Análisis de factores que influyen en el nivel del colesterol.

Aaron Mauricio Gómez Jiménez (UNAM) Francisca Vilca Sánchez (UC) Alejandro Correa Vázquez (UNAM)

1. Introducción

El colesterol es una sustancia cerosa y parecida a la grasa, en el cuerpo humano se mide a través del colesterol total y se debería medir al menos una vez cada cinco años. En el siguiente proyecto, el objetivo es identificar que componentes del cuerpo humano afectan en mayor medida al aumento del colesterol en la sangre.

Para estudiar esto, se pretende analizar cómo diferentes características como la edad, el peso y la altura,si pueden influir en la producción de colesterol. Se dispone de la base de datos *insulinoresistencia.csv* la cual cuenta con 64 muestras, cada una con 12 variables relacionadas con el colesterol y las hormonas. El objetivo principal es utilizar técnicas de regresión para identificar qué componentes tienen un impacto significativo en la producción de colesterol en el cuerpo humano.

Al encontrar un modelo de regresión que explique qué hormonas son las que afectan en mayor medida la producción de colesterol, se podrá obtener información valiosa sobre los factores que influyen en este proceso y potencialmente identificar biomarcadores útiles para prevenir o tratar desequilibrios en los niveles de colesterol. Esta investigación puede proporcionar una mejor comprensión de las interacciones hormonales y su relación con la salud cardiovascular, contribuyendo así a la prevención y manejo de enfermedades relacionadas con el colesterol.

2. Análisis exploratorio

Se hará un análisis exploratorio de los datos, ya que como se quiere hacer un modelo de regresión lineal se buscará las que variables tengan mayor correlación entre ellas, y así decidir que variables pueden aportar más al modelo de regresión. Observando la estructura de la base de datos es posible notar que hay 12 variables, con 64 observaciones, todas son variables numéricas, podemos ver las estadísticas más comunes como la media y la mediana, y algunos cuantiles.

Al analizar el resumen de estadísticas se nota que la edad promedio de la base es de 45,8 años, la persona mas joven tenía 20 años y la mas longeva 83 años, el peso promedio es de 68 kilogramos, altura promedio es de 1,56 metros, el IMC promedio es de 27,82, es decir, el promedio de las persona se encuentra en el rango de sobrepeso si tomamos en cuenta el IMC promedio, la persona con el mayor IMC se encuentra en la categoría de Obesidad Tipo II, la circunferencia abdominal promedio es 87 cm, la medición de glucemia basal promedio de la población es de 125,4, es decir, normal para una persona que esta en ayuna, como lo supone el muestreo, la variable Insulina se promedia en 8,125.

Para la variable HOMA IR el promedio es de 2,5 lo cual se considera normal, para el colesterol total que es la suma del colesterol bueno (HDL) más el colesterol malo (LDL) debe ser menor a 200 lm/dl en nuestro muestreo se encuentra en un promedio de 205, lo cual lo pone un poco arriba de los valores normales, pero en la categoría de fronterizo alto, debido a que el promedio es sensible a outliers tomamos en cuenta el tercer cuantil que se ubica en 230 lo cual se sigue manteniendo en fronterizo alto, es decir, en Hipercolesterolemia.

El colesterol bueno (HDL) tiene una medida de 44 en promedio, debido a que en promedio hombres y mujeres deben tener una medición menor a 37,5, se está arriba del estándar pero sin llegar a un nivel alto (> 60), para el colesterol malo (LDL) 118 es la media, lo cual es un poco elevado del óptimo (< 100).

Aquí se tiene una variable importante, dado que el promedio de Trigliceridos es de 214 mg/dl para la muestra, lo óptimo es que sea menor a 150 mg/dl, con 214 mg/dl se está con valores altos de los Trigliceridos.

Una vez analizado el comportamiento de los datos en términos de inferencia, se hará un gráfico para ver la correlación entre las variables, primero utilizando la función `pairs` para intuir gráficamente si algunas variable siguen alguna relación lineal y revisar el coeficiente de correlación, lo cual se observa en la Figura 1.

