Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

Actuaría

Analisis De Regresion

Profesor: Arturo Vera Moreno

Proyecto

Semestre 2024-II

Elaborado por:

Torres Ferrer Luis Noe

Soto Luna Denilson

Hernandez Estrada Yahir Emiliano

Galan Olivares Juan Miguel

Índice

1 Introducción

El peso corporal es una variable fundamental en el análisis de la salud, ya que está estrechamente relacionado con diversas condiciones y factores de riesgo. Nos llamó la atención un conjunto de datos disponible en Kaggle, el cual fue creado utilizando información de tres bases de datos provenientes de la Encuesta Nacional de Salud y Nutrición (ENSANUT), publicada por el gobierno de México.

Este conjunto de datos nos brinda la oportunidad de explorar cómo diferentes variables, como el sexo, la edad, la concentración de hemoglobina, los niveles de colesterol (HDL, LDL, total), creatinina, ácido úrico, albúmina, el índice de masa corporal (IMC) y la actividad física, entre otras, pueden influir en el peso corporal. Nuestro objetivo es construir un modelo de regresión que permita predecir el peso de una persona en función de estas características, identificando los factores más relevantes y entendiendo cómo interactúan para influir en este resultado.

Con este análisis, buscamos no solo desarrollar un modelo predictivo, sino también proporcionar información que podría ser útil para diseñar estrategias personalizadas de control de peso y promoción de un estilo de vida saludable.

2 Hipótesis

Es posible predecir el peso de una persona en función de variables relacionadas con el estado de salud de un individuo y el estilo de vida; es decir, el peso está relacionado con la edad, sexo, altura, tensión arterial, actividad física, presencia o ausencia de diabetes, colesterol y hemoglobina.

3 Objetivo

Nuestro objetivo a partir de esta investigación es poder crear un modelo de regresión lineal múltiple para el cual necesitaremos establecer variables que resulten importantes para modelar las relaciones entre el peso de una persona y sus enfermedades o padecimientos a través de técnicas estadísticas.

Las variables propuestas en la hipótesis no solo permiten estimar el peso con mayor precisión, sino que también podrían aportar información clave para comprender la relación entre el peso corporal y condiciones de salud como la hipertensión.

Al construir y analizar este modelo de regresión, esperamos identificar patrones significativos que podrían facilitar la intervención temprana y la personalización de estrategias de control de peso y prevención de enfermedades crónicas.

4 Marco Teórico

4.1 Definiciones médicas

- **Hipertensión** La hipertensión arterial es una enfermedad crónica en la que aumenta la presión con la que el corazón bombea sangre a las arterias, para que circule por todo el cuerpo. IMSS
- **Hemoglobina** Metaloproteína intraeritrocitaria formada por un grupo proteínico, la globina, unido al grupo hemo. Su función es el transporte de oxígeno a los tejidos y de dióxido de carbono desde los tejidos a los pulmones. MedlinePlus
- **Ácido úrico** Producto final del catabolismo de las purinas en los seres humanos y los primates, excretado fundamentalmente por el riñón y, en menor medida, por vía intestinal. MedlinePlus2
- **Albúmina** Proteína simple, soluble en agua y coagulable por calor, ampliamente distribuida en los tejidos de animales y plantas. Es también responsable de la presión oncótica del plasma sanguíneo. Los niveles de albúmina en plasma disminuyen en la desnutrición extrema o en enfermedades renales o hepáticas y causan edemas generalizados, presentes incluso en la cavidad abdominal (ascitis). MedlinePlus5
- **Insulina** Hormona polipeptídica segregada por las células b de los islotes pancreáticos de Langerhans. La insulina se libera en respuesta a la elevación de la glucosa sanguínea, aminoácidos y hormonas entre otros agentes secretagogos, y fomenta la conservación y el uso eficientes de los sustratos energéticos mediante el control del transporte de metabolitos y de iones a través de la membrana celular y la regulación de las vías intracelulares de biosíntesis. MedlinePlus3
- **Glucosa** Es la principal fuente energética de los tejidos, especialmente del cerebro; sus concentraciones sanguíneas están reguladas principalmente por las hormonas insulina (hipoglucemiante) y glucagón, cortisol, adrenalina y hormona del crecimiento (hiperglucemiantes). MedlinePlus4
- **transferrina** Globulina del plasma con un peso molecular 80 kDa que se sintetiza en el hígado y es capaz de ligar hierro reversiblemente y transportarlo. Cada molécula de proteína puede ligar hasta dos átomos de hierro (Fe3+) junto con bicarbonato. [Boccio (2003)]

4.2 Investigación

Determinación de Diabetes

La Organización Mundial de la Salud menciona:
La diabetes es una enfermedad crónica que se presenta cuando el páncreas no secreta suficiente insulina o cuando el organismo no logra utilizar eficazmente la insulina que produce. La insulina es una hormona que regula la concentración de glucosa en la sangre. Un efecto común de la diabetes no controlada es la hiperglucemia (es decir, la glucemia elevada) que, con el tiempo, daña gravemente muchos órganos y sistemas del organismo, sobre todo los nervios y los vasos sanguíneos. [OMS]

Mediante estudios se ha cambiado el valor mediante el cual se determina si la persona tiene o no diabetes, este valor se conoce como glucosa, y una forma de determinar es mediante la prueba PGA, que es la que tenemos en nuestros datos en la variable Resultado de Glucosa Promedio".

"La prueba de glucosa en plasma en ayunas (GPA), mide la concentración de glucosa en un determinado momento."National Institute of Diabetes and Digestive and Kidney Diseases [NIDDK].

American Diabetes Association [ADA] e [IBC] mencionan los siguientes criterios para el diagnóstico de la diabetes:

1. $A1C \geq 6,5\,\%$. La prueba debe realizarse en un laboratorio utilizando un método certificado por $NGSP$ y estandarizado según el ensayo $DCCT$.

2. Si los niveles de glucosa plasmática en ayunas $GPA \geq 126mg/dl(7,0mmol/l)$. El ayuno se define como la ausencia de ingesta calórica durante al menos 8 horas.

3. *Glucemia plasmática en 2 horas*$\geq 200mg/dl(11,1mmol/l)$ durante una prueba de tolerancia oral a la glucosa $(PTGO)$. La prueba debe realizarse según lo descrito por la Organización Mundial de la Salud, utilizando una carga de glucosa que contenga el equivalente a $75g$ de glucosa anhidra disuelta en agua.

4. En un paciente con síntomas clásicos de hiperglucemia o crisis hiperglucémica, una *glucosa plasmática aleatoria*$\geq 200mg/dl(11,1mmol/l)$

NOTA: En ausencia de hiperglucemia inequívoca, los criterios 1 a 3 deben confirmarse mediante pruebas repetidas.

Para identificar diabetes y diabetes no controlada tenemos la siguiente información de [IBC]:

De la prueba de $A1C$ se presentan como porcentajes. Mientras más alto se encuentre el nivel de $HbA1c$, mayor será el riesgo de las graves complicaciones relacionadas con la diabetes.

- El nivel normal de $HbA1c$ se encuentra por debajo del $5{,}7\,\%$, para alguien que no tiene diabetes.

- $HbA1c$ entre $5{,}7\,\%$ y $6{,}4\,\%$, se tiene pre-diabetes, lo que significa que tienes un alto riesgo de contraer diabetes en el futuro.

- $A1C$ del $6{,}5\,\%$ o más, indica que tienes diabetes.

- $HbA1c$ por encima del $8\,\%$ significa que la diabetes no está bien controlada y tienes un mayor riesgo de desarrollar complicaciones relacionadas con esta.

Síndrome metabólico por [NLM]

Es el nombre de factores de riesgo de enfermedades cardiacas, incluida la diabetes. Si se tiene al menos tres de ellos, se llama síndrome metabólico. Estos factores de riesgo incluyen:

- Obesidad abdominal, es decir, tener mucha grasa alrededor de la cintura, es una de las mayores amenazas para enfermedades del corazón.

- Un nivel alto de triglicéridos, un tipo de grasa que se encuentra en la sangre.
- Un nivel bajo de colesterol HDL: Él HDL ayuda a eliminar el colesterol de las arterias.
- Presión arterial alta: Si la presión arterial se mantiene alta en el tiempo, puede dañar sel corazón y provocar otros problemas de salud.
- Glucosa alta: El nivel de azúcar en la sangre levemente alto puede ser un signo temprano de diabetes, es decir, (GPA) de $100mg/dl(5,6mmol/l)$ a $125mg/dl(6,9mmol/l)$. [ADA]

¿Qué causa el síndrome metabólico?

- Sobrepeso y obesidad
- Un estilo de vida inactivo
- Resistencia a la insulina
- Edad: Su riesgo aumenta a medida que envejece
- Genética

Las personas que tienen síndrome metabólico también presentan un aumento en su coagulación de la sangre e inflamación en todo el cuerpo. Los investigadores no saben si estas afecciones causan el síndrome metabólico o si lo empeoran.

4.2 Estudios e Investigaciones Anteriores Similares

Se tiene un estudio realizado acerca de "Factores que afectan el estado nutricional en personas mayores mexicanas"[SPM].

Los resultados obtenidos son que de 4587 participantes. La prevalencia de desnutrición fue $16{,}1\,\%$ y está relacionada con $edad \geq 80$ años, sin pareja, sin escolaridad, sobreestimación de índice de masa corporal (IMC), dificultad motriz, dependencia funcional instrumental, hospitalización en año previo y caídas en los últimos dos años, autorreporte de fuerza prensil débil, reporte de desastre que afectó vivienda o accidente que afectó la salud. La prevalencia de exceso de peso fue $43{,}6\,\%$, relacionada con ser mujer, tener 60 a 79 años, percibirse sin sobrepeso u obesidad y subestimarlo contra IMC, tener ≥ 3 enfermedades, síntomas somáticos e inactividad física.

Se tiene una tesis [Marcuño (1987)] donde realizo regresión lineal múltiple con algunas variables que ocupamos, mencionando lo siguiente en el resumen:

Se estudian los niveles sanguíneos en 193 ancianos (113 mujeres y 79 hombres) de glucosa, urea, creatinina, a. Úrico, colesterol, triglicéridos, proteínas totales, albúmina, osmolalidad, calcio, fósforo, bilirrubinas total y directa, aspartato y alanino aminotranferasas, fosfatasa alcalina, gammaglutamil transferasa, lipidograma, hdl y ldl colesterol apolipoproteínas a y b, sodio, potasio, cloro y hierro, considerándose en cada individuo las siguientes variables biológicas: sexo, edad, peso talla y tensión arterial. Se estudia la posible variación de las variables citadas en función de la edad y del sexo. **Se constata el aumento en función de la**

edad de la tensión sistólica, glucosa y urea con nivel de significación p(0,05) y la existencia de **diferencias estadísticamente significativas entre los sexos en la talla, sobrepeso, a. Úrico, creatinina, colesterol, fósforo, hdl-colesterol, apolipoproteínas** a p(0,001) peso, albúmina, bilirrubinas totales y direp(0,01) aspartato aminotransferasa y fosfatasa alcalina ((0,05). **Los varones muestran una disminución de la albúmina al avanzar la edad** p(0,01) y **las mujeres un aumento de la osmolalidad** p(0,05). Se calculan los límites del 95 % central de los valores para cada uno de los parámetros bioquímicos, junto con sus límites de confianza, en el conjunto de individuos y en hombres y mujeres separadamente. Se estudia la correlación existente entre las distintas variables bioquímicas y entre estas y las biológicas. Finalmente, se cuantifica la dependencia de cada variable respecto a las demás por regresión linear múltiple paso a paso.

Se tiene un modelo polinómico, donde se calcula el peso, pero no de humanos, sino de bovinos, con esto obtenemos, si bien que los resultados no van a ser similares obtenemos una idea de las variables que utilizaron, además de una comparación, ya que se utilizó otro modelo como se menciona en [Predição do peso... (2008)]:

El objetivo de este estudio fue investigar las **relaciones entre el peso corporal y las medidas corporales de altura de grupa** (ag)**, longitud de grupa** (gg)**, longitud corporal** (cc) **y circunferencia torácica** (pt), en bovinos derivados principalmente del cruce de razas holandesa y Chica. Se utilizaron datos de 483 vacas, 469 novillas y 62 machos, de tres rebaños diferentes, analizados por separado para cada categoría, con el fin de establecer ecuaciones polinómicas de pesos en relación con las medidas corporales. Las correlaciones simples del peso corporal con pt, cc, cg y ag fueron $0{,}807$, respectivamente; $0{,}440; 0{,}187 y 0{,}504$ para vacas; $0{,}928; 0{,}735; 0{,}819$ y $0{,}880$ varones, y $0, 942; 0{,}748; 0{,}902$ y $0{,}573$ para novillas. Si bien las regresiones del peso corporal con relación a cc y cg fueron significativas $(P < 0{,}05)$, el aumento en el porcentaje de explicaciones de las variaciones del peso corporal obtenido con la inclusión de estas mediciones, además de pt, no parece justificar el costo de las mediciones. Las ecuaciones de predicción del peso corporal en función de los pt fueron las siguientes: para vacas, $peso = 12{,}174 - 187{,}410pt + 0,97196960pt^2 - 0,00162382pt^3$, para novillas, $peso = 1{,}717 - 35,167pt + 0,238978pt^2 - 0{,}00046260pt^3$ y, para los hombres, $peso = -3{,}862 + 76{,}014pt - 0{,}488837pt^2 + 0{,}00109755pt^3$.

Como conclusión se tiene que el peso se puede estimar mediante un modelo que incluye medidas corporales, siendo pt la variable explicativa que más contribuye a la adherencia del modelo. Los modelos polinomiales para predecir el peso con base en la circunferencia torácica, cuando se incluyeron hasta el término cúbico, mostraron alta adherencia y fueron diferentes para las categorías de vacas, novillas y machos.

Se realizó un estudio de regresión lineal múltiple se menciona que el peso es una variable para explicar enfermedades, se puede asumir que el peso tiene una relación con estas enfermedades crónico-degenerativas, pero como se relacionan estas enfermedades para determinar el peso es lo que va a tratar de realizar nuestro modelo de regresión.

Se menciona que .ºbtuvieron promedios de desviaciones absolutas menores a las obtenidas con las ecuaciones de Harris-Benedict y Mifflin-St. Jeor". Notando que en comparación con ecuaciones, **un modelo de regresión múltiple es mejor.**

Tomaron datos de México [Moreno y Sánchez (2020)]:

La Encuesta Nacional de Salud y Nutrición de Medio Camino 2016, señala que **el sobrepeso y la obesidad** son las alteraciones del estado nutricio con mayor prevalencia en la

población mexicana, esto se traduce en un **alto riesgo de desarrollar patologías crónico-degenerativas a futuro**. Una oportuna estimación del gasto energético total (GET) deriva en intervenciones preventivas en la población de jóvenes. Utilizando la metodología de minería de datos con **dos máquinas de aprendizaje, redes neuronales artificiales (RNA) y el modelo de regresión lineal múltiple (MRLM)**, donde **a partir de ocho variables explicatorias (sexo, edad, peso, talla, masa grasa, masa libre de grasa, agua corporal total y actividad física) se estimó el GET**. Los resultados obtenidos mediante ambos métodos presentan un valor promedio absoluto de desviación (e_a) de 121 y 116 respectivamente, menor al calculado por fórmulas de estimación $(e_a = 558)$ tomando como referencia el cálculo de este a partir de bioimpedancia eléctrica. Se concluye que ambas máquinas de aprendizaje, el RNA y el MRLM, obtuvieron promedios de desviaciones absolutas menores a las obtenidas con las ecuaciones de Harris-Benedict y Mifflin-St. Jeor. Hoy en día el uso de la ciencia de datos representa un área de oportunidad para lograr impactos positivos en el sistema de salud en México.

En un estudio se tomó el ejemplo de [**?**] con datos de [**?**] donde se obtuvo el siguiente resultado: Siendo el más elevado numéricamente $(-0,16)$ el del número de cigarrillos fumados, que es de signo negativo, como esperaríamos.

Fue el resultado más significativo para determinar el peso de un recién nacido, factores externos que afectan la salud del individuo, en este caso, la madre en gestación determinan el peso que podría tener al nacer.

Hasta este punto podemos notar que el peso se puede determinar mediante otras variables, incluso variables de salud. Pero para qué nos sirve determinar el peso si depende principalmente de cantidad de comida que una persona pueda comer. Si bien el principal factor que hace que una persona aumente o disminuya de peso es la cantidad de calorías que consuma, el objetivo de este modelo de regresión es determinar el peso, siendo que el sujeto consuma la misma cantidad de calorías.

Por [NLM Trastornos metabólicos]: Se puede desarrollar un trastorno metabólico si algunos órganos, como el hígado o el páncreas, se enferman o no funcionan normalmente. La diabetes es un ejemplo.

[Chitarroni (2002)]. El análisis de correlación y regresión lineal entre variables cuantitativas. El autor de la investigación aborda la correlación entre variables cuantitativas, lo que resulta crucial para nuestro análisis, ya que incluso hacen uso de las variables como el peso y la altura, las cuales se correlacionan con otras variables bastante parecidas a las nuestras.

[Cortés (2015)]: El autor realiza un modelo estadístico en la cual expresa una variable respuesta en función de otras variables predictoras. En la correlación las 2 variables en estudio tienen un papel simétrico, pero en el modelo de regresión es asimétrico: la respuesta representa la futura incógnita, y la predictora, la información que estará disponible.

Los 2 objetivos del modelo de regresión son: (1) anticipar el valor que tomará la respuesta; y (2) cuantificar la precisión de esta predicción. Y como primer ejemplo presentan que: El peso de una persona se puede adivinar, en parte, por su altura.

[Sabogal (2021)] Sistema E-Health de adquisición y almacenamiento de variables fisiológicas, obtenidas de dispositivos comerciales, necesarias para predecir el nivel de insulina. El autor realiza un trabajo muy extenso, ya que Este trabajo de grado plantea la implementación de sistema que permite la adquisición y almacenamiento de un conjunto de variables (Frecuencia cardiaca, nivel de glucosa, cantidad de paso, cantidad de calorías y el peso) de dispositivos

comerciales para determinar el nivel de insulina que el paciente diabético requiere en una sola base de datos.

5 Fundamento Matemático

5.1 Medidas

5.1 Medidas de Tendencia Central

Son medidas estadísticas que pretenden resumir en un solo valor a un conjunto de valores. Representan un centro en torno al cual se encuentra ubicado el conjunto de los datos. Las medidas de tendencia central más utilizadas son: media, mediana y moda.

