

Universidad Nacional De La Matanza

Especialización Ciencia de Datos

Fundamentos de Captura de Datos

Profesor: Silvia N. Pérez

Trabajo Práctico

Regresión Lineal Múltiple

Estimación del Porcentaje de Grasa Corporal a
partir de Variables Antropométricas en el
Conjunto de Datos 'obesidad25'

Integrantes Grupo 10:

Fica Millán, Yesica	DNI 27.624.956
Miranda Charca, Florencia	DNI 41.398.768
Petraroia, Franco	DNI 27.161.862

Contenido

1. Introducción	5
1.1. Objetivo del trabajo.....	5
1.2. Importancia de estudiar el porcentaje de grasa corporal	5
1.3 Descripción general del conjunto de datos.....	6
2. Análisis Exploratorio de Datos (EDA).....	6
2.1. Carga y examinación los datos	6
2.2. Verificación de NAs y duplicados	8
2.3. Gráficos univariados	9
2.4. Outliers	10
2.5. Análisis bivariado y correlaciones	15
2.6. Distribución del porcentaje de grasa corporal	20
3. Modelo 1: Regresión Simple.....	21
3.1. Selección de la variable predictora	21
3.2. Ajuste del modelo lineal simple	21
3.3. Interpretación del modelo	21
3.4. Visualización del ajuste	23
4. Modelo 2: Regresión Múltiple con todas las variables numéricas.....	23
4.1. Ajuste del modelo múltiple	23
4.2. Evaluación global del modelo.....	24
4.3. Ecuación del modelo ajustado	24
4.4. Análisis de significancia individual de variables	25
4.5. Diagnóstico de multicolinealidad	25
5. Comparación y Selección de Modelos de Regresión.....	26

5.1. Modelo con selección automática de variables	26
5.2. Modelo con selección manual.....	26
5.3. Modelo completo (todas las variables)	27
5.4. Comparación de modelos y selección final	27
6. Evaluación del Modelo Seleccionado	28
6.1. Verificación de supuestos del modelo lineal.....	28
6.2. Detección de observaciones influyentes.....	31
6.3. Predicción para nuevos casos.....	33
7. Conclusiones.....	35
Análisis Exploratorio de Datos (EDA):.....	35
Modelo 1: Regresión Simple:	35
Modelo 2: Regresión Múltiple con todas las variables numéricas:.....	35
Comparación y Selección de Modelos de Regresión:	36
Evaluación del Modelo Seleccionado:.....	36
8. Anexos	36

Índice de Gráficos

Ilustración 1 - Distribución de variables numéricas	10
Ilustración 2 - Boxplots de variables numéricas	10
Ilustración 3 - Porcentaje GRC vs Abdomen con outliers.....	13
Ilustración 4 - Comparación del modelo lineal con y sin outliers en abdomen	14
Ilustración 5 - Matriz de correlación entre variables numéricas.....	16
Ilustración 6 - Relación entre grc y las variables más correlacionadas	18
Ilustración 7 - Relaciones entre grc y variables predictoras.....	19

Ilustración 8 - Distribución y densidad del grc	20
Ilustración 9 - Ajuste del modelo lineal simple	23
Ilustración 10 - Gráficos de diagnóstico del modelo lineal	29
Ilustración 11 - Matriz de correlación modelo lineal	30
Ilustración 12 - Distancia de Cook	31
Ilustración 13 - Distribución del leverage.....	32

1. Introducción

Este trabajo tiene por finalidad tratar de estimar el porcentaje de grasa corporal a partir de otras variables clínicas y antropométricas que están presentes en el dataset `obesidad25`. Se busca entender qué factores influyen realmente en ese porcentaje y cómo se puede predecir con cierta precisión.

Se utilizarán diferentes técnicas de regresión, buscando construir modelos que no solo ajusten bien los datos, sino que además puedan ser herramientas útiles en contextos reales, como el seguimiento clínico de pacientes o la toma de decisiones en la salud pública.