Los resultados muestran el coeficiente de correlación de Pearson entre las variables, esto nos sirve de guía para ver que variables están relacionadas, entre las mas significativas encontramos

IMC-Peso (.780) : Esto no es sorprendente ya que por definición el IMC involucra al Peso y la Altura, por lo tanto existe relación.

Insulina-IMC (.429): Esta relación es interesante, ya que es una relación débil moderada positiva, es decir, a mayor IMC mayor Insulina, esto tiene sentido, la Insulina es una hormona que ayuda a la glucosa que se encuentra en la sangre a entrar a las células de los músculos, grasa e hígado para obtener enegía, por lo tanto suena lógico que exista esta relación.

HOMAIR-IMC (.395): Recordando que el HOMAIR es una prueba de resistencia a la insulina, se podría argumentar que mientras mayor sea el IMC mayor es la resistencia a la Insulina.

HOMAIR-Insulina (.862): Esta relación es de las más obvias a mi parecer, ya que, a mayor índice HOMAIR mayor Insulina se requiere.

LDL-Colesterol Total (.909): Bajo el conociento de que por definición el Colesterol total es la suma de LDL más HDL, era de esperarse que existiera relación, sin embargo, podemos notar que la variable HDL no tiene una relación fuerte con Colesterol Total, esto es contra intuitivo.

Trigliceridos-Colesterol Total(.477): Se nota que una relación positiva media, lo cual podría implicar que entre mayor es el número de trigliceridos mayor es el colesterol total.

Gráficamente, solo se puede observar una relación lineal entre Colesterol Total y LDL, el peso y el IMC, peso y altura pero estos últimos tienen heterocedasticidad a simple vista con una varianza grande, esto se puede apreciar en el gráfico de dispersión de la Figura 1.