5.1 Medidas de dispersión multivariada

La matriz **S** es una posible generalización multivariada de la noción univarida de varianza. La medición es más escalable en torno a la media. Sin embargo, a veces es conveniente tener un único número para medir la escala multivariada. Dos medidas comunes son:

1. La varianza generalizada, $|\mathbf{S}|$
2. La variación total tr **S**

Los valores grandes indican un alto grado de dispersión en torno a x y los valores bajos representan una concentración en torno a x.
La varianza generalizada juega un papel importante en la estimación de máxima verosimilitud y la variación total es un concepto útil en el análisis de componentes principales.

5.1 Medidas de Dispersión

- **Rango**

$$Rango = X_{\text{máx}} - X_{\text{mín}}$$

- **Desviación Estándar**

$$D_{\bar{X}} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

- **Varianza**

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

- **Coeficiente de Variación**

$$CV = \frac{\sigma}{|\bar{x}|}$$

5.1 Medidas de Correlacion Bivariadas

Covarianza

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})$$

- Positivo indica que las variables son difectamente proporcionales
- Negativo indica que las variables son inversamente proporcionales
- 0 No existe relacion entre las variables

Coeficiente de Correlación de Pearson

$$r = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$$

5.2 Modelo de Regresión Simple

5.2 Definición

El modelo de regresión lineal estudia la relación lineal que existe entre una variable dependiente y una variable independiente a través de la siguiente forma funcional:

$$y = \beta_0 + \beta_1 x + u$$

Donde:

- y es la variable dependiente o endógena.
- x es la variable independiente o exógena.
- β_0 y β_1 son los parámetros a estimar o regresores
- u es el término de error o perturbación

5.2 Estimación por Mínimos Cuadrados Ordinarios (MCO)

Sea $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ una muestra de v.a.'s en una población con la siguiente dinámica:

$$y = \beta_0 + \beta_1 x + u$$

Donde β_0 y β_1 son valores desconocidos y constantes en la población, por lo que se busca estimarlos. A partir de los valores estimados de $\hat{\beta}_0$ y $\hat{\beta}_1$ es posible estimar el valor de la perturbación como:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Así, el método de MCO consiste en encontrar los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimicen la suma de los cuadrados de las perturbaciones estimadas.

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \hat{u_i}^2 = \min \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Condiciones de Primer Orden:

$$C.P.O = \begin{cases} \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} \hat{u_i}^2 = \sum_{i=1}^{n} (-2)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} \hat{u_i}^2 = \sum_{i=1}^{n} (-2x_i)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{cases}$$

Resolviendo el sistema de ecuaciones:

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{cases}$$

5.2 Teorema de Gauss-Markov

Bajo los supuestos de Gauss-Markov, el estimador de Mínimos Cuadrados Ordinarios (MCO) es el mejor estimador lineal insesgado (BLUE) para el Modelo de Regresión Lineal.

Supuestos de G-M:

1. Linealidad
2. Muestra Aleatoria
3. Exógeneidad de la v.i.'s
4. Perturbaciones con esperanza condicional igual a 0
5. Homoscedasticidad condicional de las perturbaciones
6. No correlación condicional de las perturbaciones

5.2 Propiedades

Para una colección de v.a.i.i.d.'s $u_1, u_2, ..., u_n$ que cumplen con el teorema de G-M, los estimadores de MCO del MRLS cumplen que:

1. $\bar{\hat{u}} = 0$
2. $S_{x\hat{u}} = 0$
3. $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$
4. $S_{\hat{y}\hat{u}} = 0$

5.2 Normalidad y estimación por máxima verosimilitud

Teorema 1. *Para una colección de v.a.i.i.d.'s $u_1, u_2, ..., u_n$ que cumplen con los supuestos de G-M y $u_i \sim N(0, \sigma^2)$:*

1. $\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$
2. $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$
3. $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
4. $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$
5. $(\hat{\beta}_0, \hat{\beta}_1)$ *son independientes de* σ^2

5.3 Modelo de Regresión Lineal Generalizado (Múltiple)

5.3 Construcción

El modelo de regresión lineal múltiple es una generalización del modelo simple con $k-1$ variables independientes:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_{k-1} x_{1(k-1)} + u_1$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_{k-1} x_{2(k-1)} + u_2$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_{k-1} x_{n(k-1)} + u_n$$

El sistema de n ecuaciones con k parámetros puede construirse usando álgebra lineal:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{13} \\ 1 & x_{21} & x_{22} & \cdots & x_{23} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

Es decir $Y_n = X_{n\times k} b_k + u_n$

Al igual que en el modelo simple, para el modelo múltiple se busca minimizar la suma de los cuadrados de las perturbaciones estimadas:

$$\min_b \sum_{i=1}^{n} \hat{u_i}^2 = \min_b \hat{u}'\hat{u} = \min_b (Y - X\hat{b})'(Y - X\hat{b})$$

Utilizando el criterio de la primera derivada:

$$\frac{\partial}{\partial \hat{b}}(Y - X\hat{b})'(Y - X\hat{b}) = -2Y'X + 2\hat{b}X'X = 0$$

Resolviendo la ecuación matricial obtenemos:

$$\hat{b}_{MCO} = (X'X)^{-1}(X'X) = b + (X'X)^{-1}X'u$$

Podemos observar también:

$$\hat{\sigma}_{u,MCO} = \frac{1}{n-k}\sum_{j=1}^{n}\hat{u_j}^2 = \frac{1}{n-k}(Y - X\hat{b}_{MCO})'(Y - X\hat{b}_{MCO})$$

Teorema 2 (Teorema de Gauss-Markov). *Para* $\left\{\begin{pmatrix} y_i \\ x_i \end{pmatrix}\right\}_{j=1}^{n}$ *una sucesión de vectores aleatorios i.i.d.'s tales que* $Y = Xb+u$ *con* $E(u) = 0$ *y* $Var(u) = \sigma_u^2 I_n$ *se cumple que el estimador de MCO de* b $(\hat{b}_{MCO})$ *es BLUE (mejor estimador lineal insesgado)*

5.3 Propiedades (asumiendo normalidad)

Para $\left\{\begin{pmatrix} y_i \\ x_i \end{pmatrix}\right\}_{j=1}^{n}$ una sucesión de vectores aleatorios i.i.d.'s tales que $Y = Xb + u$ con X y u independientes y $u_i \sim N_n(0, \sigma_u^2 I_n)$ se cumple que:

- $\hat{b}_{MCO} \sim N_k(b, \sigma_u^2(X'X)^{-1})$
- $\frac{\hat{u}'\hat{u}}{\sigma_u^2} = \frac{1}{\sigma_u^2}u'[I_n - X'(X'X)^{-1}X]u = \left(\frac{n-k}{\sigma_u^2}\hat{\sigma}_{u,MCO}^2\right) \sim \chi_{(n-k)}^2$
- $E(\hat{\sigma}_{u,MCO}^2) = \sigma_u^2$
- $\hat{b}_{MCO}$ y $\hat{\sigma}_{u,MCO}^2$ son independientes.

5.3 Supuestos del Modelos de regresión lineal múltiple

Recordemos que partimos de los siguientes supuestos de modelo de regresión múltiple

- $\epsilon \sim N(0, \sigma^2)$
- σ^2 es constante
- Rango de X es completos
- Colinealidad
- Autocorrelación

5.3 Aplicaciones del Modelo de Regresión Lineal

- **Predicción** La regresión lineal múltiple se puede utilizar para predecir el valor de una variable en el futuro.
- **Explicación** La regresión lineal múltiple se puede utilizar para explicar la relación entre dos o más variables.
- **Control** La regresión lineal múltiple se puede utilizar para controlar el efecto de una variable sobre otra

5.4 Análisis de componentes principales

Sea $a'x$ la combinación lineal estandarizada (SLC) donde $a'a = 1$ y $a = c_1\gamma_{(1)} + ... + c_p\gamma_{(p)}$ considerando a $\gamma_{(1)}, ..., \gamma_{(p)}$ los vectores propios de $\sum$

Teorema 3. *No existe una combinación lineal estandarizada (SLC) del vector aleatorio x que tenga una varianza "más grande"que la varianza del primer componente principal λ_1. De forma análoga, el último componente principal de x tiene la varianza "más pequeña"que cualquier combinación lineal estandarizada (SLC) del vector aleatorio.*

Teorema 4. *Si $a'x$ es una combinación lineal estandarizada (SLC) de x la cual no está correlacionada con las primeras k componentes principales de x, entonces la varianza de $a'x$ es maximizada cuando $a'x$ es la $(k+1)$-ésima componente principal de x.*

5.5 Transformaciones

5.5 La transformación de escala

Sea $y_r = D^{-1}(x_r - x), r = 1, ..., n$, donde $D = diag(s_i)$. Esta transformación escala cada variable para que tenga varianza unitaria y, por lo tanto, elimina la arbitrariedad en la elección de la escala (No importaria la escala). Note que: $S_y = R$

5.5 transformación de Mahalanobis

Si $\mathbf{S} > 0$ entonces S^{-1} tiene una raíz cuadrada definida positiva simétrica única $S^{-1/2}$. La transformación de Mahalanobis se define por:

$$z_r = S^{1/2}(x_r - x), r = 1, ..., n.$$

Entonces $S_z = 1$ de modo que esta transformación elimina la correlación entre las variables y estandariza la varianza de cada variable.

5.5 Transformación de componente principal

Por el teorema de la descomposición espectral podemos escribir a la matriz de covarianza S como

$$S = GLG'$$

donde

(1) G es una matriz ortogonal.

(2) L es una matriz diagonal de los valores propios de S, con $l_1 \geq l_2 \geq \cdots \geq l_p \geq 0$.

El principal componente de la transformación se define por la rotación $w_r = G'(x_r - \bar{x})$ con $r = 1, ..., n$.

Dado que $L = G'SG = S_w$ es una matriz diagonal, las columnas de W (llamadas componentes principales), representan combinaciones lineales no correlacionadas de las variables. La idea es resumir la variabilidad en los datos usando únicamente los componentes principales con las varianzas más altas, por lo que se definen las funciones simétricas y monótonas crecientes:

- $|S| = |L| = \prod l_i$
- $tr(S) = tr(L) = \sum l_i$

Por lo que la rotación a componentes principales provee más medidas de la dispersión multivariada.

5.6 Valores Atipicos (Outliers)

Formas de detectarlos:

- Histograma
- Diagrama de caja y bigote
- z-score
- z-modificada
- RIC (Rango Inter Cuantil)
- PCA

5.6 Histograma

Este método se basa en revisar de manera grafica si la distribución cuenta con valores extremos en baja proporción, por ejemplo:

5.6 Diagrama de Caja

Los diagramas de caja a su vez podemos identificar la media y de manera visual podemos ver si hay .ºutliers"que son los puntos que caen a más de 1.5 veces el IQR, a partir de la caja.

6 Análisis descriptivo y explorativo

6.1 Descripción del set de datos

Es una union de tres conjuntos de datos provenientes de la página de gobierno llamada Encuesta Nacional de Salud y Nutrición (ENSANUT) https://ensanut.insp.mx/encuestas/ensanutcontinua2022/descargas.php. Los conjuntos de datos utilizados son los siguientes:

- ensaantro2022_entrega_w.csv (Cuestionario de antropometría y tensión arterial)
- Determinaciones_bioquímicas_cronicas_deficiencias_9feb23.csv (Determinaciones para enfermedades crónicas y deficiencias)
- ensafisica2022_adultos_entrega_w.csv (Actividad física - Adolescentes y adultos)

6.2 Clasificación de las Variables

Tenemos los siguientes tipos de variables:

- **Continuas** Toma valores en un intervalo dado y siempre se mantiene en ese intervalo a lo largo del tiempo por ejemplo el peso y la altura de una persona.
- **Discreta** Toman valores enteros positivos por ejemplos la calificación de una materia.
- **Dicotómicas** Son 2 valores mutuamente excluyentes entre si por ejemplo una persona puede estar viva o muerte pero nunca ambas, generalmente se ocupa el 1 y 0 para representar estas variables.

Para separar a las variables de acuerdo a su clasificación recordemos que, como se menciono anteriormente, el dataset cuenta con variables de tipo entero, flotante y objeto por lo que en esta sección nos dedicamos a separarlas para su adecuado tratamiento.
Como resultado de esta clasificación tenemos dos dataframes:

- **Variables Categóricas:**
  - FOLIO_I
  - sexo
  - riesgo_hipertension
- **Variables Numéricas:**
  - **Enteras (int64):**
    - edad
    - temperatura_ambiente
    - valor_colesterol_hdl
    - valor_colesterol_total
    - valor_trigliceridos
    - resultado_glucosa_promedio
    - tension_arterial
    - sueno_horas
    - actividad_total
  - **Decimales (float64):**
    - concentracion_hemoglobina

- valor_acido_urico
- valor_albumina
- valor_colesterol_ldl
- valor_creatina
- resultado_glucosa
- valor_insulina
- valor_hemoglobina_glucosilada
- valor_ferritina
- valor_folato
- valor_homocisteina
- valor_proteinac_reactiva
- valor_transferrina
- valor_vitamina_bdoce
- valor_vitamina_d
- peso
- estatura
- medida_cintura
- segundamedicion_peso
- segundamedicion_estatura
- distancia_rodilla_talon
- circunferencia_de_la_pantorrilla
- segundamedicion_cintura
- masa_corporal

6.2 Resumen estadistico basico general

Calculamos algunas medidas estadísticas básicas para cada una de las variables de manera general:

	count	mean	std	min	0.25	0.5
sexo	4363.00	1.61	0.49	1.00	1.00	2.00
edad	4363.00	47.32	14.45	4.00	36.00	48.00
concentracion_hemoglobina	4363.00	14.23	1.15	5.50	14.20	14.20
temperatura_ambiente	4363.00	21.39	3.17	2.00	22.00	22.00
valor_acido_urico	4363.00	4.79	0.82	0.20	4.80	4.80
valor_albumina	4363.00	3.87	0.45	1.00	4.00	4.00
valor_colesterol_hdl	4363.00	36.03	8.08	9.00	34.00	34.00
valor_colesterol_ldl	4363.00	87.81	17.53	11.10	86.00	86.00
valor_colesterol_total	4363.00	144.14	28.23	40.00	139.00	139.00
valor_creatina	4363.00	0.61	0.20	0.06	0.58	0.58
resultado_glucosa	4363.00	96.89	45.59	10.40	92.00	92.00
valor_insulina	4363.00	6.74	9.80	0.60	4.00	4.00
valor_trigliceridos	4363.00	137.27	77.75	23.00	123.00	123.00
resultado_glucosa_promedio	4363.00	110.31	32.61	65.00	103.00	103.00
valor_hemoglobina_glucosilada	4363.00	5.45	1.01	3.90	5.20	5.20
valor_ferritina	4363.00	7.65	24.60	0.70	2.70	2.70
valor_folato	4363.00	22.69	4.17	4.20	23.40	23.40
valor_homocisteina	4363.00	5.22	1.38	1.56	4.90	4.90
valor_proteinac_reactiva	4363.00	0.09	0.36	0.02	0.02	0.02
valor_transferrina	4363.00	1.13	0.37	0.10	1.10	1.10
valor_vitamina_bdoce	4363.00	200.83	331.23	33.50	167.00	167.00
valor_vitamina_d	4363.00	21.23	2.95	0.01	20.80	20.80
peso	4363.00	58.29	33.21	2.00	49.30	67.60
estatura	4363.00	156.17	14.71	0.00	151.00	154.50
medida_cintura	4363.00	68.43	46.56	0.00	0.00	90.10
segundamedicion_peso	4363.00	65.79	7.45	2.00	64.70	64.70
segundamedicion_estatura	4363.00	153.82	7.85	0.00	154.00	154.00
distancia_rodilla_talon	4363.00	48.46	3.49	0.00	48.50	48.50
circunferencia_de_la_pantorrilla	4363.00	34.06	4.03	0.00	33.50	33.50
segundamedicion_cintura	4363.00	19.99	40.44	0.00	0.00	0.00
tension_arterial	4363.00	123.73	22.61	0.00	111.00	121.00
sueno_horas	4363.00	3.24	2.78	1.00	2.00	3.00
masa_corporal	4363.00	22.45	12.31	1.00	19.62	26.23
actividad_total	4363.00	481.82	673.70	10.00	240.00	380.00
riesgo_hipertension	4363.00	0.65	0.48	0.00	0.00	1.00

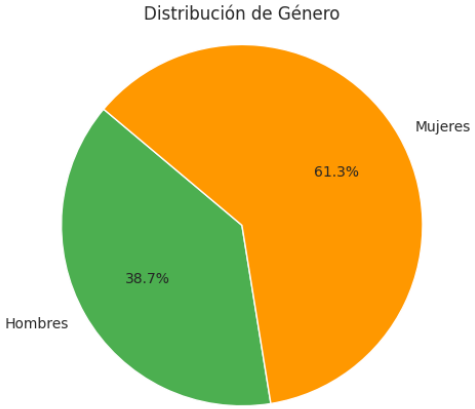
	0.75	max
sexo	2.00	2
edad	58.00	93
concentracion_hemoglobina	14.20	19.9
temperatura_ambiente	22.00	35
valor_acido_urico	4.80	11
valor_albumina	4.00	5.3
valor_colesterol_hdl	34.00	279
valor_colesterol_ldl	86.00	303
valor_colesterol_total	139.00	681
valor_creatina	0.58	8.27
resultado_glucosa	92.00	2372
valor_insulina	4.70	264.1
valor_trigliceridos	123.00	1320
resultado_glucosa_promedio	103.00	1114
valor_hemoglobina_glucosilada	5.20	17.2
valor_ferritina	2.70	613.2
valor_folato	23.40	153
valor_homocisteina	4.90	50
valor_proteinac_reactiva	0.02	10.78
valor_transferrina	1.10	16.2
valor_vitamina_bdoce	167.00	7520
valor_vitamina_d	20.80	50.9
peso	79.60	168.8
estatura	162.90	192
medida_cintura	101.60	189.3
segundamedicion_peso	64.70	151.2
segundamedicion_estatura	154.00	182.6
distancia_rodilla_talon	48.50	97.3
circunferencia_de_la_pantorrilla	33.50	105.2
segundamedicion_cintura	0.00	165
tension_arterial	136.00	200
sueno_horas	4.00	99
masa_corporal	30.29	60.51347647
actividad_total	585.00	17820
riesgo_hipertension	1.00	1

