21 Apr 2024	**Deviance:**	1675.8
Time:	22:55:44	**Pearson chi2:**	1.29e+03
No. Iterations:	5	**Pseudo R-squ. (CS):**	0.3792
Covariance Type:	nonrobust		

	coef	std err	z	P> |z|	[0.025	0.975]
const	-1.0973	0.167	-6.572	0.000	-1.425	-0.770
Glu	-0.7412	0.086	-8.654	0.000	-0.909	-0.573
Cho	1.0060	0.093	10.818	0.000	0.824	1.188
Hemo	1.2253	0.086	14.227	0.000	1.057	1.394
Pla	-1.2022	0.065	-18.613	0.000	-1.329	-1.076
WBC	-0.0481	0.069	-0.702	0.482	-0.182	0.086
RBC	0.0726	0.077	0.943	0.345	-0.078	0.224
Hema	-0.2172	0.074	-2.928	0.003	-0.363	-0.072
MCV	1.0621	0.073	14.639	0.000	0.920	1.204
MCH	0.8519	0.065	13.120	0.000	0.725	0.979
Ins	0.2609	0.088	2.959	0.003	0.088	0.434
BMI	1.4450	0.115	12.513	0.000	1.219	1.671
SBP	-0.2603	0.081	-3.213	0.001	-0.419	-0.102
DBP	-0.7371	0.091	-8.073	0.000	-0.916	-0.558
Tri	-0.5094	0.082	-6.178	0.000	-0.671	-0.348
HbA1c	0.2607	0.074	3.547	0.000	0.117	0.405
LDL_Ch	-0.9188	0.084	-10.991	0.000	-1.083	-0.755
HDL_Ch	-0.2771	0.073	-3.793	0.000	-0.420	-0.134
HR	0.2113	0.076	2.780	0.005	0.062	0.360
Cre	0.5466	0.091	6.031	0.000	0.369	0.724
CRP	0.5485	0.083	6.625	0.000	0.386	0.711

3.5.2 E.2) Modelo Poisson corregido.

Dep. Variable:	Disease	**No. Observations:**	2351
Model:	GLM	**Df Residuals:**	2332
Model Family:	Poisson	**Df Model:**	18
Link Function:	Log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3129.5
Date:	Sun, 21 Apr 2024	**Deviance:**	1677.1
Time:	22:55:44	**Pearson chi2:**	1.30e+03
No. Iterations:	5	**Pseudo R-squ. (CS):**	0.3788
Covariance Type:	nonrobust		

	coef	std err	z	P> |z|	[0.025	0.975]
const	-1.0753	0.163	-6.581	0.000	-1.396	-0.755
Glu	-0.7476	0.086	-8.743	0.000	-0.915	-0.580
Cho	1.0081	0.092	10.910	0.000	0.827	1.189
Hemo	1.2169	0.085	14.382	0.000	1.051	1.383
Pla	-1.1994	0.065	-18.590	0.000	-1.326	-1.073
Hema	-0.2341	0.072	-3.264	0.001	-0.375	-0.094
MCV	1.0664	0.073	14.707	0.000	0.924	1.208
MCH	0.8478	0.064	13.231	0.000	0.722	0.973
Ins	0.2456	0.086	2.866	0.004	0.078	0.414
BMI	1.4396	0.115	12.512	0.000	1.214	1.665
SBP	-0.2663	0.081	-3.298	0.001	-0.425	-0.108
DBP	-0.7338	0.091	-8.066	0.000	-0.912	-0.555
Tri	-0.5043	0.081	-6.231	0.000	-0.663	-0.346
HbA1c	0.2859	0.069	4.162	0.000	0.151	0.421
LDL_Ch	-0.8976	0.081	-11.025	0.000	-1.057	-0.738
HDL_Ch	-0.2620	0.071	-3.673	0.000	-0.402	-0.122
HR	0.2054	0.076	2.720	0.007	0.057	0.353
Cre	0.5236	0.088	5.979	0.000	0.352	0.695
CRP	0.5400	0.082	6.551	0.000	0.378	0.702

3.6 Anexo F - Anexos para pregunta 7:

3.6.1 F.1) Resultados de test de sobredispersión

Dep. Variable:	Disease	**R-squared (uncentered):**	0.389
Model:	OLS	**Adj. R-squared (uncentered):**	0.389
Method:	Least Squares	**F-statistic:**	1499.
Date:	Sun, 21 Apr 2024	**Prob (F-statistic):**	4.87e-254
Time:	22:55:44	**Log-Likelihood:**	-1975.3
No. Observations:	2351	**AIC:**	3953.
Df Residuals:	2350	**BIC:**	3958.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
x1	-0.2470	0.006	-38.712	0.000	-0.260	-0.234

Omnibus:	622.846	**Durbin-Watson:**	1.472
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1385.165
Skew:	1.502	**Prob(JB):**	1.64e-301
Kurtosis:	5.263	**Cond. No.**	1.00

Notes:
[1] R^2 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.