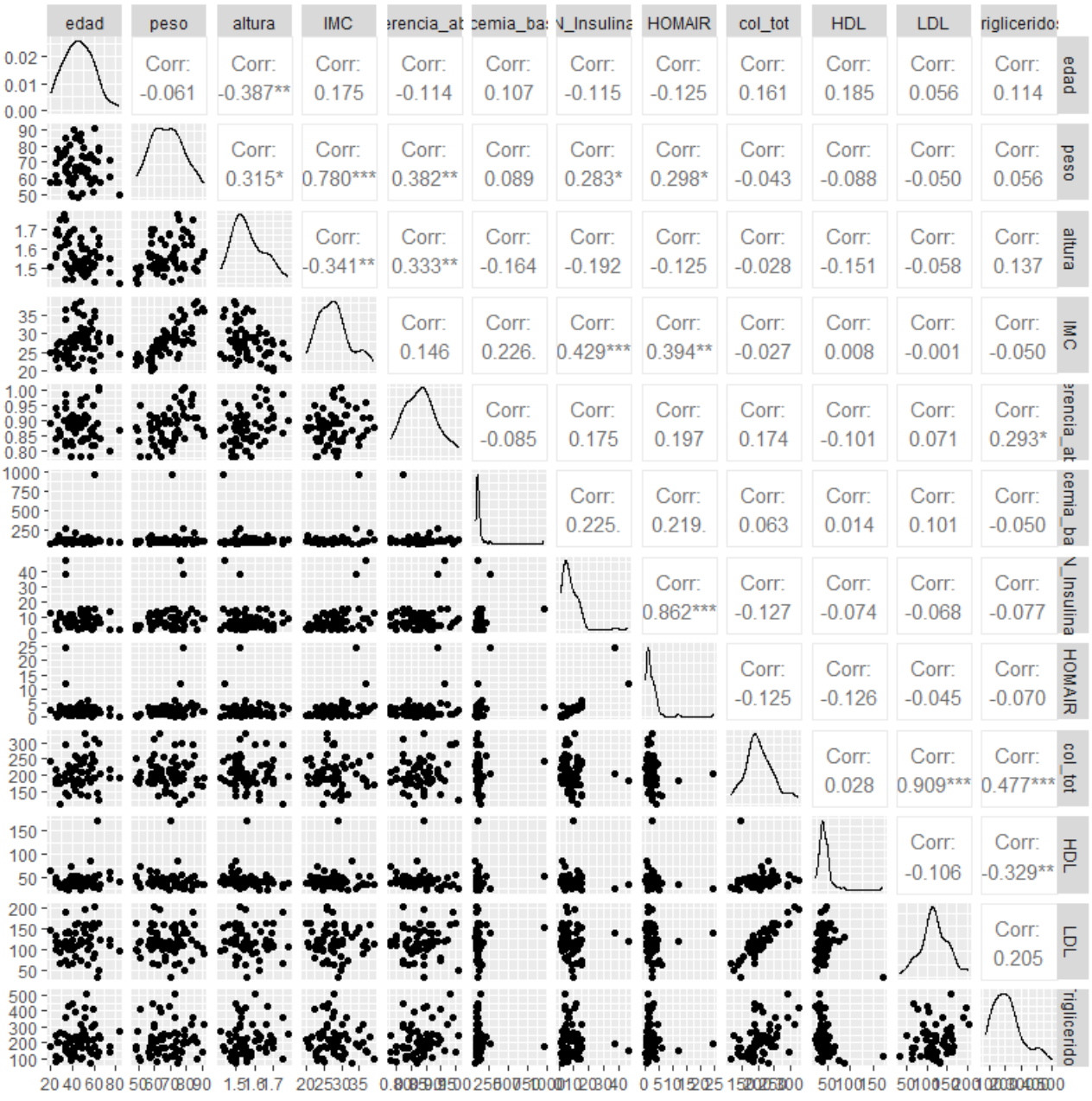

Figura 1: Gráfico de Dispersión

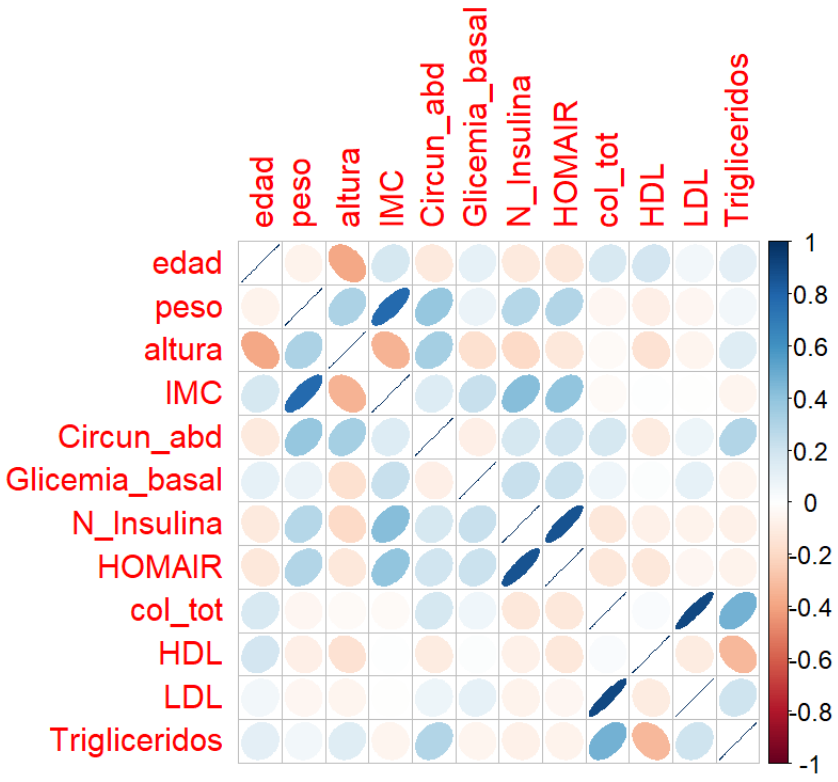

Figura 2: Correlograma

En el correlograma observamos las relaciones antes descritas pero de forma mas sencilla visualmente, ya que, mientras más grande sea el círculo, mayor es la varianza y su color dice el nivel de correlación que hay entre las variables.