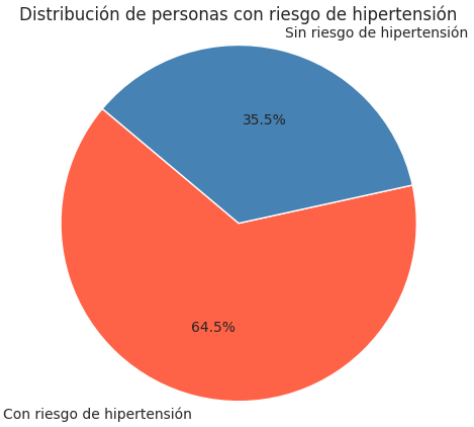
6.2 Resumen estadistico basico de personas con riesgo de hipertención

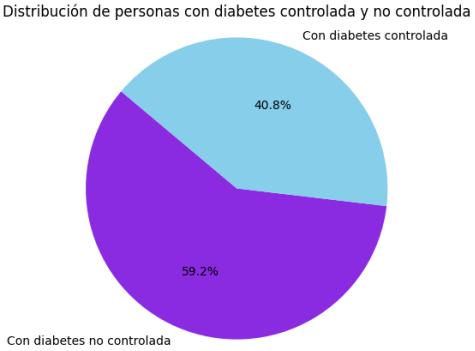
	count	mean	std	min	0.25	0.5
edad	4363.00	47.32	14.45	4.00	36.00	48.00
concentracion_hemoglobina	4363.00	14.23	1.15	5.50	14.20	14.20
temperatura_ambiente	4363.00	21.39	3.17	2.00	22.00	22.00
valor_acido_urico	4363.00	4.79	0.82	0.20	4.80	4.80
valor_albumina	4363.00	3.87	0.45	1.00	4.00	4.00
valor_colesterol_hdl	4363.00	36.03	8.08	9.00	34.00	34.00
valor_colesterol_ldl	4363.00	87.81	17.53	11.10	86.00	86.00
valor_colesterol_total	4363.00	144.14	28.23	40.00	139.00	139.00
valor_creatina	4363.00	0.61	0.20	0.06	0.58	0.58
resultado_glucosa	4363.00	96.89	45.59	10.40	92.00	92.00
valor_insulina	4363.00	6.74	9.80	0.60	4.00	4.00
valor_trigliceridos	4363.00	137.27	77.75	23.00	123.00	123.00
resultado_glucosa_promedio	4363.00	110.31	32.61	65.00	103.00	103.00
valor_hemoglobina_glucosilada	4363.00	5.45	1.01	3.90	5.20	5.20
valor_ferritina	4363.00	7.65	24.60	0.70	2.70	2.70
valor_folato	4363.00	22.69	4.17	4.20	23.40	23.40
valor_homocisteina	4363.00	5.22	1.38	1.56	4.90	4.90
valor_proteinac_reactiva	4363.00	0.09	0.36	0.02	0.02	0.02
valor_transferrina	4363.00	1.13	0.37	0.10	1.10	1.10
valor_vitamina_bdoce	4363.00	200.83	331.23	33.50	167.00	167.00
valor_vitamina_d	4363.00	21.23	2.95	0.01	20.80	20.80
peso	4363.00	58.29	33.21	2.00	49.30	67.60
estatura	4363.00	156.17	14.71	0.00	151.00	154.50
medida_cintura	4363.00	68.43	46.56	0.00	0.00	90.10
segundamedicion_peso	4363.00	65.79	7.45	2.00	64.70	64.70
segundamedicion_estatura	4363.00	153.82	7.85	0.00	154.00	154.00
distancia_rodilla_talon	4363.00	48.46	3.49	0.00	48.50	48.50
circunferencia_de_la_pantorrilla	4363.00	34.06	4.03	0.00	33.50	33.50
segundamedicion_cintura	4363.00	19.99	40.44	0.00	0.00	0.00
tension_arterial	4363.00	123.73	22.61	0.00	111.00	121.00
sueno_horas	4363.00	3.24	2.78	1.00	2.00	3.00
masa_corporal	4363.00	22.45	12.31	1.00	19.62	26.23
actividad_total	4363.00	481.82	673.70	10.00	240.00	380.00

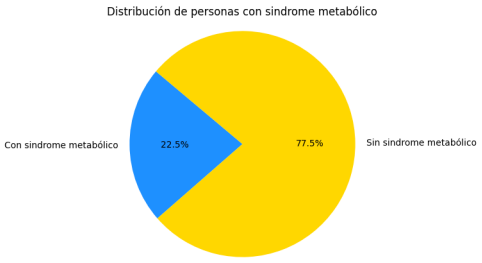
	0.75	max
edad	58.00	93.00
concentracion_hemoglobina	14.20	19.90
temperatura_ambiente	22.00	35.00
valor_acido_urico	4.80	11.00
valor_albumina	4.00	5.30
valor_colesterol_hdl	34.00	279.00
valor_colesterol_ldl	86.00	303.00
valor_colesterol_total	139.00	681.00
valor_creatina	0.58	8.27
resultado_glucosa	92.00	2372.00
valor_insulina	4.70	264.10
valor_trigliceridos	123.00	1320.00
resultado_glucosa_promedio	103.00	1114.00
valor_hemoglobina_glucosilada	5.20	17.20
valor_ferritina	2.70	613.20
valor_folato	23.40	153.00
valor_homocisteina	4.90	50.00
valor_proteinac_reactiva	0.02	10.78
valor_transferrina	1.10	16.20
valor_vitamina_bdoce	167.00	7520.00
valor_vitamina_d	20.80	50.90
peso	79.60	168.80
estatura	162.90	192.00
medida_cintura	101.60	189.30
segundamedicion_peso	64.70	151.20
segundamedicion_estatura	154.00	182.60
distancia_rodilla_talon	48.50	97.30
circunferencia_de_la_pantorrilla	33.50	105.20
segundamedicion_cintura	0.00	165.00
tension_arterial	136.00	200.00
sueno_horas	4.00	99.00
masa_corporal	30.29	60.51
actividad_total	585.00	17820.00

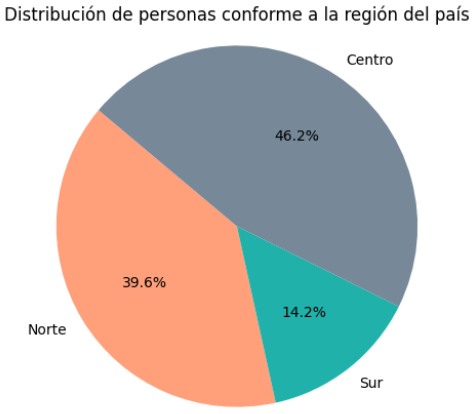
6.2 Graficas para visualización de nuestros datos

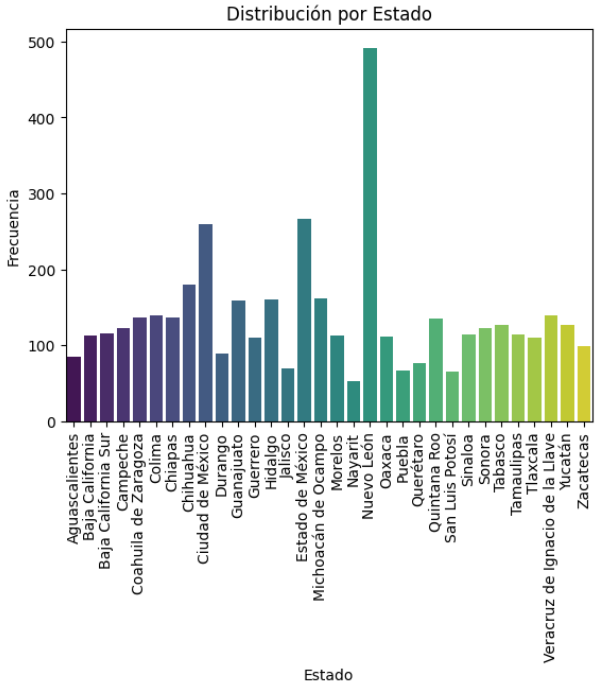










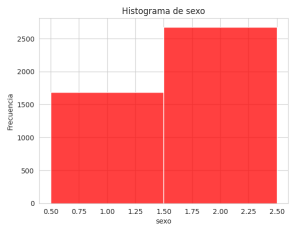


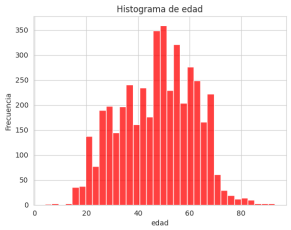
Notamos la presecia de mayor mujeres con 61.3 % que hombres, así como mayor personas con riesgo de hipertension con 64.5 % una mayor cantidad de personas de diabetes por nuestra investigación anticipada con 94.5 % de personas sin diabetes, mediante la investigación previa igual determinamos pacientes con diabetes no controlada que es el mayor porcentaje que tenemos con 59.3 %, 77.5 % de nuestra población no tiene sindrome metabólico y hay mayor observaciones del centro del pais, que es en porcentaje 46.2 %.

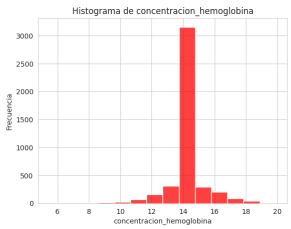
La división de en los estados pertenecientes al norte, sur y centro. Esta dada de la siguiente manera:

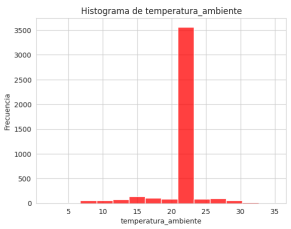
- `"norte": [2: "Baja California", 3: "Baja California Sur", 6: Çolima", 7: Çhiapas", 8: Çhihuahua", 10: "Durango", 19: "Nuevo León", 26: "Sonora", 27: "Tabasco", 28: "Tamaulipas", 32: "Zacatecas"]`
- `çentro": [1: .Aguascalientes", 4: Çampeche", 5: Çoahuila de Zaragoza", 9: Çiudad de México", 11: "Guanajuato", 12: "Guerrero", 13: "Hidalgo", 14: "Jalisco", 15: .Estado de México", 16: "Michoacán de Ocampo", 17: "Morelos", 18: "Nayarit", 21: "Puebla", 22: "Querétaro", 24: "San Luis Potosí", 25: "Sinaloa"]`
- `"sur": [20: .Oaxaca", 23: "Quintana Roo", 29: "Tlaxcala", 30: "Veracruz de Ignacio de la Llave", 31: "Yucatán"]`