1.1. Objetivo del trabajo

El objetivo principal es construir un modelo que permita predecir el porcentaje de grasa corporal (grc) usando las demás variables disponibles en el conjunto de datos. Para tal fin se realizará:

- Análisis exploratorio, considerando a cada variable por separado (univariado) y cómo se relacionan entre sí (bivariado).
- Ajuste del modelo de regresión lineal simple con la variable que tenga la mayor correlación con el porcentaje de grasa corporal.
- Desarrollar un modelo múltiple que incluya todas las variables numéricas, comparando diferentes formas de seleccionar las más relevantes, tanto de forma automática y manual, con criterio humano.
- Finalmente, se pondrá a prueba el modelo seleccionado. Se observará si se cumple con los supuestos, se identificarán los datos influyentes, para concluir con el cálculo de intervalos de confianza y predicción de nuevos casos.

Este enfoque permitirá arribar a un modelo que funcione bien y, además, evaluar la significación y robustez del modelo ajustado.

1.2. Importancia de estudiar el porcentaje de grasa corporal

El porcentaje de grasa corporal no solo es un número estético o deportivo, es un indicador significativo del estado de salud general de una persona. Cuando ese valor está fuera de rango, activa varias alarmas.

- **Para prevenir y diagnosticar:** un porcentaje graso elevado suele ir de la mano con riesgos importantes para la salud, como enfermedades cardiovasculares, diabetes tipo 2 y otros problemas metabólicos. Tener una estimación confiable puede marcar la diferencia entre un diagnóstico a tiempo y uno tardío.
- **Para hacer seguimiento clínico:** en el caso de que una persona está haciendo un cambio importante en su estilo de vida. Poder monitorear cómo varía su grasa corporal permite evaluar si el tratamiento es efectivo o no.

- **Para investigar en salud:** en estudios más amplios, el porcentaje de grasa corporal se utiliza como un indicador clave para entender patrones poblacionales y vínculos entre hábitos, genética y riesgo.

En resumen, conocer este valor proporciona una herramienta necesaria para cuidar y mejorar la salud, tanto a nivel individual como colectivo.

1.3 Descripción general del conjunto de datos

El conjunto de datos a utilizar, llamado `obesidad25`, reúne información real de pacientes evaluados en estudios sobre obesidad.

Las variables que contiene son medidas antropométricas: el peso, la altura, el perímetro abdominal y, principalmente, el porcentaje de grasa corporal (`grc`).

Este conjunto de variables permitirá iniciar por una descripción general, observar cómo se distribuyen los datos y si hay valores anómalos, para comprender la distribución y comportamiento de cada variable. Luego, se iniciará un análisis más profundo, con un ajuste de modelos de regresión que explore la relación entre ellas.

Cabe señalar que el conjunto de datos no explicita las unidades de medida asociadas a las variables numéricas. En función de los rangos de valores observados y el contexto clínico del estudio, se parte del siguiente **supuesto**:

- Las medidas de `peso` están expresadas en kilogramos (kg).
- Las variables: `altura`, `cuello`, `pecho`, `abdomen`, `cadera`, `muslo` y `rodilla` están expresadas en centímetros (cm).
- La variable `grc` corresponde al porcentaje de grasa corporal (%).

Esta suposición se mantendrá a lo largo del trabajo, permitiendo interpretar los resultados en un marco clínico coherente.

2. Análisis Exploratorio de Datos (EDA)

2.1. Carga y examinación los datos

Para comenzar con el análisis, se procedió a cargar el conjunto de datos `obesidad25.xlsx` el cual contiene información antropométrica de 260 individuos, estructurada en 9 variables numéricas.

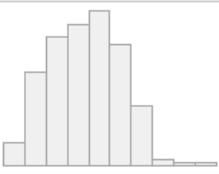
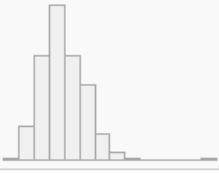
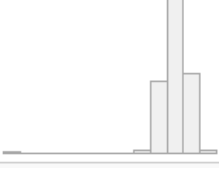
A continuación, se presentan los principales hallazgos de la exploración preliminar:

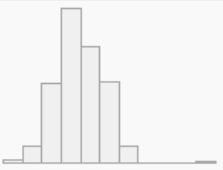
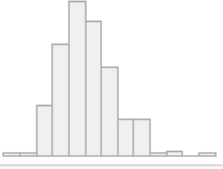
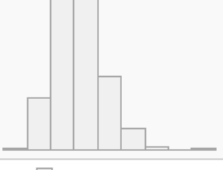
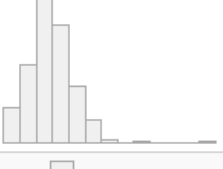
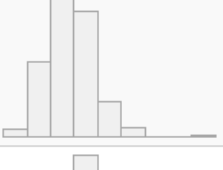
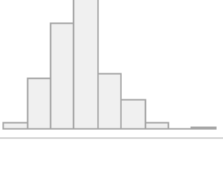
- **Vista preliminar de los datos:** al observar las primeras filas del conjunto, se identificaron variables como el porcentaje de grasa corporal (`grc`), el peso, la altura, y distintas medidas

corporales (cuello, pecho, abdomen, cadera, muslo y rodilla). La siguiente tabla muestra los primeros seis registros del dataset:

Primeras 6 observaciones								
grc	peso	altura	cuello	pecho	abdomen	cadera	muslo	rodilla
12.3	154.25	169.38	36.2	93.1	85.2	94.5	59.0	37.3
6.1	173.25	180.62	38.5	93.6	83.0	98.7	58.7	37.3
25.3	154.00	165.62	34.0	95.8	87.9	99.2	59.6	38.9
10.4	184.75	180.62	37.4	101.8	86.4	101.2	60.1	37.3
6.3	155.25	173.12	37.5	89.3	78.4	96.1	56.0	37.4
20.9	210.25	186.88	39.0	104.5	94.4	107.8	66.0	42.0

- **Dimensiones del dataset:** el conjunto cuenta con 260 observaciones (filas) y 9 variables (columnas).
- **Nombres de las variables:** las variables presentes son:
 - grc: porcentaje de grasa corporal (variable objetivo)
 - peso, altura: medidas generales del cuerpo
 - cuello, pecho, abdomen, cadera, muslo, rodilla: medidas específicas por zona corporal.
- **Resumen estadístico:** a continuación, se presenta un resumen de las estadísticas descriptivas para cada variable:

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	grc [numeric]	Mean (sd) : 19.1 (8.3) min ≤ med ≤ max: 0 ≤ 19.2 ≤ 47.5 IQR (CV) : 12.8 (0.4)	176 distinct values		260 (100.0%)	0 (0.0%)
2	peso [numeric]	Mean (sd) : 178.7 (29.2) min ≤ med ≤ max: 118.5 ≤ 176.5 ≤ 363.1 IQR (CV) : 38.6 (0.2)	197 distinct values		260 (100.0%)	0 (0.0%)
3	altura [numeric]	Mean (sd) : 175.2 (9.1) min ≤ med ≤ max: 73.8 ≤ 175 ≤ 194.4 IQR (CV) : 10 (0.1)	48 distinct values		260 (100.0%)	0 (0.0%)

4	cuello [numeric]	Mean (sd) : 38 (2.4) min ≤ med ≤ max: 31.1 ≤ 38 ≤ 51.2 IQR (CV) : 3 (0.1)	90 distinct values		259 (99.6%)	1 (0.4%)
5	pecho [numeric]	Mean (sd) : 100.9 (8.6) min ≤ med ≤ max: 79.3 ≤ 99.7 ≤ 136.2 IQR (CV) : 11.4 (0.1)	174 distinct values		260 (100.0%)	0 (0.0%)
6	abdomen [numeric]	Mean (sd) : 92.5 (10.7) min ≤ med ≤ max: 69.4 ≤ 91 ≤ 148.1 IQR (CV) : 14.5 (0.1)	185 distinct values		260 (100.0%)	0 (0.0%)
7	cadera [numeric]	Mean (sd) : 99.8 (7.2) min ≤ med ≤ max: 85 ≤ 99.3 ≤ 147.7 IQR (CV) : 7.9 (0.1)	152 distinct values		260 (100.0%)	0 (0.0%)
8	muslo [numeric]	Mean (sd) : 59.4 (5.2) min ≤ med ≤ max: 47.2 ≤ 59 ≤ 87.3 IQR (CV) : 6.3 (0.1)	139 distinct values		260 (100.0%)	0 (0.0%)
9	rodilla [numeric]	Mean (sd) : 38.6 (2.4) min ≤ med ≤ max: 33 ≤ 38.5 ≤ 49.1 IQR (CV) : 3 (0.1)	90 distinct values		259 (99.6%)	1 (0.4%)

Se observa que las variables `cuello` y `rodilla` presentan un valor faltante cada una (NA).