3. Formulación del modelo

Ahora que ya se ha estudiado la relación entre las variables, es más fácil saber que variables incluir en el modelo de regresión, además al observar el correlograma, en especifico la variable de colesterol total; las variables LDL y Trigliceridos son las que más relación positiva tienen.

Para tener un análisis más completo haremos un modelo de regresión con todas las variables solo para confirmar que las variables que se cree que tienen relación en efecto la tienen.

Despúes se hará la regresión con las variables significativas para finalmente ajustar el modelo. Asumiendo los supuestos de Linealidad, Homocedasticidad, No autocorrelación, Distribución Normal de los errores, No Multicolinealidad y que no existan valores atípicos significativos, se hará el modelo de regresión lineal.

4. Ajuste o estimación del modelo

Regresión con todas las variables.

Empezando con una regresión lineal con todas las variables para ver si explican a la variable de colesterol total se obtiene.

```
Call:
lm(formula = col_tot ~ ., data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-17.956  -3.785  -0.244   2.761  41.522

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.740e+02  1.416e+02   1.229    0.225
edad                       8.940e-02  8.986e-02   0.995    0.324
peso                       1.477e+00  1.077e+00   1.371    0.176
altura                    -1.208e+02  8.901e+01  -1.357    0.181
IMC                       -3.633e+00  2.588e+00  -1.404    0.166
Circunsferencia_abdominal  3.052e+01  2.304e+01   1.324    0.191
Glicemia_basal             4.182e-03  1.042e-02   0.401    0.690
N_Insulina                 1.584e-01  2.907e-01   0.545    0.588
HOMAIR                    -6.805e-01  6.303e-01  -1.080    0.285
HDL                        5.264e-01  5.909e-02   8.909  4.8e-12 ***
LDL                        1.117e+00  3.176e-02  35.159  < 2e-16 ***
Trigliceridos              1.594e-01  1.222e-02  13.037  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.282 on 52 degrees of freedom
Multiple R-squared:  0.972,     Adjusted R-squared:  0.9661
F-statistic: 164.2 on 11 and 52 DF,  p-value: < 2.2e-16
```

Figura 3: Regresión con todas las variables

Solo tres varaibles son significativas HDL, LDL Y Trigliceridos, las otras variables no son significativas incluyendo al intercepto, tal como se ve en la Figura 3.

Regresión con variables significantes.

Utilizando el método forward para construir el modelo, se obtiene el modelo con las 3 variables significativas, ya que, agregar otra variable no mejoraba el modelo. Se comprueba que es un modelo bueno, ya que todas las variables son significantes al 5 % por lo tanto, con el modelo resultante, se comprobará en la siguiente sección que tan bueno es y si cumple con los supuestos que se hicieron al inicio.

```
Call:
lm(formula = col_tot ~ LDL + Trigliceridos + HDL, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-19.573  -3.278  -0.863   2.926  49.070

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.34118    5.25087   2.160   0.0348 *
LDL           1.11429    0.03072  36.275  < 2e-16 ***
Trigliceridos 0.17151    0.01117  15.353  < 2e-16 ***
HDL           0.56140    0.05617   9.995 2.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.245 on 60 degrees of freedom
Multiple R-squared:  0.968,     Adjusted R-squared:  0.9664
F-statistic:   605 on 3 and 60 DF,  p-value: < 2.2e-16
```