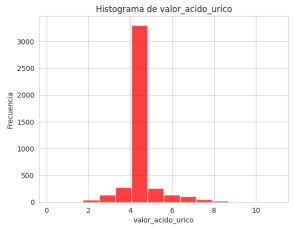
6.2 Histogramas

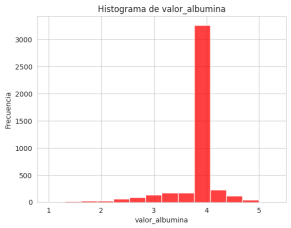


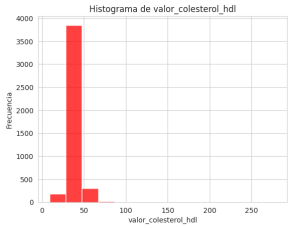


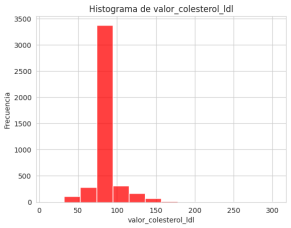


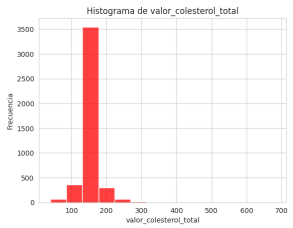


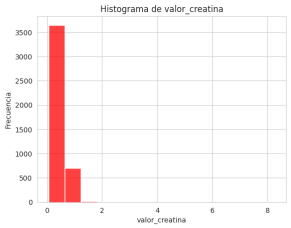


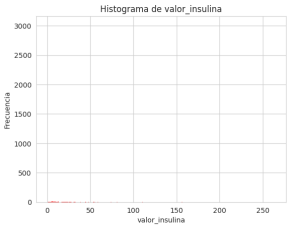


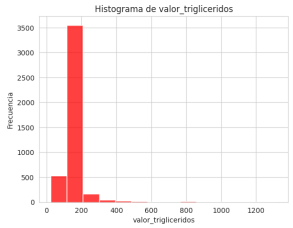


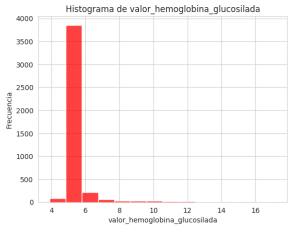


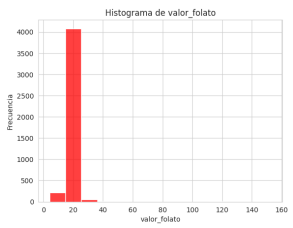


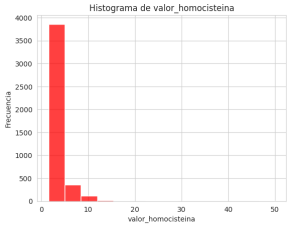


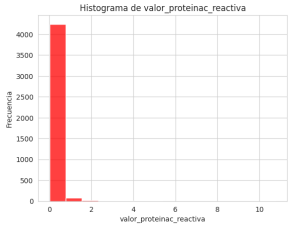


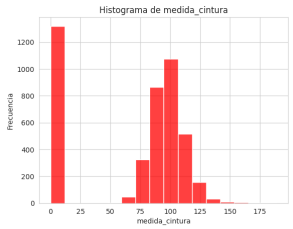


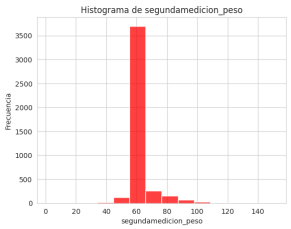


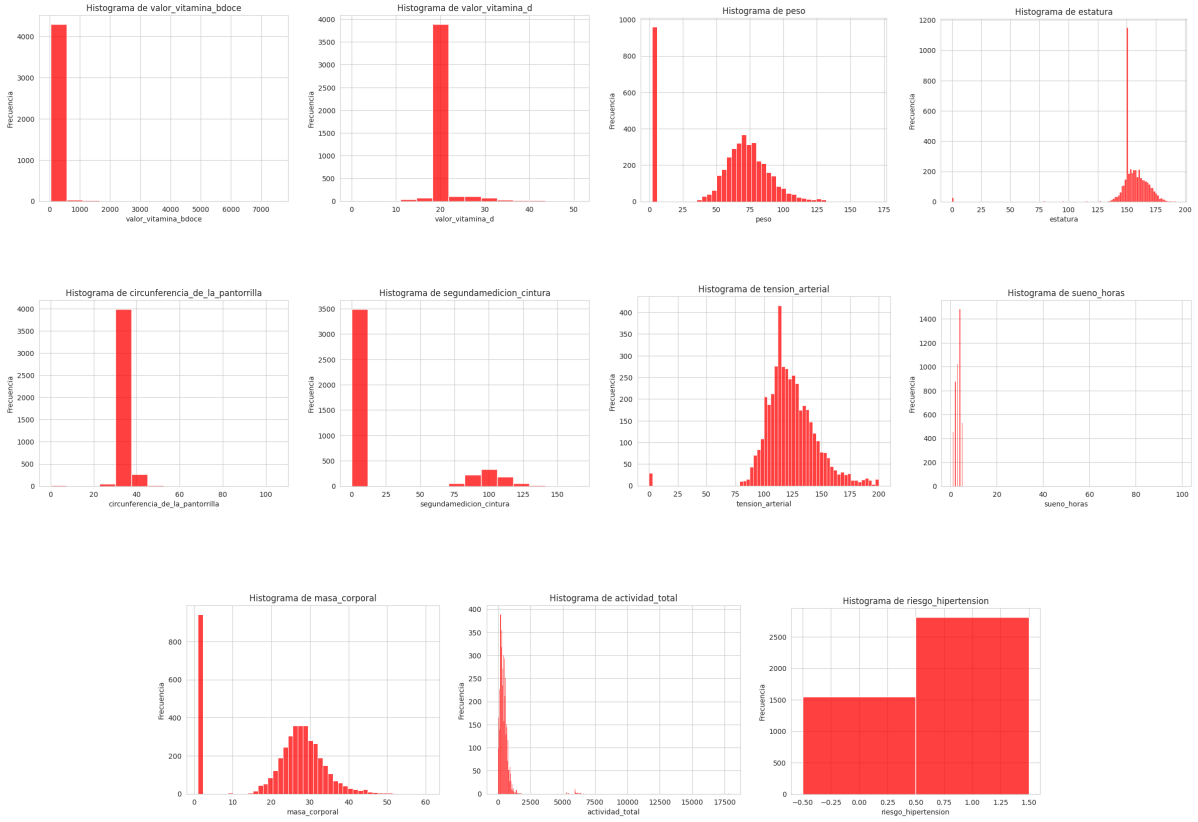




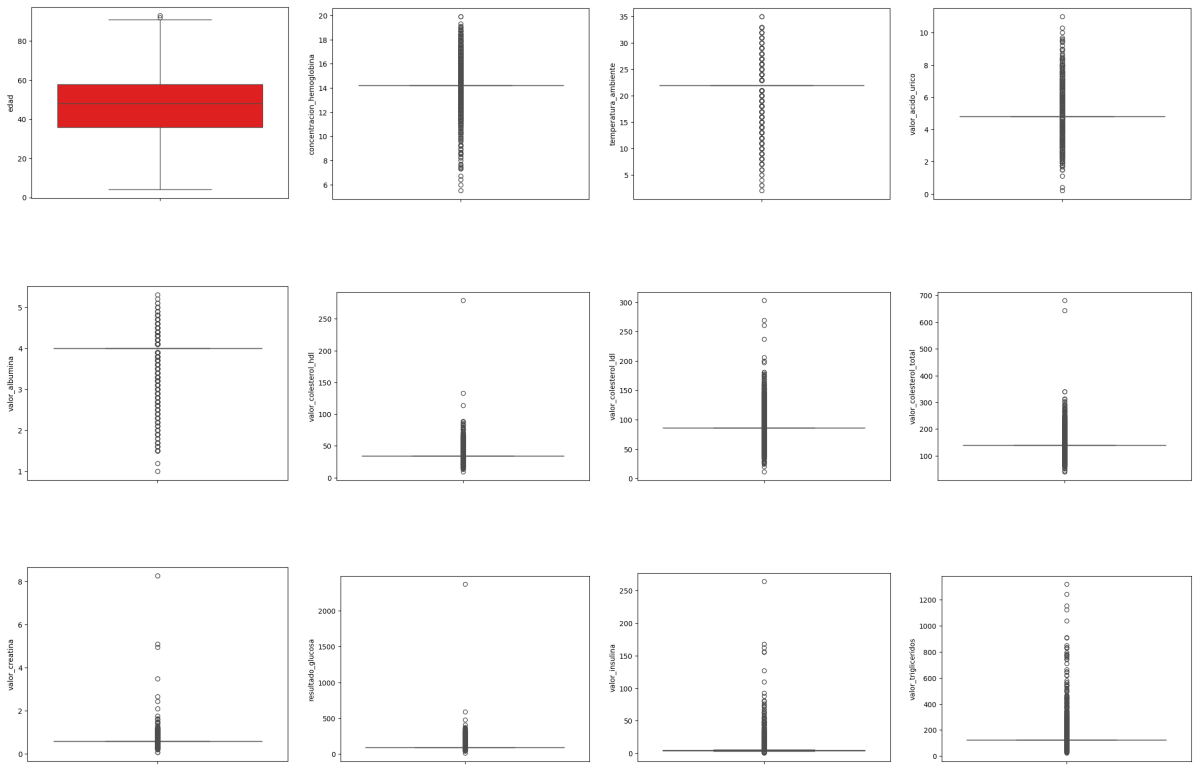


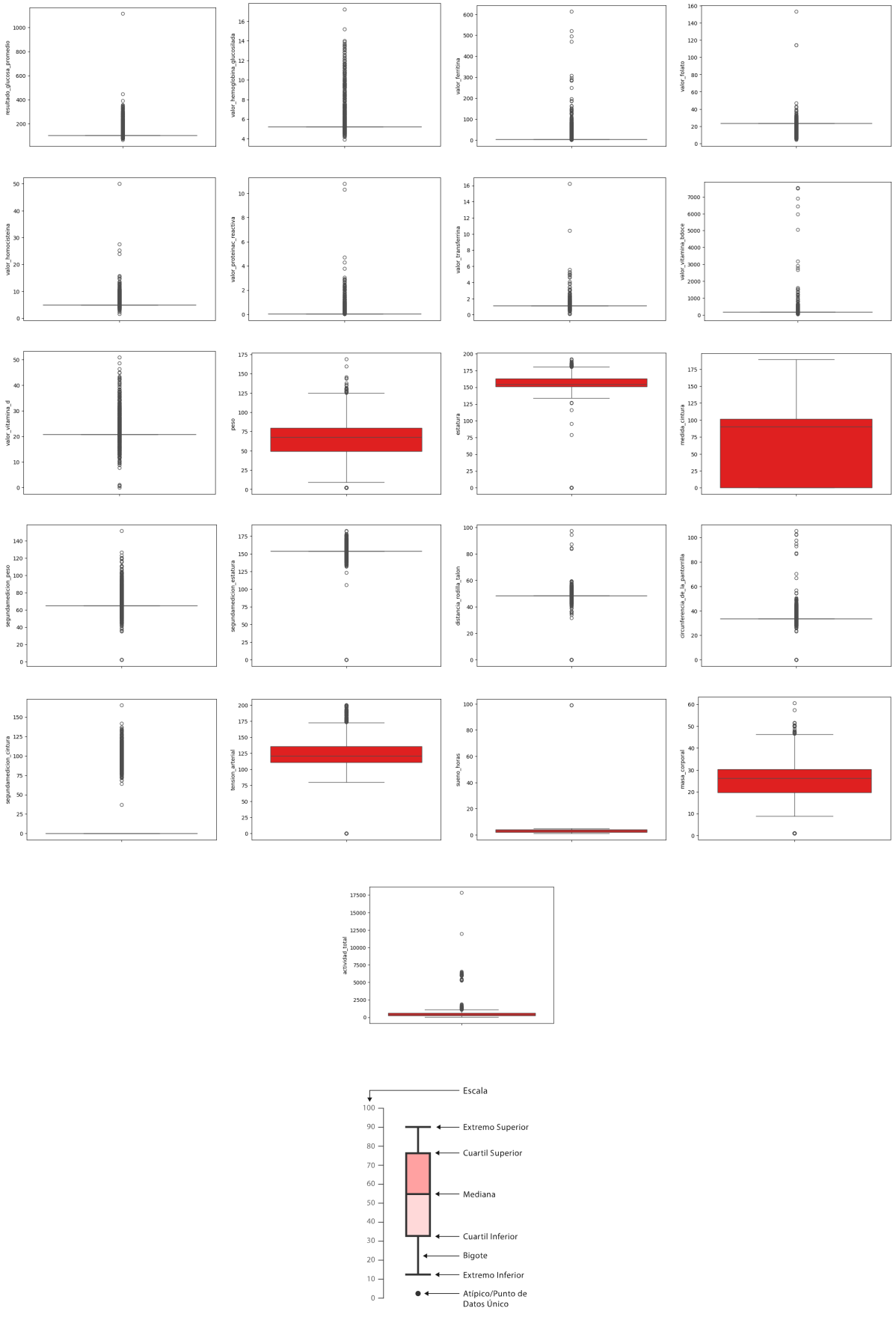






6.2 Graficos de cajas y bigotes




resultado_glucosa_promedio
valor_hemoglobina_glucosilada
valor_ferritina
valor_folato
valor_homocisteina
valor_proteinac_reactiva
valor_transferrina
valor_vitamina_b doce
valor_vitamina_d
peso
estatura
medida_cintura
segundamedicion_peso
segundamedicion_estatura
distancia_rodilla_talon
circunferencia_de_la_pantorrilla
segundamedicion_cintura
tension_arterial
sueno_horas
masa_corporal
actividad_total

Escala
Extremo Superior
Cuartil Superior
Mediana
Cuartil Inferior
Bigote
Extremo Inferior
Atípico/Punto de Datos Único

INTERPRETACIÓN DE LOS DIAGRAMAS DE CAJA Y BIGOTE E HISTOGRAMAS COMO AYUDA VISUAL:

- **Sexo:** Hay mas individuos del sexo Femenino en los datos analizados.
- **Edad:** Es aproximadamente simétrica, con un pico en torno a los 50 años. La mayoría de los datos se concentran entre los 20 y 80 años, indicando que este es el grupo etario predominante en la muestra.
- **Concentracion de Hemoglobina:** La mayoría de los valores se encuentran entre 14 y 15. Representa una baja variación en los niveles de hemoglobina, posiblemente reflejando valores normales en la población.
- **Temperatura Ambiente:** La mayoría de los valores se encuentran entre 21 y 23 grados Celsius. Esto representa un rango típico de temperaturas.
- **Acido Urico:** El valor de ácido úrico muestra una concentración mayoritaria en un rango típico, aunque se identifican valores atípicos altos por encima de 7 mg/dL. Esto podría reflejar individuos con niveles anormales, como en el caso de hiperuricemia.
- **Valor de Albúmina:** Los datos muestran poca variabilidad, concentrándose en un rango estrecho de 3.5 a 4. Esto indica niveles normales de albúmina en la mayoría de la población.
- **Colesterol HDL y LDL**
  - **Colesterol HDL:** La mayoría de los valores de colesterol HDL están agrupados en un rango típico (alrededor de 40-60 mg/dL). Sin embargo, se observan valores atípicos que superan los 100 mg/dL y algunos por debajo del rango saludable.
  - **Colesterol LDL:** Los valores de colesterol LDL muestran un rango intercuartílico ajustado, pero hay varios valores atípicos elevados que superan los 190 mg/dL, lo que puede indicar hipercolesterolemia en algunos individuos.
- **Colesterol Total:** La mediana del colesterol total está dentro de un rango normal, pero se identifican valores atípicos por encima de los 300 mg/dL, lo cual podría indicar hipercolesterolemia severa.
- **Creatinina:** La mayoría de los valores de creatinina están por debajo de 2 mg/dL, lo que es normal, pero se observan valores atípicos elevados que podrían sugerir insuficiencia renal o problemas metabólicos.
- **Glucosa:** La distribución del resultado de glucosa es amplia, con algunos valores atípicos superiores a 1000 mg/dL, lo cual podría reflejar casos severos de hiperglucemia o diabetes no controlada.
- **Insulina:** La mayoría de los valores no tienen una frecuencia alta, se nota la mayor cantidad de datos entre 0 y 50 nivel normal, lo que podría indicar coasos de hipoinsulinemia cuando es menor que 2 y mayores a 50 necesitan atencion medica.

- **Valor Triglicéridos:** Los valores de triglicéridos están principalmente agrupados por debajo de 200 mg/dL, lo que corresponde al rango normal. Sin embargo, hay varios valores atípicos que exceden los 400 mg/dL, indicando hipertrigliceridemia en ciertos individuos.
- **Resultado Glucosa Promedio:** La glucosa promedio presenta una distribución donde la mayoría de los datos están por debajo de 200 mg/dL, pero hay valores atípicos significativos que superan los 1000 mg/dL, lo cual podría reflejar casos severos de hiperglucemia.
- **Hemoglobina Glucosilada:** La hemoglobina glucosilada muestra una mediana en un rango saludable cercano a 6 %, pero se identifican valores atípicos superiores al 12 %, lo cual es un indicativo de un mal control de la glucosa en algunos pacientes.
- **Valor Ferritina:** La ferritina tiene un rango amplio de datos con algunos valores atípicosque exceden los 400 ng/mL, lo que podría estar asociado a trastornos como la sobrecarga de hierro o inflamación crónica.
- **Valor Folato:** Los valores predominan entre 0 y 20, lo cual es un rango típico. Sin embargo, se observan algunos valores más altos que podrían estar relacionados con suplementación.
- **Valor Homocisteína:** La mayoría de los valores están por debajo de 10, lo que refleja niveles normales. Algunos valores altos podrían indicar riesgo cardiovascular aumentado.
- **Valor Proteína C Reactiva:** La mayoría de los valores están cercanos a cero, lo que sugiere baja inflamación en la mayoría de los individuos. Sin embargo, hay algunos valores elevados, posiblemente asociados con inflamación o infecciones.
- **Valor Transferrina:** Los valores se agrupan entre 0-2, lo cual es típico. Algunos valores más altos podrían sugerir deficiencia de hierro o alta capacidad de transporte de hierro.
- **Medida Cintura:** Los valores predominan entre 70 y 110 cm, lo que es típico en adultos. Algunos valores más altos podrían estar relacionados con obesidad abdominal. Notar que hay alta freciencia en medida de cinturas muy pequeñas que se trabajaran en le codigo.
- **Segunda Medición Peso:** La distribución está centrada entre 60 kg, lo cual es un rango típico para el peso corporal en adultos. Algunos valores extremos podrían representar casos de obesidad o bajo peso.
- **Segunda Medición Estatura:** La mayoría de los valores están cerca de 150-175cm, lo que corresponde al rango promedio de estatura para adultos.
- **Distancia Rodilla-Talón:** Los valores se concentran cerca de 40 y 60cm. meida normal, pero hay valores pequeños cerca de cero.
- **Valor Vitamina B12:** Se encuentran mayormente entre 250 lo cual los valores preocupantes son los mayores a 950.

- **Valor Vitamina D:** Se concentra el 20 loa niveles son normales de 20 a 40. Notando poblacion con niveles bajos.
- **Peso:** Se distribuye casi como una normal, se necesitaria hacer una prueba de hipotesis para asegurarlo. Notando valores cercanos a cero que no prodrian ser posibles, ya que delimitamos la poblacion a mayores de 14 años.
- **Estatura:** Igual tiene distribución parecida a la normal pero con una mayor freciencia en 150 aproximadamente.
- **Circunferencia de la pantorrilla:** Se concentra en 30
- **Segunda medicion de Cintura:** Notamos igual valores cercanos a cero que no pueden ser posibles
- **Tension Arterial:** Se distribuye como una normal o t-student, notando valores atipicos cercanos a cero.
- **Sueño en horas:** Notamos horas menores a 10 e incluso 8, i agunos cercanos a cero.
- **Masa Corporal:** Notamos una ditribución parecida a la normal o t-student pero con valores atipicos cercanos a cero debido a nuestra población definida.
- **Actividad Total:** Notamos que la cantidad de minutos de actividad se concentran mayormente cercanos a cero y disminuye.
- **Riesgo de Hipertensión:** Hay más población con riesgo de hipertensión.

6.3 Z-Score

Para este modelo utilizamos mahalanibis robusta, ya que nos parece la mejor manera para quitar outliers. Sin embargo, tambien hicimos el calculo mediante Z-core modificado para tener una referencia.
El Z-score modificado es una variante del Z-score diseñada para ser más robusta frente a outliers, ya que no depende de la media ni de la desviación estándar, sino de la mediana y la desviación absoluta de la mediana (MEDA). Esto lo hace menos sensible a los valores extremos.
El umbral empleado fue de 5, identificando los valores que exceden este límite como atípicos. A continuación, se presenta una tabla que resume el número de valores atípicos detectados por columna:

Columna	Número de Atípicos
Concentración de Hemoglobina	218
Temperatura Ambiente	193
Valor Ácido Úrico	268
Valor Albúmina	113
Colesterol HDL	96
Colesterol LDL	199
Colesterol Total	129
Creatinina	81
Resultado Glucosa	140
Insulina	98
Triglicéridos	113
Glucosa Promedio	148
Hemoglobina Glucosilada	148
Ferritina	130
Folato	229
Homocisteína	164
Proteína C Reactiva	128
Transferrina	145
Vitamina B12	68
Vitamina D	241
Estatura	25
Segunda Medición Peso	211
Segunda Medición Estatura	286
Tensión Arterial	29
Sueño Horas	3
Actividad Total	45
Diabetes	238

Cuadro 1: Resumen de valores atípicos por columna utilizando el Z-score modificado con umbral 5.

- Las columnas con mayor número de valores atípicos son **Segunda Medición de Estatura (286)**, **Vitamina D (241)** y **Diabetes (238)**. Esto indica una alta variabilidad en estas variables, posiblemente debido a errores de medición o características particulares de la muestra analizada.
- Variables como **Sueño Horas (3)** y **Estatura (25)** presentan un bajo número de valores atípicos, lo que sugiere que la mayoría de los datos están dentro de un rango esperado.
- Las mediciones bioquímicas como **Colesterol HDL**, **Colesterol LDL** y **Triglicéridos** tienen un número moderado de atípicos, lo cual podría reflejar la presencia de individuos con trastornos metabólicos.

7 Tratamiento de datos erróneos

7.1 ¿En busca de errores o de metahumanos?...

En esta sección eliminaremos aquellos datos que, de ser correctos, implicarían graves afectaciones para el ser humano que puedan complicar el hecho mismo de ser entrevistados o inclusive, de ser humanos. Se realiza un análisis exhaustivo para identificar y corregir datos erróneos en variables críticas como niveles de glucosa, concentración de hemoglobina, y otros indicadores clínicos. Esto tiene como objetivo asegurar la calidad y la confiabilidad de los datos utilizados en la investigación.

Tensión arterial y peso corporal: Se eliminaron registros con valores de *tensión arterial* menores o iguales a 0, y *peso corporal* menor o igual a 2 kg, ya que representan errores en los datos:
El tratamiento aplicado permitió identificar y corregir valores atípicos o inconsistentes en las variables seleccionadas. Estos pasos son esenciales para evitar sesgos en los análisis posteriores y asegurar la validez de las conclusiones del estudio. Por ejemplo, eliminar datos fuera del rango fisiológico esperado reduce el riesgo de interpretar errores como hallazgos clínicamente significativos.
En conclusión, este proceso no solo depuró el conjunto de datos, sino que también destacó posibles áreas de mejora en la recolección de datos futuros, como un mayor control en el ingreso de medidas físicas (ej. *estatura, medida cintura*) y resultados clínicos (ej. *diabetes*).

7.1 Medida de Cintura

La variable **medida de cintura** presenta valores nulos o igual a cero en algunas observaciones, lo que requiere un proceso de limpieza para garantizar la calidad de los datos y su utilidad en los análisis posteriores.
Para los registros en los que la primera medición (*medida_cintura*) es igual a cero, se sustituyó este valor con la segunda medición disponible (*segundamedicion_cintura*). De manera inversa, si la segunda medición es cero, se utilizó la primera medición para completarla. Este procedimiento asegura que al menos una medición válida esté presente para cada observación. Después de realizar el reemplazo de valores faltantes, se verificó la consistencia entre las dos

mediciones. La comparación indica que **en todas las observaciones las dos mediciones coinciden perfectamente**, lo que confirma que siempre se cuenta con una medición válida:

- Resumen estadístico de la diferencia entre las dos mediciones: Resultados:
  - **Promedio:** 0.0
  - **Desviación estándar:** 0.0
  - **Máximo:** 0.0
  - **Mínimo:** 0.0

  Resultados:
  - **Total de observaciones:** 3361
  - **Frecuencia de coincidencia:** 100 %

Dado que ambas mediciones son ahora idénticas, se elimina la columna *segundamedicion_cintura* para mantener únicamente la variable *medida_cintura*: Resumen estadístico de la variable *medida_cintura*:

- **Media:** 88.01 cm
- **Desviación estándar:** 32.67 cm
- **Mínimo:** 0.0 cm (valores potencialmente problemáticos)
- **Máximo:** 189.3 cm

Se identificaron valores igual a 0, considerados erróneos, los cuales fueron eliminados: Posteriormente, se aplicó el **Z-score modificado** con un umbral > 5 para detectar valores atípicos. Estos valores representan mediciones extremadamente altas de la circunferencia de cintura, que podrían reflejar casos reales extremos o errores en los datos:

- **Valores atípicos detectados:** 12 observaciones.
- **Rango de valores atípicos:** entre 152.1 cm y 189.3 cm.

Resultado del cálculo:

- **Mediana de la desviación absoluta (MEDA):** 10.57.
- Valores atípicos detectados:

```
[183.2, 152.1, 162.6, 154.0, 157.8, 158.0,
 154.0, 189.3, 164.0, 168.0, 179.9, 174.3]
```

El tratamiento de la variable *medida_cintura* garantiza la eliminación de valores faltantes o erróneos, manteniendo únicamente registros confiables para los análisis posteriores. La identificación de valores atípicos también destaca casos extremos que podrían requerir análisis adicionales para determinar su naturaleza, ya sea como mediciones válidas de casos reales o errores en el registro.

7.1 Medida de la estatura

Para la variable *estatura*, se utilizó el **Z-score modificado** con el objetivo de identificar valores atípicos. El análisis inicial arrojó los siguientes resultados:

- **Valores atípicos detectados:** $\{0.0, 0.0, 0.0, 116.3\}$.