3.7 Anexo G - Anexos para pregunta 8:

3.7.1 G.1) Modelo Binomial Negativo inicial con todas las variables

Dep. Variable:	Disease	**No. Observations:**	2351
Model:	GLM	**Df Residuals:**	2330
Model Family:	NegativeBinomial	**Df Model:**	20
Link Function:	Log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3688.8
Date:	Sun, 21 Apr 2024	**Deviance:**	1006.1
Time:	22:55:44	**Pearson chi2:**	687.
No. Iterations:	10	**Pseudo R-squ. (CS):**	0.2104
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-1.1068	0.246	-4.499	0.000	-1.589	-0.625
Glu	-0.7116	0.132	-5.381	0.000	-0.971	-0.452
Cho	1.1378	0.137	8.275	0.000	0.868	1.407
Hemo	1.1645	0.124	9.367	0.000	0.921	1.408
Pla	-1.3526	0.100	-13.546	0.000	-1.548	-1.157
WBC	-0.1654	0.102	-1.615	0.106	-0.366	0.035
RBC	-0.0610	0.115	-0.529	0.597	-0.287	0.165
Hema	-0.3009	0.110	-2.730	0.006	-0.517	-0.085
MCV	1.2804	0.108	11.899	0.000	1.070	1.491
MCH	0.9520	0.097	9.769	0.000	0.761	1.143
Ins	0.2418	0.132	1.829	0.067	-0.017	0.501
BMI	1.4706	0.161	9.111	0.000	1.154	1.787
SBP	-0.2449	0.123	-1.995	0.046	-0.485	-0.004
DBP	-0.7894	0.134	-5.899	0.000	-1.052	-0.527
Tri	-0.6016	0.121	-4.986	0.000	-0.838	-0.365
HbA1c	0.3966	0.113	3.506	0.000	0.175	0.618
LDL_Ch	-0.8886	0.125	-7.092	0.000	-1.134	-0.643
HDL_Ch	-0.3805	0.113	-3.381	0.001	-0.601	-0.160
HR	0.3395	0.116	2.919	0.004	0.112	0.567
Cre	0.5963	0.135	4.426	0.000	0.332	0.860
CRP	0.5305	0.129	4.108	0.000	0.277	0.784

3.7.2 G.2) Modelo Binomial Negativo corregido.

Dep. Variable:	Disease	**No. Observations:**	2351
Model:	GLM	**Df Residuals:**	2332
Model Family:	NegativeBinomial	**Df Model:**	18
Link Function:	Log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3690.3
Date:	Sun, 21 Apr 2024	**Deviance:**	1009.1
Time:	22:55:44	**Pearson chi2:**	673.
No. Iterations:	10	**Pseudo R-squ. (CS):**	0.2094
Covariance Type:	nonrobust		

	coef	std err	z	P> |z|	[0.025	0.975]
const	-1.1536	0.242	-4.776	0.000	-1.627	-0.680
Glu	-0.7014	0.132	-5.301	0.000	-0.961	-0.442
Cho	1.1080	0.137	8.098	0.000	0.840	1.376
Hemo	1.1769	0.124	9.509	0.000	0.934	1.419
Pla	-1.3523	0.100	-13.528	0.000	-1.548	-1.156
Hema	-0.3334	0.106	-3.152	0.002	-0.541	-0.126
MCV	1.2706	0.108	11.812	0.000	1.060	1.481
MCH	0.9293	0.097	9.599	0.000	0.740	1.119
Ins	0.2838	0.129	2.206	0.027	0.032	0.536
BMI	1.4345	0.161	8.897	0.000	1.118	1.750
SBP	-0.2677	0.123	-2.180	0.029	-0.508	-0.027
DBP	-0.7643	0.133	-5.729	0.000	-1.026	-0.503
Tri	-0.6218	0.120	-5.201	0.000	-0.856	-0.388
HbA1c	0.3830	0.107	3.588	0.000	0.174	0.592
LDL_Ch	-0.8789	0.123	-7.167	0.000	-1.119	-0.639
HDL_Ch	-0.3904	0.110	-3.550	0.000	-0.606	-0.175
HR	0.3105	0.115	2.699	0.007	0.085	0.536
Cre	0.5727	0.132	4.342	0.000	0.314	0.831
CRP	0.5354	0.128	4.187	0.000	0.285	0.786