- **Estructura de los datos:** mediante la función `str()`, se confirmó que todas las variables fueron importadas como numéricas y que el objeto `datos` es un `data.frame` con formato `tibble`.

Esta etapa de exploración inicial permitió comprender la estructura del conjunto de datos y detectar aspectos relevantes como valores atípicos o faltantes, los cuales serán tratados en el análisis exploratorio siguiente.

2.2. Verificación de NAs y duplicados

Se realizó una revisión del conjunto de datos en busca de valores faltantes (NA) y observaciones duplicadas, con el fin de asegurar la calidad y consistencia de los datos utilizados en los análisis posteriores.

- **Valores faltantes:** se detectó la presencia de datos faltantes en dos variables: `cuello` y `rodilla`, cada una con un único valor ausente. En total, el conjunto contenía 2 valores faltantes distribuidos en diferentes observaciones. Para evitar interferencias en los modelos de regresión, se optó por eliminar las filas con valores faltantes, trabajando a partir de allí con un subconjunto de datos completos.
- **Datos duplicados:** también se evaluó la existencia de observaciones duplicadas. El resultado fue cero casos duplicados, lo que indica que no hay registros repetidos dentro del dataset.

Con esta limpieza inicial, se conformó un nuevo objeto con **258 observaciones completas**, sobre el cual se desarrollarán los análisis siguientes.

2.3. Gráficos univariados

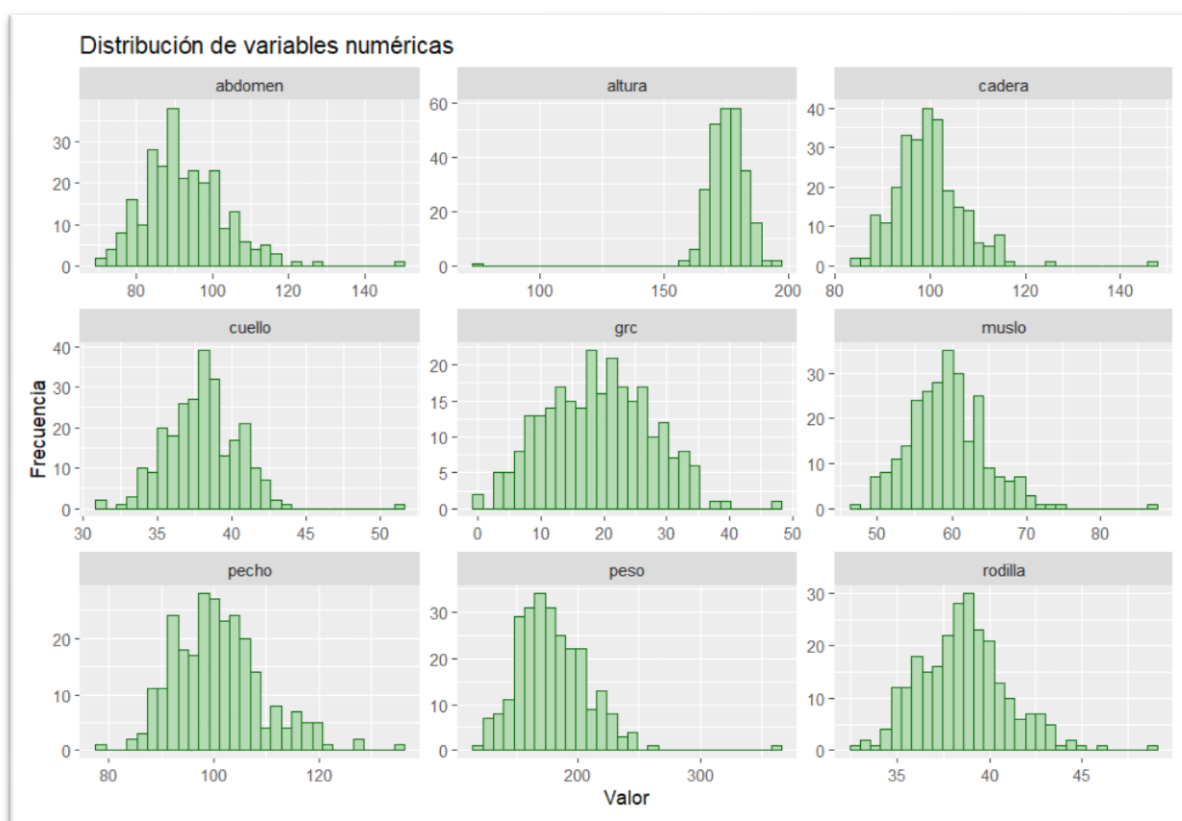
Para explorar la distribución de las variables numéricas del conjunto de datos, se construyeron histogramas individuales agrupados en una única visualización.