Los valores iguales a $0{,}0$ sugieren datos faltantes o errores de registro, mientras que el valor $116{,}3$ destaca como un *outlier* significativo.
Para abordar los valores de $0{,}0$, se implementó un procedimiento que aprovecha la variable *segundamedicion_estatura*, bajo la premisa de que esta medición secundaria puede complementar los datos faltantes en *estatura*. El procedimiento consistió en lo siguiente:

1. Para cada observación con valor $0{,}0$ en *estatura*, este fue reemplazado por el valor correspondiente de *segundamedicion_estatura*.
2. De manera análoga, para cada observación con valor $0{,}0$ en *segundamedicion_estatura*, este fue reemplazado por el valor correspondiente de *estatura*.

Después de aplicar el procedimiento descrito, se realizó nuevamente el cálculo del Z-score modificado para identificar valores atípicos restantes. Los resultados obtenidos son los siguientes:

- **Mediana de desviaciones absolutas (MAD):** $7{,}62$.
- **Valores atípicos detectados:** $\{116{,}3\}$.

Esto confirma que el único valor considerado atípico después del tratamiento es $116{,}3$, lo cual es consistente con el análisis inicial. Los valores cero fueron eliminados correctamente.
Después de realizar el tratamiento de datos faltantes o erróneos en las variables *estatura* y *segundamedicion_estatura*, se llevó a cabo un análisis de las diferencias entre ambas mediciones con el objetivo de evaluar la consistencia entre estas. El análisis estadístico de las diferencias entre las dos mediciones de *estatura* arrojó los siguientes resultados:

Estadístico	Valor
Cantidad de observaciones (count)	3014
Media (mean)	4.7827
Desviación estándar (std)	9.4045
Mínimo (min)	-37.7
Primer cuartil (25 %)	-2.1
Mediana (50 %)	4.0
Tercer cuartil (75 %)	11.5
Máximo (max)	38.0

Cuadro 2: Estadísticos descriptivos de la diferencia entre *estatura* y *segundamedicion_estatura*.

- La media de la diferencia entre las dos mediciones es de $4{,}78$ cm, lo que sugiere una ligera discrepancia sistemática entre ambas mediciones.

- La mediana es de $4{,}0$ cm, lo cual indica que la mayor parte de las diferencias están distribuidas alrededor de este valor.
- El rango intercuartílico (IQR) oscila entre $-2{,}1$ y $11{,}5$ cm, mostrando que el 50 % central de los datos presenta diferencias moderadas.
- La diferencia mínima es de $-37{,}7$ cm y la máxima es de $38{,}0$ cm, valores que podrían ser atribuibles a errores de medición o casos atípicos específicos.

Se generó un histograma para visualizar la distribución de las diferencias entre las dos mediciones:

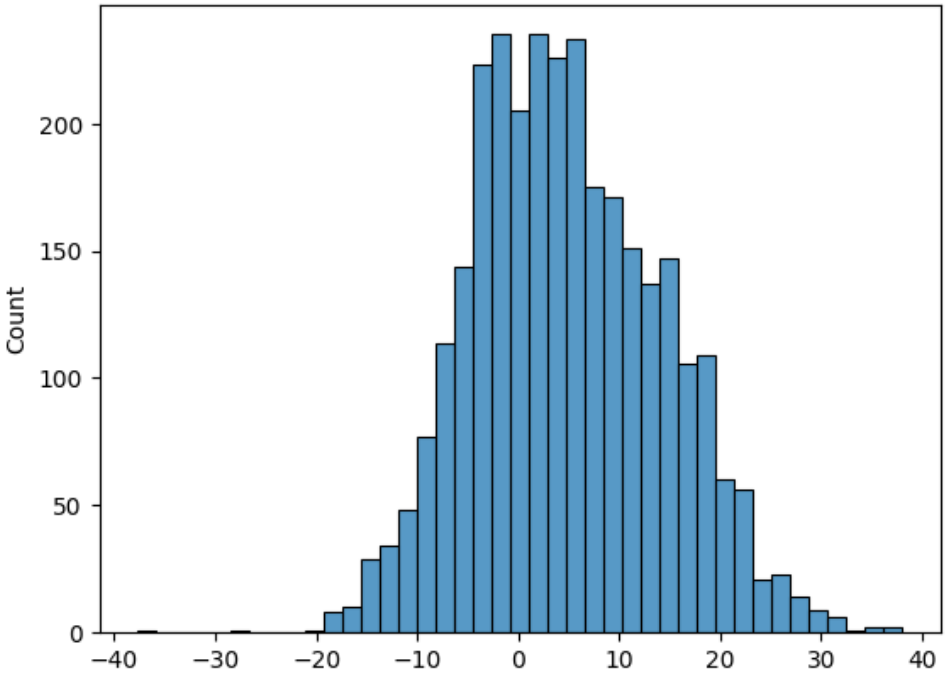


Figura 24: Histograma de la diferencia entre *estatura* y *segundamedicion_estatura*.

La mayoría de las diferencias entre las dos mediciones se concentran cerca del promedio ($4{,}78$ cm), aunque existen casos extremos que deben ser evaluados para garantizar la calidad de los datos. Este análisis refuerza la necesidad de un tratamiento cuidadoso de los datos antes de utilizarlos en análisis posteriores.

7.1 Respecto al peso

Se realizó un análisis descriptivo de las variables *peso* y *segundamedicion_peso*. Los estadísticos obtenidos son los siguientes:

Estadístico	Peso	Segunda Medición de Peso
Cantidad de observaciones (count)	3014	3014
Media (mean)	75.08	64.70
Desviación estándar (std)	16.03	0.00
Mínimo (min)	32.05	64.70
Primer cuartil (25 %)	64.15	64.70
Mediana (50 %)	73.20	64.70
Tercer cuartil (75 %)	84.18	64.70
Máximo (max)	168.80	64.70

Cuadro 3: Estadísticos descriptivos de las mediciones de peso.

- Se observó que la variable *segundamedicion_peso* es constante con un valor de $64{,}70$ kg para todas las observaciones. Por lo tanto, esta columna fue eliminada del conjunto de datos.
- Posteriormente, se graficó un histograma para visualizar la distribución de la variable *peso*.

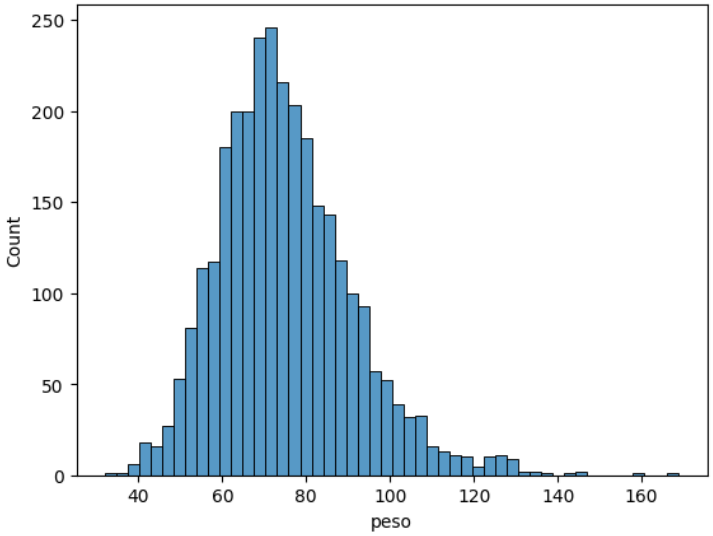


Figura 25: Histograma de la variable *peso*.

El Índice de Masa Corporal (IMC) se calculó utilizando la fórmula:

$$\text{IMC} = \frac{\text{peso}}{\left(\frac{\text{estatura}+\text{segundamedicion\_estatura}}{2} \times \frac{1}{100}\right)^2}$$

Se generó un histograma para observar la distribución del *IMC*:

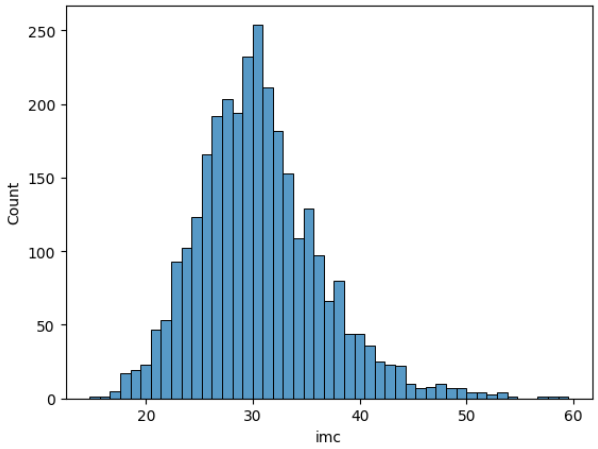


Figura 26: Histograma de la variable *IMC*.

Para validar la consistencia de los datos, se calculó la diferencia entre el *IMC* recién calculado y la variable *masa_corporal*. La diferencia fue graficada:

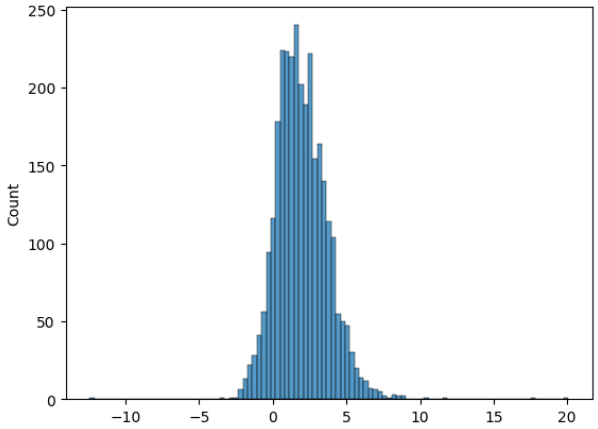


Figura 27: Histograma de la diferencia entre *IMC* calculado y *masa_corporal*.

Adicionalmente, se graficó la distribución de *masa_corporal*:

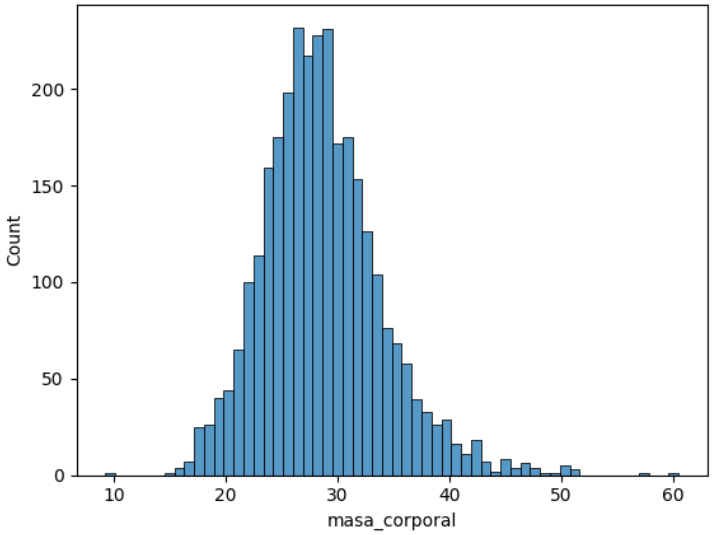


Figura 28: Histograma de la variable *masa_corporal*.

- El *IMC* presenta una distribución acorde a lo esperado, mostrando su utilidad para la categorización de los datos en términos de normopeso, sobrepeso y obesidad.
- Las diferencias entre el *IMC* calculado y la variable *masa_corporal* fueron mínimas, lo que sugiere una coherencia en los datos.
- Se eliminó la columna *segundamedicion_peso* debido a su falta de variabilidad.

7.1 Respecto a las horas de sueño

Para identificar posibles valores atípicos en la variable *sueno_horas*, se aplicó el método de Z-modificado. Este método utiliza la desviación absoluta de la mediana (MAD) para calcular un criterio robusto y menos sensible a valores extremos.

El cálculo resultó en un valor de meda $= 0{,}9768$, pero no se detectaron valores atípicos significativos (Z-modificado $= []$) según el umbral adoptado (> 5).

Se calculó la moda de la variable *sueno_horas*, excluyendo los valores atípicos codificados como 99. El resultado fue:

$$\text{Moda}(\text{sueno\_horas}) = 4\,\text{horas}.$$

En el conjunto de datos original, los valores 99 en la columna *sueno_horas* representan valores anómalos o faltantes. Para su tratamiento, estos valores se reemplazaron con la moda de la variable (4 horas):

Posteriormente, se reiniciaron los índices del DataFrame para mantener la coherencia estructural.

Después de la imputación de valores faltantes, la variable *sueno_horas* fue actualizada y lista para su análisis. Los pasos realizados aseguran que no haya valores atípicos ni inconsistencias que puedan afectar los análisis posteriores.

- La variable *sueno_horas* no presentó valores atípicos según el método de Z-modificado.
- Los valores faltantes representados por 99 se imputaron utilizando la moda, con lo cual se preservó la consistencia en los datos.
- Estos ajustes permiten que la variable sea utilizada de manera confiable en análisis posteriores.

8 Dependencias

El análisis de dependencia entre variables se realizó utilizando dos enfoques principales: un mapa de calor de la matriz de correlación y la identificación de pares de variables con alta correlación. Se utilizó el método de correlación de Spearman, que es robusto frente a relaciones no lineales.

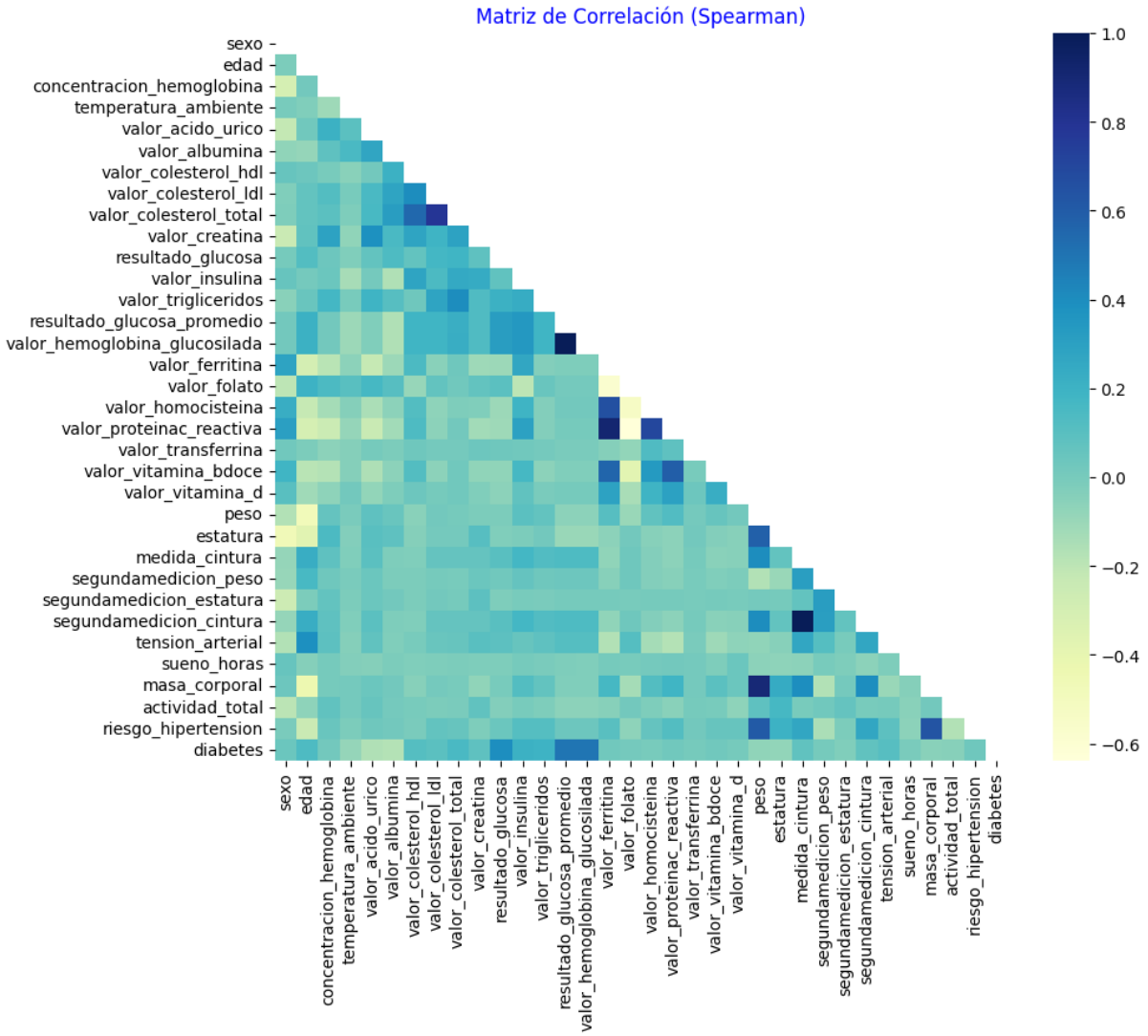


Figura 29: Histograma de la variable *masa_corporal*.

A continuación se presentan las correlaciones más relevantes encontradas en los datos:

- **Colesterol LDL y Colesterol Total**:
    - El valor_colesterol_ldl está altamente correlacionado con valor_colesterol_total. Este resultado es esperado, ya que el colesterol LDL es una fracción del colesterol total y, por lo tanto, debe mostrar una relación significativa con él.
- **Ferritina y Proteína C Reactiva**:
    - Se observa una alta correlación entre valor_ferritina y valor_proteinac_reactiva. Esto puede sugerir una relación entre los niveles de ferritina (un marcador de almacenamiento de hierro) y la inflamación en el cuerpo, ya que la proteína C reactiva es un indicador de inflamación.
- **Peso, Cintura y Masa Corporal**:
    - El peso está altamente correlacionado con la medida_cintura y la masa_corporal. Este hallazgo es esperado, ya que el peso corporal suele estar relacionado de manera directa con la masa corporal total y las mediciones de cintura, que son indicadores comunes de la obesidad o el exceso de peso.
- **Masa Corporal y Medida de Cintura**:
    - La masa_corporal muestra una fuerte correlación con peso y medida_cintura. Esto es lógico, ya que tanto el peso como las mediciones de la cintura son factores directamente asociados a la masa corporal.
- **Actividad total y horas de sueño**:
    - Algunas variables como sueno_horas y actividad_total parecen mostrar correlaciones débiles con las demás variables, lo que podría indicar una menor influencia sobre los factores fisiológicos medidos.

9 Particion

Dado que los rangos adecuados de mediciones tales como la concentración de hemoglobina o el nivel de colesterol hdl son diferentes para ambos sexos y, teniendo en cuenta que las mujeres conforman más del 60 por ciento de los datos, se decide particionar la muestra haciendo distinción entre hombres y mujeres.

- Se crea el DataFrame df_H que contiene los datos correspondientes a los hombres (sexo $= 1$), y se elimina la columna sexo de este DataFrame. Posteriormente, se restablece el índice.