Figura 4: Regresión con las variables significantes

5. Verificación del modelo

Para verificar el modelo, se debe revisar que los estimadores encontrados tengan sentido, para ello usando el comando `summary()`, de *R*, posible comprobar que este modelo tiene sentido, si se observa en la Figura 5, es claro notar, los estimadores tienen sentido:

	Estimador	Error Estándar	Prueba T	Valor-p
Intercepto	11.341	5.251	2.160	0.035
LDL	1.114	0.031	36.275	0.000
Trigliceridos	0.172	0.011	15.353	0.000
HDL	0.561	0.056	9.995	0.000

Figura 5: Tabla resumen de los regresores estimados

De la Figura 2, lo que nos interesa notar es que el error estándar es pequeño y que para un nivel de signifancia del 5 % todos los regresores son significativos, vale decir, el modelo construido tiene sentido.

6. Evaluación del modelo

Para comprobar que el modelo creado tiene sentido, es decir, la selección de los parámetros es certera, hay que revisar la bondad o calidad del modelo, para ello hay diferentes medidas que ayudan en ello, las cuales se obtienen de manera sencilla en R usando el comando `summary()`:

1. **Error estándar de la regresión**
   Este valor se obtiene como la distancia entre las observaciones reales y las estimadas, para el modelo creado el valor de error estándar es de 8,25, lo que al ser un valor bastante pequeño nos confirma que nuestro modelo es bastante preciso, lo cual es lo que se busca en un modelo de regresión.

2. **Coeficiente de determinación**
   Este valor se puede entender como la proporción de variabilidad total de la variable respuesta explicada según los regresores, se como como R^2, el cual para este modelo es de: 96,8 %

3. **Coeficiente de determinación ajustado**
   Este valor, al igual que el anterior, tambien busca es la proporción de variabilidad total de la variable respuesta explicada según los regresores, sin embargo, este castiga por la cantidad de regresores ocupado, con el fin de no sobrestimar el modelo, para este caso el R^2 ajustado es de 96,6 %, el cual disminuye un poco en comparación al anterior, pero al ser tan poca su diferencia, aún se puede afirmar que es un buen modelo de regresión.

4. **Estadística F y su valor-p**
   Una sencilla prueba estadística para saber si los regresores tienen sentido es la estadística F, la cual cuya hipótesis nula se basa en comprobar si los regresores son significativos para el modelo, es decir,

   $$H_0 : \beta_i = 0 \qquad \forall i \in \{1, ..., k\}$$
   $$H_a : \exists j \in \{1, ..., k\} \qquad \beta_j \neq 0$$

   Para este caso, el valor de la estadística F es de 605 y su valor-p asociado es de: $< 2{,}2e-16$, lo que se interpreta que con un 5 % de significancia, todos los regresores son significativos.

7. Verificación de supuestos

Dado que al construir el modelo, se tomaron algunas decisiones arbitrarias de como son los datos una vez que se comprobó que es un buen modelo, solo basta revisar que estas decisiones son correctas, a continuación se pasarán a revisar cada una:

1. **Linealiadad de los parámetros**
   Se decidió usar un modelo de regresión lineal, lo cual se debe netamente a la forma lineal de las variables, si se observa la Figura al inicio, se puede comprobar este punto.

2. **Homocedasticidad**
   Para comprobar esta parte se usarán 2 metodologías las pruebas estadísticas y el análisis gráfico,

a) **Prueba Breusch-Pagan**
Esta prueba es útil para comprobar que hay homocedasticidad en los residuos, lo cual es lo que queremos, ya que si estos tienen alguna relación perdería el sentido, por lo que obteniendo su estadística con el comando `bptest()` en *R*, este es de 5,2649 y su valor-p asociado es de 0,1534 lo que con un nivel de significancia del 5 % no permite rechazar la hipótesis nula de que los datos son homocedasticos, por lo que se puede decir que tienen varianza constante

b) **Graficamente**
Usando los gráficos de Tukey-Anscombe es posible ver que los datos si se distribuyen como una nube de puntos luego de eliminar 2 valores un poco alejados