Estos gráficos permiten observar el comportamiento general de cada variable, identificar asimetrías, concentraciones de valores, y posibles valores atípicos.

A continuación, se presentan los principales hallazgos:

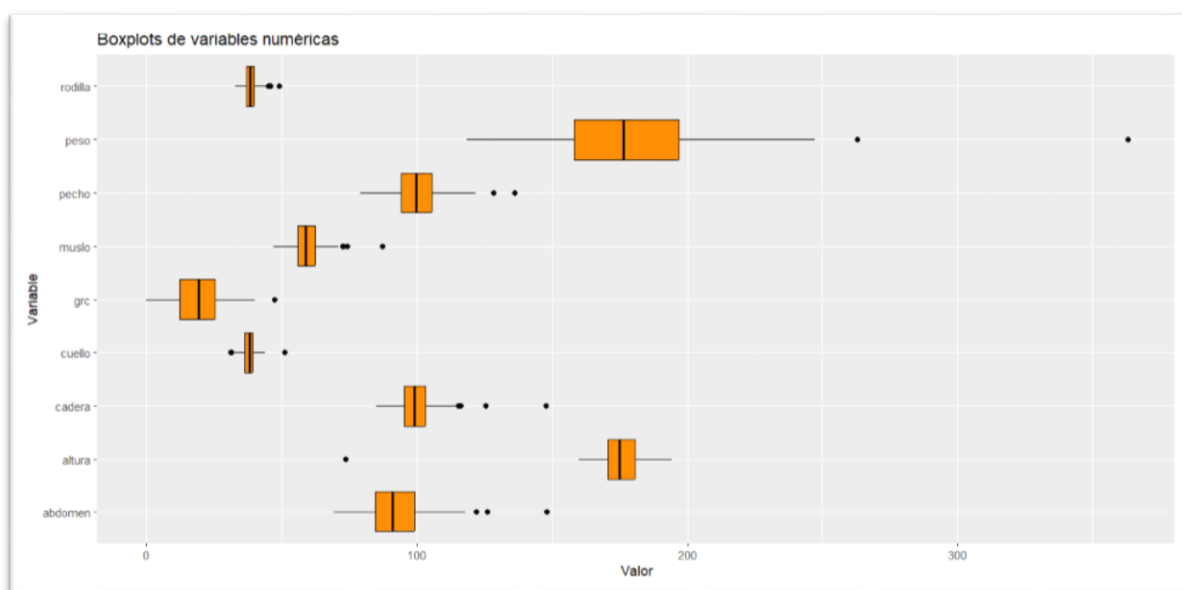
- **grc (porcentaje de grasa corporal):** muestra una distribución relativamente simétrica, con una mayor concentración entre el 10% y el 30%. No obstante, se observan algunos valores extremos hacia ambos lados.
- **abdomen:** presenta una leve asimetría hacia la derecha, con una acumulación principal entre 80 y 100 cm, pero también con valores extremos por encima de 120 cm.
- **peso:** la distribución es asimétrica hacia la derecha, con una mayoría de individuos entre los 150 y 220 kg, aunque existen casos con valores muy elevados (más de 300 kg) que podrían ser considerados atípicos.
- **altura:** se presenta una distribución levemente sesgada a la izquierda, con un grupo de valores inusualmente bajos (por debajo de los 100 cm), lo cual puede indicar errores de medición o registro.
- **cuello, pecho, cadera, muslo, rodilla:** todas estas variables presentan distribuciones aproximadamente normales, con formas de campana relativamente simétricas, aunque en algunos casos con colas alargadas o leves asimetrías.