- De forma similar, se crea el DataFrame df_M para las mujeres (sexo = 2), también eliminando la columna sexo y restableciendo el índice.

Además, se identifican las columnas con varianza baja en el DataFrame de hombres df_H. Esto es útil para descartar variables que no aportan información significativa debido a la falta de variabilidad en los datos. Se considera que una columna tiene varianza baja si su varianza es menor que un umbral basado en el 1 % del rango de la columna.

- Las columnas con varianza baja identificadas en `df_H` son las siguientes:
  - `valor_creatina`
  - `valor_ferritina`
  - `valor_folato`
  - `valor_homocisteina`
  - `valor_proteinac_reactiva`
  - `valor_transferrina`
  - `valor_vitamina_bdoce`
  - `valor_vitamina_d`
  - `segundamedicion_peso`
  - `segundamedicion_estatura`

  Estas variables muestran una varianza extremadamente baja, lo que sugiere que sus valores son constantes o muy limitados, y por lo tanto no proporcionan suficiente información para el análisis.
- Las columnas constantes en el DataFrame `df_H` son:
  - `valor_ferritina`
  - `valor_folato`
  - `valor_homocisteina`
  - `valor_proteinac_reactiva`
  - `valor_transferrina`
  - `valor_vitamina_bdoce`
  - `valor_vitamina_d`
  - `segundamedicion_peso`
  - `segundamedicion_estatura`

  Estas variables deben ser consideradas para su posible eliminación del análisis posterior, ya que no aportan variabilidad significativa y podrían no ser útiles en los modelos predictivos.

9.1 Varianza Baja

Al analizar las columnas con baja varianza en las particiones por sexo, se identificaron las siguientes:

- Para **hombres** (`df_H`), las columnas con varianza baja son:
  - `valor_creatina`
  - `valor_ferritina`
  - `valor_folato`
  - `valor_homocisteina`
  - `valor_proteinac_reactiva`
  - `valor_transferrina`
  - `valor_vitamina_bdoce`
  - `valor_vitamina_d`
  - `segundamedicion_peso`
  - `segundamedicion_estatura`
- Para **mujeres** (`df_M`), las columnas con varianza baja son:
  - `valor_creatina`
  - `segundamedicion_peso`
  - `segundamedicion_estatura`

Dado que estas variables tienen una varianza extremadamente baja, se eliminaron de ambos subconjuntos (`df_H` y `df_M`) para evitar que afecten el análisis posterior, ya que no aportan información significativa.
Se identificaron las siguientes correlaciones significativas (*coeficiente Spearman mayor o igual a 0.76*) en el conjunto de datos de hombres:

- valor_colesterol_ldl altamente correlacionado con valor_colesterol_total.
- valor_colesterol_total altamente correlacionado con valor_colesterol_ldl.
- resultado_glucosa_promedio altamente correlacionado con valor_hemoglobina_glucosilada.
- valor_hemoglobina_glucosilada altamente correlacionado con resultado_glucosa_promedio.
- peso altamente correlacionado con medida_cintura y masa_corporal.

- medida_cintura altamente correlacionado con peso y masa_corporal.

De manera similar, las correlaciones significativas en el conjunto de datos de mujeres son:

- valor_colesterol_ldl altamente correlacionado con valor_colesterol_total.
- valor_colesterol_total altamente correlacionado con valor_colesterol_ldl.
- resultado_glucosa_promedio altamente correlacionado con valor_hemoglobina_glucosilada.
- valor_hemoglobina_glucosilada altamente correlacionado con resultado_glucosa_promedio.
- peso altamente correlacionado con medida_cintura y masa_corporal.
- medida_cintura altamente correlacionado con peso y masa_corporal.
- masa_corporal altamente correlacionado con peso y medida_cintura.

La eliminación de variables con baja varianza permitió simplificar los datos, centrándose en aquellas que aportan información relevante. Las correlaciones encontradas muestran patrones claros entre variables relacionadas, como el colesterol total y LDL, o el peso y las medidas corporales. Estas relaciones son consistentes entre hombres y mujeres, lo que puede ser útil para modelar factores de riesgo comunes o específicos del sexo en estudios posteriores.

9.2 Variables para la matriz de Varianza

Para que la matriz de varianza sea invertible, es necesario que tenga rango completo, lo que implica que no debe contener columnas linealmente dependientes. Por esta razón, se eliminaron las variables que presentaron una alta correlación entre sí, conservando aquellas con una mayor correlación con la variable de interés: el riesgo de hipertensión (representado en este caso por el índice de masa corporal `imc`).

Se calcularon las correlaciones de las variables seleccionadas con el índice de masa corporal (`imc`) utilizando el método de Spearman. Los resultados fueron los siguientes:

- resultado_glucosa_promedio: correlación con imc = 0.0953
- valor_hemoglobina_glucosilada: correlación con imc = 0.0953
- medida_cintura: correlación con imc = 0.9005
- valor_colesterol_ldl: correlación con imc = 0.0679
- valor_colesterol_total: correlación con imc = 0.0365

Dado que `medida_cintura` presentó la correlación más alta con el `imc`, se conservará esta variable. Las variables `resultado_glucosa_promedio` y `valor_colesterol_total`, que tienen correlaciones más bajas, fueron eliminadas del conjunto de datos para hombres (`df_H`).

De manera similar, se calcularon las correlaciones de las variables con el índice de masa corporal (`imc`) en el conjunto de datos de mujeres (`df_M`), obteniéndose los siguientes valores:

- `resultado_glucosa_promedio`: correlación con `imc` = 0.0896
- `valor_hemoglobina_glucosilada`: correlación con `imc` = 0.0897
- `medida_cintura`: correlación con `imc` = 0.8687
- `valor_colesterol_ldl`: correlación con `imc` = 0.0431
- `valor_colesterol_total`: correlación con `imc` = 0.0330

Nuevamente, medida_cintura mostró la correlación más alta con el imc, por lo que se conservó esta variable. Las variables resultado_glucosa_promedio y valor_colesterol_total fueron eliminadas del conjunto de datos para mujeres (df_M).

9.3 Valores atípicos y correlaciones

Para identificar valores atípicos en el conjunto de datos, se utilizó la distancia de Mahalanobis robusta, que es resistente a la influencia de los outliers gracias al estimador de covarianza mínimo determinante (`MinCovDet`). El proceso se describe a continuación:

9.3 Distancia de Mahalanobis Robusta

La distancia de Mahalanobis robusta fue utilizada para identificar y eliminar valores atípicos en las variables `peso`, `medida_cintura`, `estatura` e `imc`. Este enfoque asegura que el análisis subsiguiente esté basado en datos limpios y representativos. A continuación, se describen los pasos del proceso:

1. Se eliminó la columna `FOLIO_I`, ya que no aporta información para el cálculo de distancias.
2. Las variables de interés (`peso`, `medida_cintura`, `estatura`, `imc`) fueron seleccionadas del conjunto de datos `df_H`.
3. Se utilizó el estimador de covarianza mínimo determinante (`MinCovDet`) con un estado aleatorio fijo (`random_state=8`) para garantizar reproducibilidad.
4. Se calculó la distancia de Mahalanobis para cada observación:

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

   donde μ es el centroide robusto y Σ^{-1} es la inversa de la matriz de covarianza robusta.
5. Se calcularon los *p-valores* asociados a estas distancias usando la distribución χ^2 con grados de libertad igual al número de columnas del subconjunto de datos.
6. Las observaciones con un *p-valor* menor a 0.05 fueron identificadas como valores atípicos.

Los siguientes valores atípicos fueron detectados y eliminados del conjunto de datos:

Índice	FOLIO_I	Peso	Medida Cintura	Estatura
270	2022_08019067	117.05	183.2	183.2
313	2022_09005027	75.40	162.6	162.4
761	2022_19031104	124.70	189.3	170.8
841	2022_20339014	87.80	116.3	116.3
845	2022_20499002	71.20	168.0	168.0
985	2022_26030029	98.80	179.9	179.9
990	2022_26030121	89.10	174.3	174.3

Cuadro 4: Observaciones identificadas como outliers (Iteración 1).

Estas observaciones fueron eliminadas del conjunto de datos y los índices fueron reiniciados.

En una segunda iteración con las variables mencionadas, se detectó el siguiente valor atípico:

Índice	FOLIO_I	Peso	Medida Cintura	IMC
44	2022_02003004	168.8	147.2	56.694707

Cuadro 5: Observación identificada como outlier (Iteración 2).

Esta observación también fue eliminada y los índices fueron reiniciados.

Se analizaron las variables relacionadas con la *tensión arterial*, el *riesgo de hipertensión*, *peso*, *medida de cintura*, y *estatura*. A continuación, se presentan los valores atípicos detectados en diferentes iteraciones: Los siguientes valores fueron identificados como outliers:

Índice	ID	Tensión Arterial	Riesgo Hipertensión
40	2022_02002149	193	0
159	2022_05030037	189	0
304	2022_08064001	200	0

Cuadro 6: Valores atípicos eliminados en la primera iteración.

Estas observaciones fueron eliminadas, y los índices fueron reiniciados para garantizar la consistencia del conjunto de datos. El mismo proceso se realizó para el conjunto de datos de mujeres con distintas combinaciones de variables. Los resultados son los siguientes:

Índice	FOLIO_I	Peso	Medida Cintura	Estatura
409	2022_08037175	72.40	152.1	152.1
926	2022_16019001	58.70	147.0	147.0
1096	2022_19018029	80.25	157.8	157.8
1165	2022_19031077	74.55	158.0	158.0
1215	2022_19046008	95.40	164.0	164.0
1775	2022_31101015	77.05	149.8	149.8

Cuadro 7: Valores atípicos eliminados en la primera iteración.

Índice	FOLIO_I	Peso	Medida Cintura	IMC
1511	2022_26042021	126.0	125.9	42.839

Cuadro 8: Valor atípico eliminado en la segunda iteración.

Índice	FOLIO_I	Tensión Arterial	Riesgo Hipertensión
23	2022_01002021	191	0
107	2022_03008060	179	0
177	2022_05024007	193	1
209	2022_05030087	200	1
642	2022_12001126	197	1
785	2022_14098021	191	1
1419	2022_24037025	177	0
1510	2022_26042010	200	1
1516	2022_26047007	200	1
1699	2022_30159025	188	0
1787	2022_32024009	200	1

Cuadro 9: Valores atípicos eliminados en la tercera iteración.

El uso de la distancia de Mahalanobis robusta permitió identificar múltiples valores atípicos en el conjunto de datos. La eliminación de estas observaciones asegura un análisis posterior más robusto y menos influenciado por valores extremos. Finalmente graficamos los mapas de calor respecto a cada data frame (hombre y mujeres).

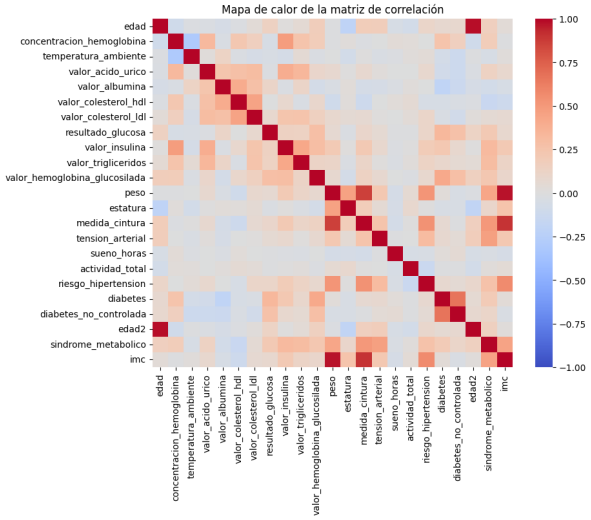


Figura 30: DataFrame de Hombres

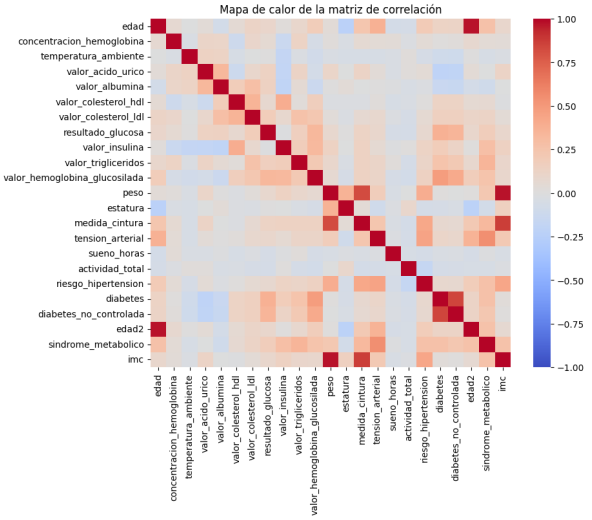


Figura 31: DataFrame de Mujeres

10 Modelo de regresión lineal

Para evaluar la relación entre las variables predictoras y el peso, se implementaron tres modelos de regresión lineal. El primer modelo incluye a hombres y mujeres mientras que los otros dos se aplicaron de manera separada, uno exclusivamente para hombres y otro para mujeres. Cabe recalcar que, para evitar la heterocedasticidad que se observó sobre los tres modelos iniciales se eligió la raíz cuadrada de la variable predictora (peso) como transformación, pues la variabilidad de los residuos aumentaba conforme lo hacía el peso, por ende, la transformación raíz ayudó a estabilizar la varianza de los errores hacía un único valor.

10.1 Herramientas

1. **Heterocedasticidad:** Se utilizó la prueba de Breusch-Pagan para verificar la constancia de la varianza de los residuos. Un p-valor menor a 0.05 indicaría heterocedasticidad.
2. **Multicolinealidad:** Se calcularon:
    - Índice de Condición: Valores superiores a 30 sugieren multicolinealidad severa mientras que un valor entre 10 y 30 sugiere una multicolinealidad moderada y aquellos por debajo de 10 indican que no hay problema respecto a la multicolinealid.
    - Factor de Inflación de la Varianza (VIF): Indicadores mayores a 10 sugieren problemas de redundancia entre predictores.
3. **Autocorrelación:** La prueba de Durbin-Watson y su respectivo valor en tablas evaluó la independencia de los residuos.

11 Resultados

11.1 Población de hombres y mujeres

Las variables predictoras son:

- Edad
- Estatura
- Medida de cintura
- Riesgo de hipertensión (binaria)
- Sexo (codificada como 0 = femenino, 1 = masculino)
- Diabetes no controlada (0 = negativo, 1 = positivo)

Los resultados obtenidos para el DataFrame General mostraron que:

- La esperanza de los errores es cero.
- $R^2 = 0{,}875$, lo que indica que el modelo explica el $87{,}5\,\%$ de la variabilidad en la variable dependiente.

- $CriteriodeAkaike = 1653$
- $CriteriodeBayesiano = 1695$
- Prueba de Breusch-Pagan respecto a la heterocedasticidad:
  - Estadístico LM: $35{,}4637$
  - P-valor LM: $0{,}0000$
  - Estadístico F: $5{,}9674$
  - P-valor F: $0{,}0000$

  Estos resultados indican que existe evidencia de heterocedasticidad $p - value < 0{,}05$ en al menos una prueba.
- $IC = 2{,}29968$
- $PruebadeDurbin - Watson = 1{,}895$ Este valor indica que hay evidencia de autocorrelación positiva en los residuos.

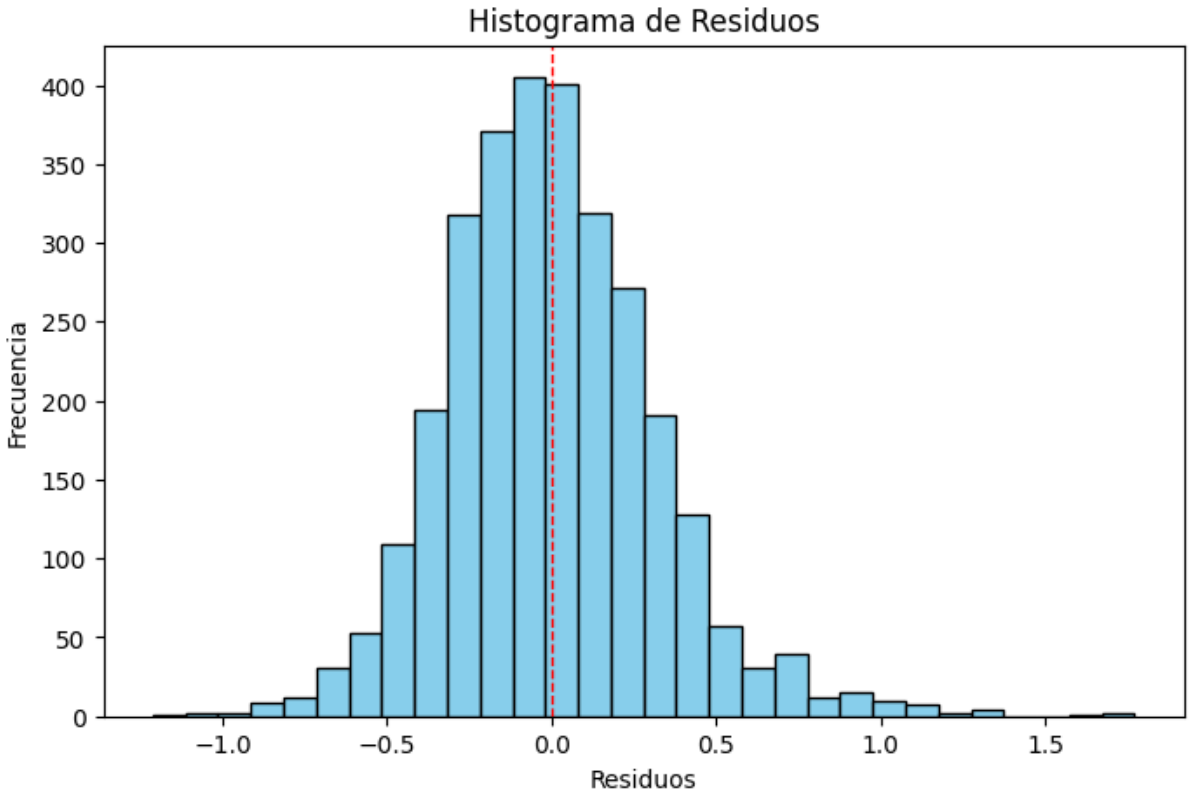


Figura 32: Histograma de residuos.