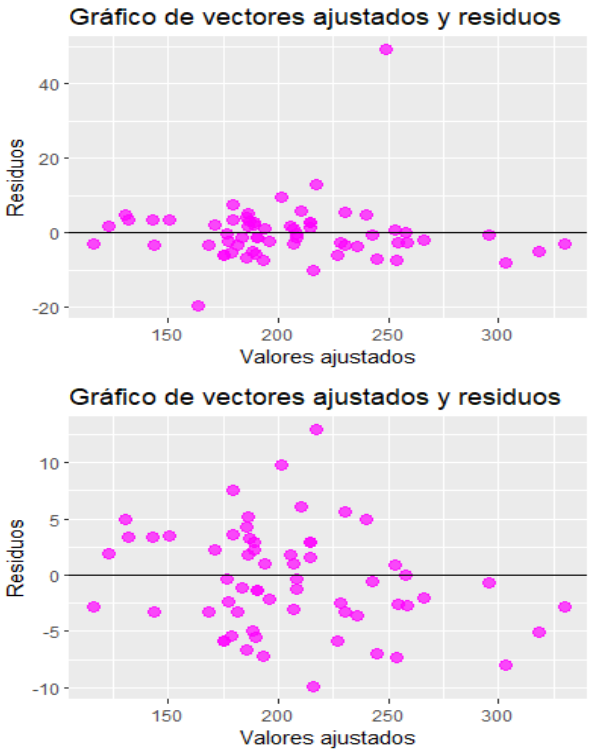

Figura 6: Comparación para ver la Homocedasticidad de los datos

En general la Figura 6, se puede apreciar la Homocedasticidad de los datos, que es lo que se quiere.

3. **No autocorrelación**
   Uno de los supuestos también es que no hay correlación entre los residuos, por lo que una forma de comprobar la autocorrelación de los datos es graficando usando el comando `acf()`, el cual se puede observar en la Figura 7:

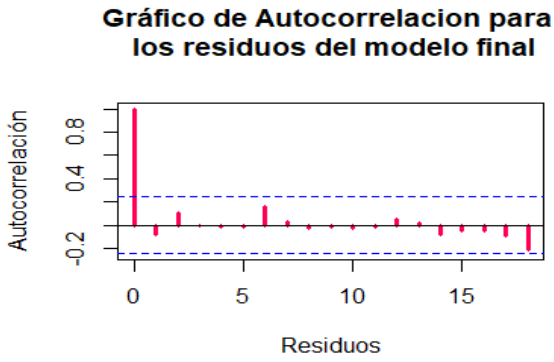

Figura 7: Autocorrelograma para los residuos del modelo

Gráficamente, vemos que los datos no poseen autocorrelación, sin embargo, al hacer la prueba de Durbin-Watson, esta no se rechaza, ya que su estadística obtenida con el comando `dwtest()` es de 2,1625 y su valor-p asociado es de 0,7512, por lo que para un nivel de significancia del 5 % no hay información estadística suficiente para rechazara la hipótesis nula, lo cual se puede deber a los valores atípicos que se tuvo que eliminar en el supuesto anterior.

4. **Errores normalmente distribuidos**
   Otro supuesto usado en el modelo es que los errores distribuyen normales, por lo que para comprobarlo se usará las pruebas de Kolmogorov-Smirnov, Shapiro-Wilks y el método gráfico:

   a) **Prueba Kolmogorov-Smirnov**
   Esta prueba es útil para comprobar que los residuos distribuyen normal, lo cual es lo que queremos que ocurra, obteniendo su estadística con el comando `kstest()` en *R*, este es de 0,41674 y su valor-p asociado es de $4{,}432e-10$ lo que con un nivel de significancia del 5 % si se permite rechazar la hipótesis nula de que los datos no distribuyen normal.