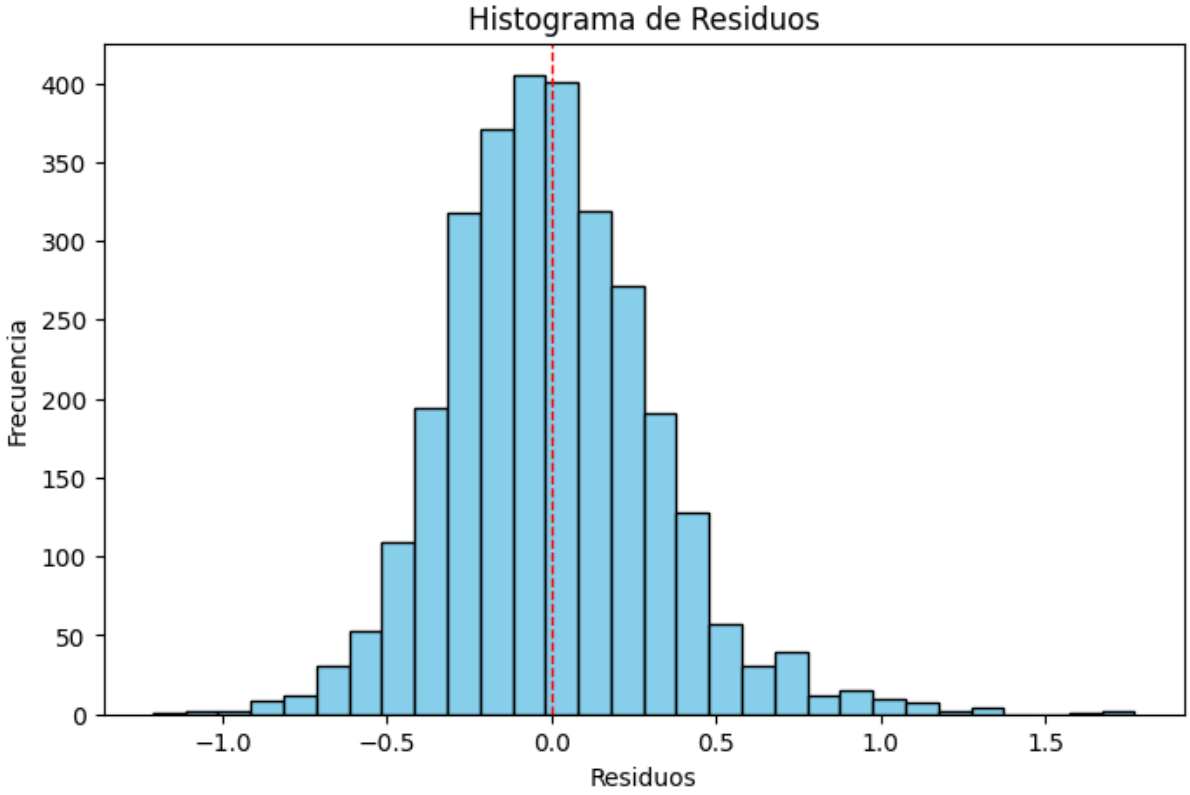


Figura 33: Histograma de residuos.

Variable	Coeficiente	Error Estándar	Estadístico t	P-valor
Constante	-2.2530	0.135	-16.639	0.000
Edad	-0.0075	0.001	-13.416	0.000
Estatura	0.0368	0.001	43.731	0.000
Medida cintura	0.0537	0.001	104.641	0.000
Riesgo hipertensión	0.1832	0.016	11.206	0.000
Sexo	-0.0936	0.016	-5.997	0.000
Diabetes no controlada	-0.1118	0.035	-3.201	0.001

Cuadro 10: Resultados de la regresión lineal múltiple.

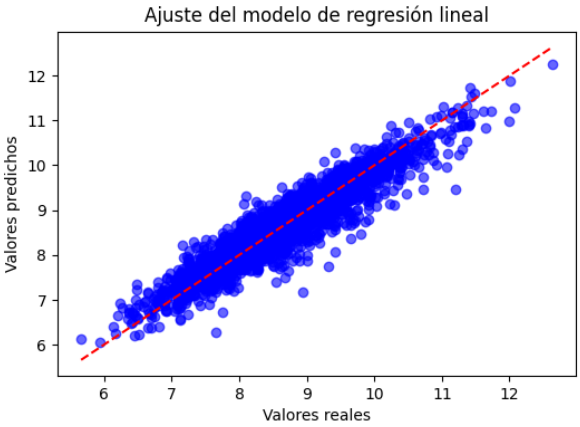


Figura 34: Ajuste del modelo

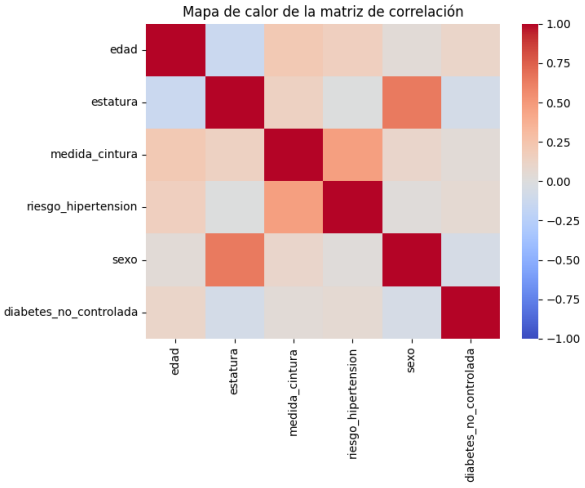


Figura 35: Correlación

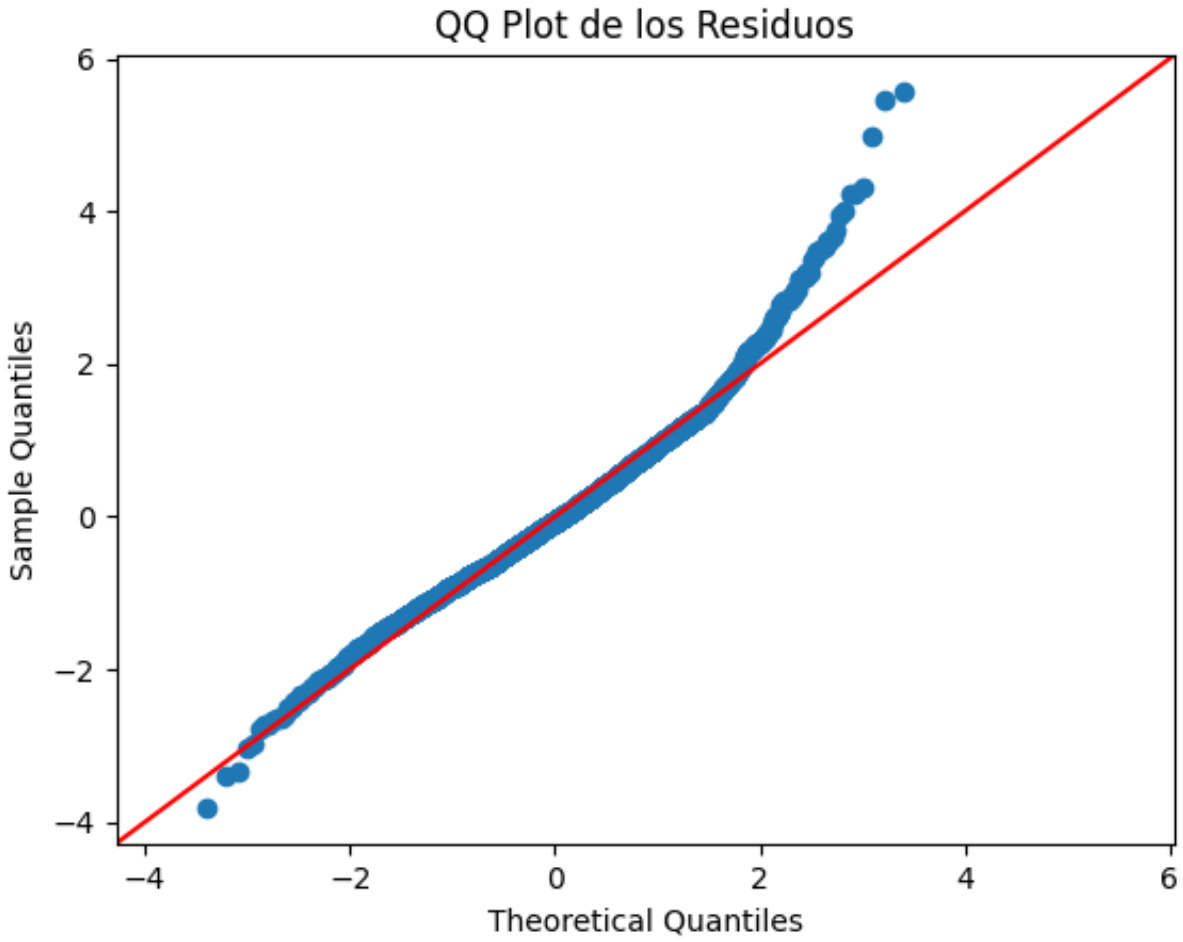


Figura 36: Grafica Q-Q

Dificultades

Heterocedasticidad. La prueba de Breusch-Pagan indica heterocedasticidad ($p < 0{,}05$). Esto sugiere que la varianza de los errores no es constante, lo que podría afectar la eficiencia de los estimadores.

Autocorrelación. $PruebadeDurbin - Watson = 1{,}895$ indica que hay evidencia de autocorrelación positiva en los residuos.

Variable	VIF
Constante	541.9
Edad	1.1
Estatura	1.8
Medida cintura	1.4
Riesgo hipertensión	1.3
Sexo	1.7
Diabetes no controlada	1.0

Cuadro 11: Factores de Inflación de la Varianza (VIF).

11.2 DataFrame Hombres

Las variables predictoras son:

- Edad
- Estatura
- Medida de cintura
- Riesgo de hipertensión
- R-cuadrado: El coeficiente de determinación ajustado es $R^2 = 0{,}898$, indicando que el $89{,}8\,\%$ de la variabilidad en el peso se explica por las variables predictoras seleccionadas.
- Índice de Condición: $1{,}9487$ lo que indica que no hay problemas significativos de multicolinealidad.
- Prueba de Breusch-Godfrey para Autocorrelación:
  - Estadístico LM: $6{,}6808$, $p = 0{,}5714$
  - Estadístico F: $0{,}8306$, $p = 0{,}5756$

  No se detecta autocorrelación en los residuos ($p > 0{,}05$).

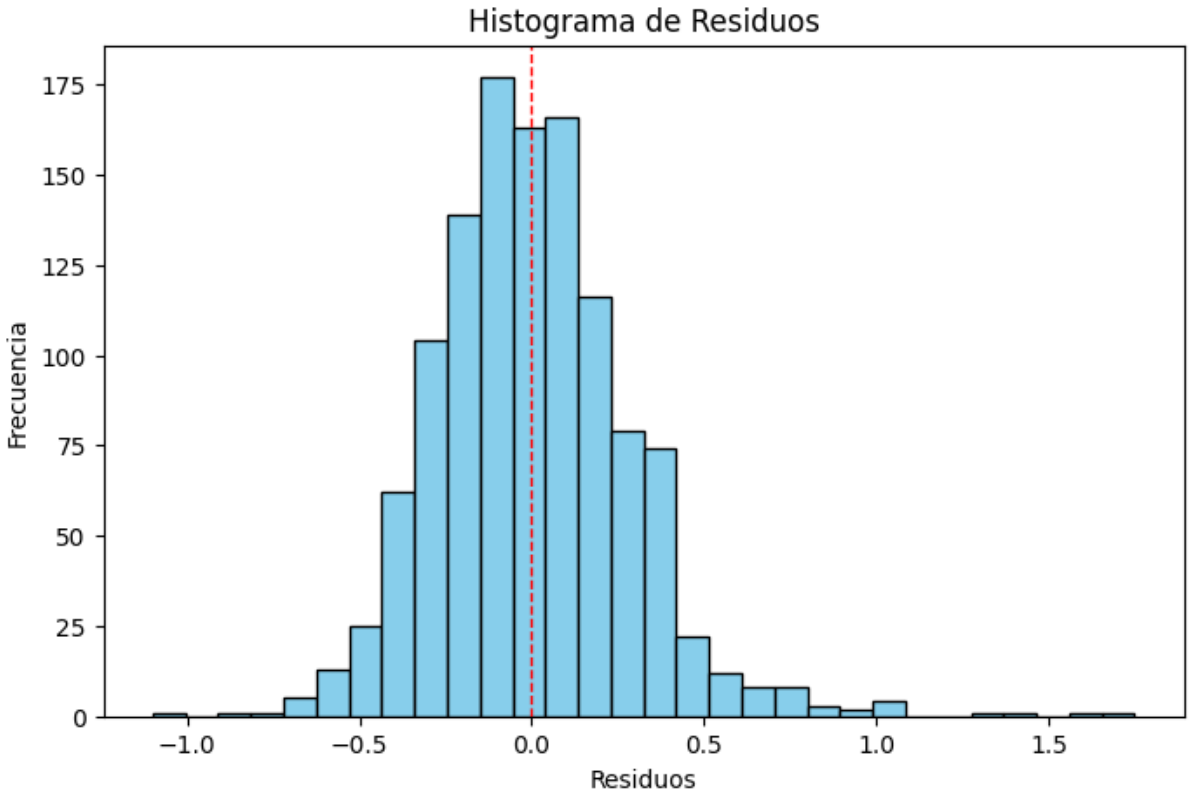


Figura 37: Histograma de residuos.

- La esperanza de los errores es cero.
- Criterios de información:
  - AIC: 384.2
  - BIC: 409.6

Variable	Coeficiente	Error Estándar	Estadístico t	P-valor
Constante	-2.0008	0.191	-10.459	0.000
Edad	-0.0081	0.001	-10.364	0.000
Estatura	0.0345	0.001	30.562	0.000
Medida cintura	0.0541	0.001	70.244	0.000
Riesgo hipertensión	0.2123	0.024	8.718	0.000

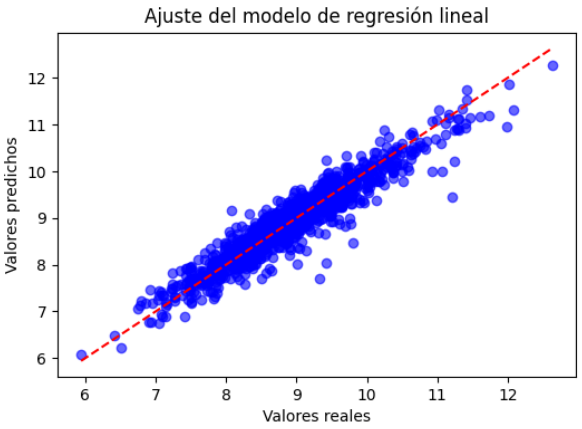
Cuadro 12: Resultados de la regresión lineal múltiple para hombres.

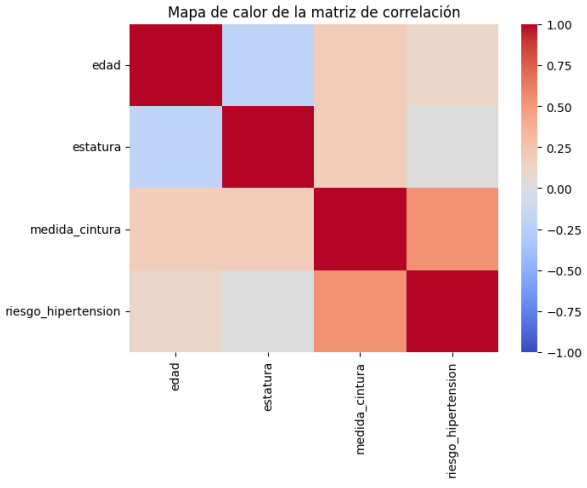
Autocorrelación. La prueba de Breusch-Godfrey no detecta evidencia de autocorrelación en los residuos ($p > 0{,}05$).

Multicolinealidad. Los factores de inflación de la varianza (VIF) se reportan en la Tabla 13. Exceptuando la constante (VIF > 10), todas las variables predictoras presentan valores aceptables (VIF < 2).

Variable	VIF
Constante	540.2
Edad	1.1
Estatura	1.1
Medida cintura	1.5
Riesgo hipertensión	1.4

Cuadro 13: Factores de Inflación de la Varianza (VIF) para hombres.





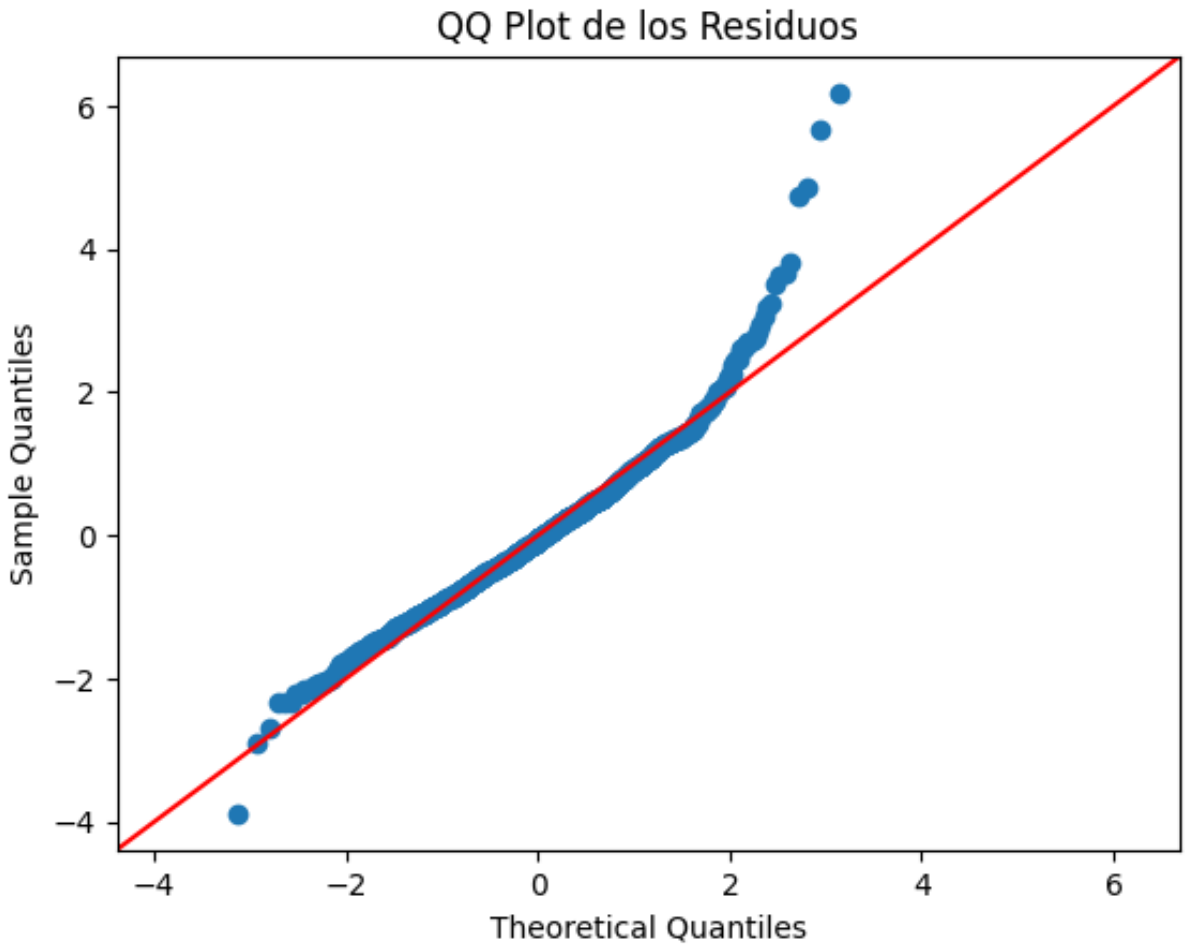


Figura 38: Grafica Hombres Q-Q

11.3 DataFrame de Mujeres

Para este modelo las variables predictoras son:

- Edad
- Estatura
- Medida de cintura
- Riesgo de hipertensión
- Diabetes no controlada
- Norte
- Sur

De la misma manera, utilizamos el mismo modelo de regresion MRL3 para el dataframe de mujeres. Los resultados fueron los siguientes:

- R-cuadrado: $R^2 = 0{,}843$, lo que indica que el $84{,}3\,\%$ de la variabilidad en el peso se explica mediante las variables del modelo.
- Prueba de heterocedasticidad (Breusch-Pagan):
  - Estadístico LM: $9{,}3286$, $p = 0{,}1559$.
  - Estadístico F: $1{,}5568$, $p = 0{,}1560$.