   b) **Prueba Shapiro-Wilks**
   Otra prueba de hipótesis para saber si existe una distribución normal con los residuales es la Shapiro-Wilks, al hacer la prueba de hipótesis obtenemos un p-value de $4{,}219e-10$, es decir se rechaza la hipotesis nula, por lo tanto no se distribuye Normal.

   c) **Gráficamente**
   Usando los gráficos para comprobar la normalidad el qqpllot en este es posible ver que los residuos si se distribuyen como una distribución Normal, pero las pruebas de hipótesis nos dicen que no es así, por eso es importante hacer varias pruebas para tener mayor seguridad al aceptar o rechazar los supuestos, a priori se podría decir que gráficamente se distribuye normal,tal como se muestra en la Figura 8, pero ya vimos que no es así:

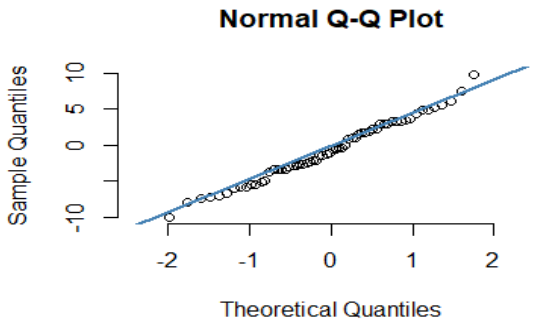

Figura 8: QQplot de los residuos

5. **No hay multicolinealidad perfecta**
   Una forma de estudiar la multicolinealidad y evitarla es revisando el estadístico VIF, el cual se construye en base del valor de R^2, por lo que valores de VIF muy grandes significa que se esta siendo redundante con el modelo, para este caso se tiene:

	VIF
LDL	1.046
Trigliceridos	1.160
HDL	1.124

Figura 9: Valores del VIF para cada variable

   Como para todas las variables el valor es muy pequeño, se ve que no hay multicolinealidad en los datos.

6. **Valores Atípicos**
   Una forma sencilla de detectar outliers, es con una gráfico de caja o boxplot, el cual gráficamente muestra cuales son los outliers, en la Figura 10, se ve claramente el caso:

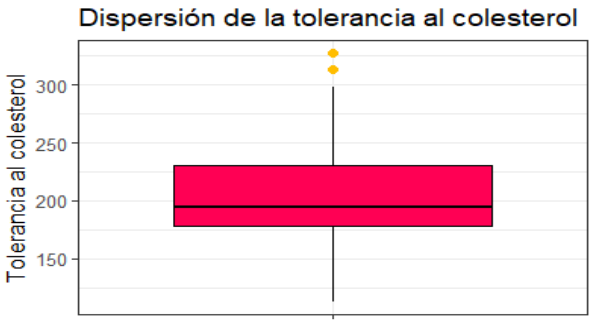

Figura 10: Boxplot de la variable Tolerancia al colesterol

Aquí nuevamente se observa los 2 valores atipicos que aparecieron en este análisis de supuestos.

8. Conclusiones

En conclusión al realizar un análisis de regresión lineal con todas las variables de la base, notamos que solo eran significativas las que en el análisis exploratorio tenían mayor relación con la variable que queríamos explicar; al efectuar una nueva regresión con las variables significativas obtuvimos una regresión con una R-cuadrada de 96 por ciento, la cual es muy buena y por lo tanto decidimos quedarnos con ese modelo.

El colesterol total se puede explicar de muy buena forma con las variables HDL(Colesterol bueno) LDL(Colesterol malo) y los trigliceridos, es decir existe una relación lineal múltiple,es decir entre mayor sea el colesterol total en teoría las variables deberían crecer también en la misma dirección, sería interesante hacer un análisis multivariado para ver en que proporción afecta cada variable al colesterol total.

Además, el estudio también investigó la relación de estos componentes del perfil lipídico están estrechamente relacionados y juegan un papel crucial en la salud cardiovascular. El análisis de su interacción puede proporcionar información valiosa sobre el equilibrio lipídico y su impacto en la salud.