  No se detecta evidencia de heterocedasticidad ($p > 0{,}05$).
- Índice de Condición: $2{,}0356$, indicando un nivel aceptable de estabilidad numérica.
- Prueba de autocorrelación (Breusch-Godfrey):
  - Estadístico LM: $20{,}0492$, $p = 0{,}0102$.
  - Estadístico F: $2{,}5132$, $p = 0{,}0102$.

  Existe evidencia de autocorrelación en los residuos ($p < 0{,}05$).
- La esperanza de los errores es: $-0{,}00$.
- Criterios de información:
  - AIC: 1212.
  - BIC: 1256.

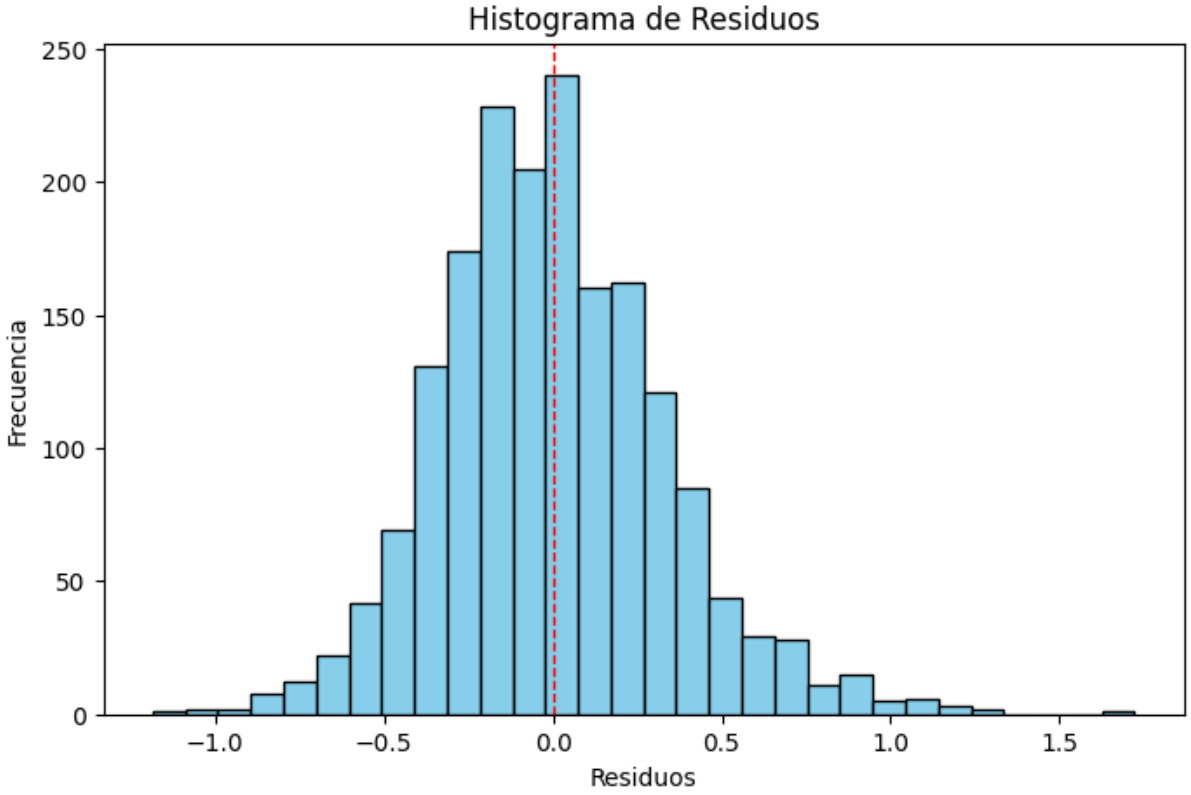


Figura 39: Histograma de residuos.

Variable	Coeficiente	Error Estándar	Estadístico t	P-valor
Constante	-2.5214	0.196	-12.873	0.000
Edad	-0.0075	0.001	-8.490	0.000
Estatura	0.0388	0.001	32.088	0.000
Medida cintura	0.0534	0.001	77.510	0.000
Riesgo hipertensión	0.1634	0.022	7.491	0.000
Diabetes no controlada	-0.1169	0.043	-2.751	0.006
Edad^2	0.0219	0.019	1.148	0.251

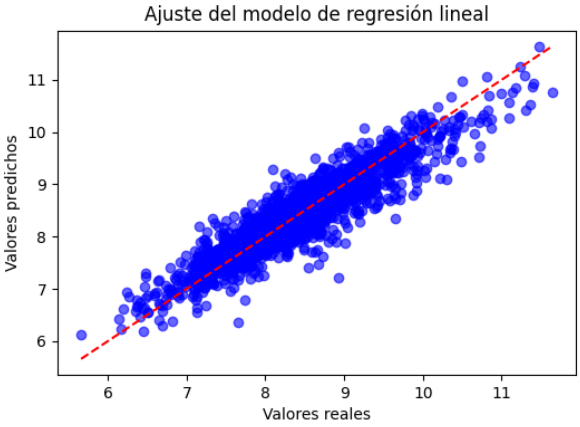
Cuadro 14: Resultados del modelo de regresión lineal múltiple para mujeres.

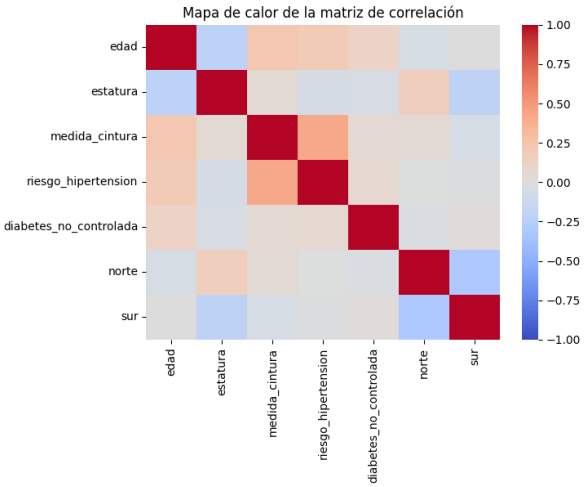
Dificultades

Autocorrelación. La prueba de Durbin-Watson indica evidencia de autocorrelación en los residuos ($p < 0{,}05$), lo que podría afectar la eficiencia de las estimaciones.

Variable	VIF
Constante	604.6
Edad	1.5
Estatura	1.1
Medida cintura	1.3
Riesgo hipertensión	1.2
Diabetes no controlada	1.0
Edad2	1.4

Cuadro 15: Factores de Inflación de la Varianza (VIF) para mujeres.





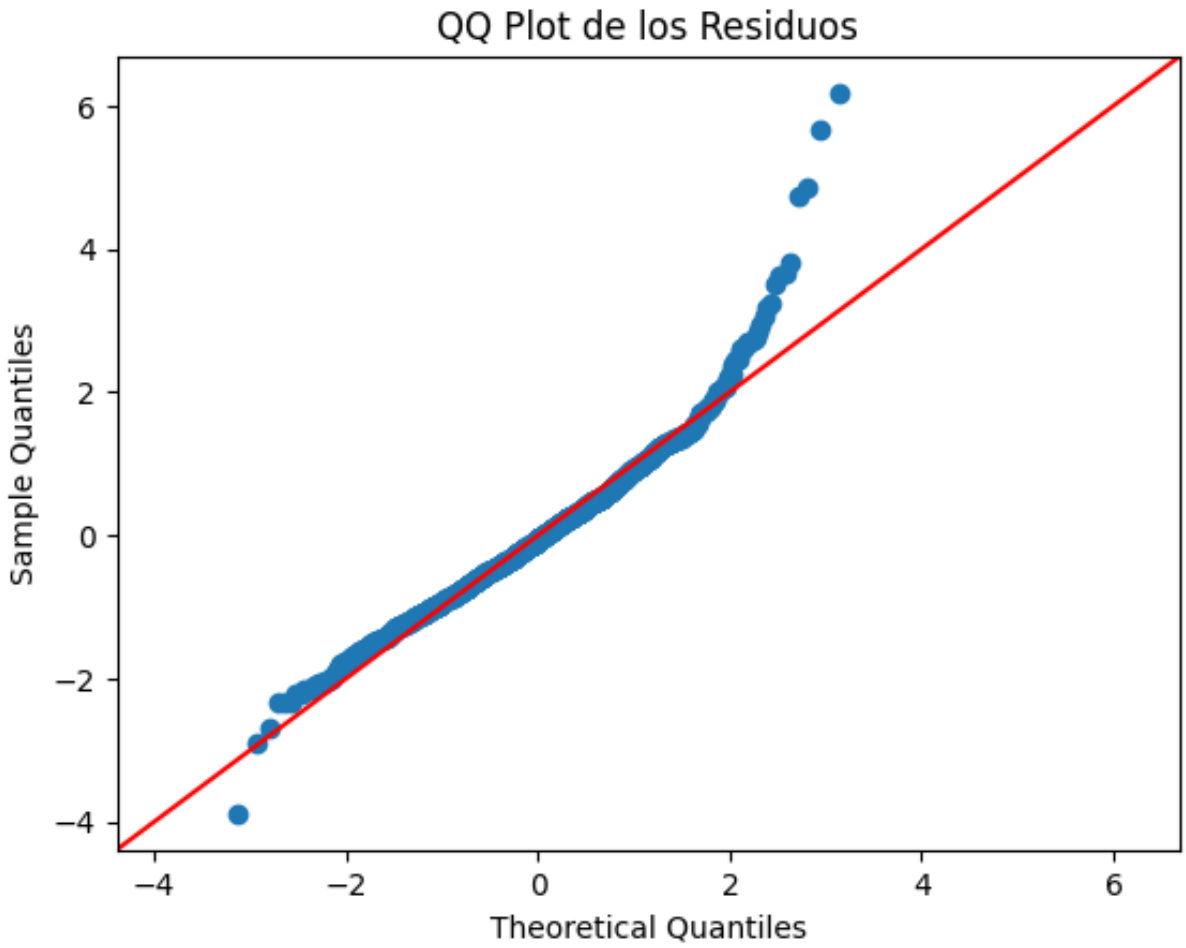


Figura 40: Grafica Mujeres Q-Q

12 Conclusiones

La conclusión general de nuesto modelo para hombres aporta información contundente sobre los factores a los que se le puede atribuir su peso son la edad, medida de cintura y el riesgo de hipertensión. Por la naturaleza del modelo, es desafiante asignarle una interpretación exacta en la forma que contribuyen al peso del varón, pues el modelo resulta explicativo para la raíz cuadrada del peso y no solamente el peso. Sin embargo, podemos afirmar que el peso disminuye conforme se envejece y, naturalmente el peso aumenta de acuerdo a la altura y respecto a la medida de la cintura (indicador de grasa abdominal). Aproximadamente, el hecho de tener riesgo de hipertensión es un factor de aumento de peso casi de cinco kilos extra. Respecto al modelo general, dado el problema de heterocedasticidad y autocorrelación se debe ser cauteloso con la interpretación del modelo, pues las pruebas de hipótesis son menos confiables así como también puede ser que se sobreestime el coeficiente R^2 por lo que nos limitaremos de realizar conclusiones de este modelo. El único porblema que tuvo el modelo para mujeres es el de autocorrelación el cual, a pesar de haber intentado atenderlo mediante la división en regiones del país (norte,centro y sur) puede conllevar sobreestimar el coeficiente

R^2 y hacer de los estadísticos medidas menos confiables. Por otra parte, este modelo nos dice que la diabetes no controlada es un factor de pérdida de peso mientras que el riesgo de hipertensión un factor de aumento. Por la construcción, aquellas mujeres pertenecientes a la zona norte del país no reciben ninguna penalización sobre el peso, siendo que las mujeres de la región centro y sur sí lo hacen de tal manera que las mujeres de la región sur son las que reciben un mayor aumento de peso. Por último, vale la pena recalcar que el origen de esta autocorrelación debe ser estuddiada más a fondo, ya sea mediante la aplicación de otro modelo o bien, indagando sobre los factores que pueden llevar a esta autocorrelación y que no pueden ser modelados con nuestros datos, pues carecemos de mediciones que nor permiten identificar enfermedades tiroideas, como el hipotiroidismo e hipertiroidismo, y trastornos como la anemia o la bulimia.

Norte

Sur

Estatura

Medida de cintura

Riesgo de hipertensión

Referencias

[Benzadon 2014] Benzadón, Mariano, Forti, Luján, & Sinay, Isaac. (2014). Update on the diagnosis of diabetes. Medicina (Buenos Aires), 74(1), 64-68. Recuperado en 22 de noviembre de 2024, de https://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S0025-76802014000100016&lng=es&tlng=en.

[NIDDK] Pruebas y diagnóstico de la diabetes - NIDDK. (n.d.). National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/informacion-de-la-salud/diabetes/informacion-general/pruebas-diagnostico

[OMS] World Health Organization: WHO. (2024, November 14). Diabetes. https://www.who.int/es/news-room/fact-sheets/detail/diabetes

[ADA] Diagnosis and classification of diabetes mellitus. (2010). Diabetes Care, 33(Supplement_1), S62–S69. https://doi.org/10.2337/dc10-s062

[IBC] Hemoglobina Glicosilada: importancia para el diagnóstico y control de la Diabetes Mellitus. (n.d.). Instituto De Bioquímica Clínica.https://www.ibcrosario.com.ar/articulos/diabetes-2020-pacientes.html

[NLM] National Library of Medicine. (n.d.). Síndrome metabólico. https://medlineplus.gov/spanish/metabolicsyndrome.html

[SPM] Tinajero-Delgado, J., Martínez-Ezquerro, J. D., Moreno-Tamayo, K., Curcio-Borrero, C. L., Arias-Merino, E. D., Sánchez-García, S., Espinel-Bermúdez, M. C., & Valencia-Rico, C. L. (2023). Factores que afectan el estado nutricional en personas mayores mexicanas: Enasem, 2018. Salud Pública De México, 65(5, sept-oct), 493–503. https://doi.org/10.21149/14753

[Marcuño (1987)] Chomon Barredo, B. (1987). Parámetros bioquímicos en ancianos. Documents - Universidade De Santiago De Compostela. https://investigacion.usc.es/documentos/5d1df65f29995204f766a055?lang=en

[Predição do peso... (2008)] Reis, G. L., Albuquerque, F. H. M. a. R., Valente, B. D., Martins, G. A., Teodoro, R. L., Ferreira, M. B. D., Monteiro, J. B. N., De Almeida E Silva, M., & Madalena, F. E. (2008). Predição do peso vivo a partir de medidas corporais em animais mestiços Holandês/Gir. Ciência Rural, 38(3), 778–783. https://doi.org/10.1590/s0103-84782008000300029

[Moreno y Sánchez (2020)] Moreno-Cruz, F., & Sánchez-Herrera, J. (2020). Redes neuronales artificiales y modelo de regresión lineal múltiple: nuevas alternativas para mejorar la estimación de gasto energético total en jóvenes universitarios. Revista Mexicana De Endocrinología Metabolismo Y Nutrición. https://doi.org/10.24875/rme.19001887

[NIH (2022)] Causas y factores de riesgo | NHLBI, NIH. (2022, March 24). NHLBI, NIH. https://www.nhlbi.nih.gov/es/salud/sobrepeso-y-obesidad/causas

[NLM Trastornos metabólicos] National Library of Medicine. (n.d.). Trastornos metabólicos. https://medlineplus.gov/spanish/metabolicdisorders.html

[Mayo Clinic] ¿Puedes acelerar el metabolismo? (n.d.). Mayo Clinic. https://www.mayoclinic.org/es/healthy-lifestyle/weight-loss/in-depth/metabolism/art-20046508

[MedlinePlus4] Glucosa en la sangre. (n.d.). https://medlineplus.gov/spanish/bloodglucose.html

[Boccio (2003)] Boccio, Jose, Salgueiro, Jimena, Lysionek, Alexis, Zubillaga, Marcela, Goldman, Cinthia, Weill, Ricardo, Caro, Ricardo. (2003). Metabolismo del hierro: conceptos actuales sobre un micronutriente esencial. Archivos Latinoamericanos de Nutrición, 53(2), 119-132. Recuperado en 26 de noviembre de 2024, de http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S0004-06222003000200002&lng=es&tlng=es.

[Chitarroni (2002)] Chitarroni, H. (2002). El análisis de correlación y regresión lineal entre variables cuantitativas.

[Cortés (2015)] Cortés, J., Bielsa, N., Cobo, E., Muñoz, P., González, A. (2015). Regresión lineal simple. Barcelona: Universitat Politècnica de Catalunya, 3–35.

[Sabogal (2021)] Sabogal-Céspedes, G. (2021). Sistema E-Health de adquisición y almacenamiento de variables fisiológicas, obtenidas de dispositivos comerciales, necesarias para predecir el nivel de insulina.