En general, la mayoría de las variables muestra un comportamiento razonable desde el punto de vista clínico. Sin embargo, algunas distribuciones revelan la existencia de posibles outliers, como en los casos de `altura`, `peso` y `cadera`, que serán abordados en el próximo apartado.

*Ilustración 1 - Distribución de variables numéricas*

2.4. Outliers

Se construyeron diagramas de caja (boxplots) para cada una de las variables numéricas, con el fin de identificar posibles valores atípicos y examinar la dispersión de los datos.

*Ilustración 2 - Boxplots de variables numéricas*

A partir del gráfico se observan lo siguiente:

- **Presencia de outliers:** varias variables presentan puntos fuera de los bigotes del boxplot, lo que indica la presencia de valores atípicos. Estos outliers se detectan en *peso*, *altura*, *abdomen*, *cadera*, *cuello*, *pecho*, *muslo* y *rodilla*.
- **Peso:** es una de las variables con mayor dispersión y con valores atípicos notoriamente elevados, algunos por encima de los 300 kg, lo cual podría representar casos extremos clínicos o errores de registro.
- **Altura:** se identifica un valor atípicamente bajo, por debajo de lo esperable para una población adulta. Esto sugiere que podría tratarse de errores de carga o casos excepcionales.
- **Abdomen y cadera:** también presentan una cantidad moderada de outliers hacia el extremo superior, que podrían corresponder a pacientes con obesidad abdominal más pronunciada.
- **Rodilla y cuello:** muestran pocos outliers, pero con rangos estrechos, lo que indica menor dispersión.
- **Porcentaje de grasa corporal (grc) y muslo:** presentan distribuciones más simétricas y con menor cantidad de valores atípicos en comparación con otras variables.

Este análisis visual resulta fundamental para decidir si conviene conservar, corregir o eliminar ciertos valores extremos antes de ajustar los modelos.

A continuación, se evaluarán estos casos cuidadosamente en función de su impacto sobre los resultados y del criterio clínico.

Análisis de valores atípicos

Se aplicó la **regla de Tukey** para detectar valores atípicos en cada variable numérica del conjunto de datos. Esto consiste en identificar observaciones que se ubican por fuera del rango intercuartílico ampliado:

$$[Q1 - 1.5 \times IQR ; Q3 + 1.5 \times IQR]$$

A continuación, se detallan los valores atípicos detectados:

Variable	Valores_Atípicos
grc	47.5
peso	363.15 - 262.75
altura	73.75
cuello	51.2 - 31.5 - 31.1
pecho	136.2 - 128.3 - 128.3
abdomen	148.1 - 126.2 - 122.1
cadera	116.1 - 147.7 - 125.6 - 115.5
muslo	87.3 - 72.5 - 74.4 - 72.9
rodilla	49.1 - 45 - 46

Análisis multivariado de casos con valores atípicos

Con el objetivo de contextualizar mejor a los individuos con medidas antropométricas fuera de rango, se identificaron los casos que presentaban valores atípicos en al menos una de las siguientes variables: `pecho`, `cadera`, `muslo` o `rodilla`. A estos casos se les analizó su circunferencia abdominal y el porcentaje de grasa corporal (`grc`).

La siguiente tabla muestra un resumen de estas observaciones:

abdomen	pecho	cadera	muslo	rodilla	grc
113.7	115.2	112.4	68.5	45.0	38.1
148.1	136.2	147.7	87.3	49.1	35.2
126.2	128.3	125.6	72.5	39.6	34.5
115.9	114.9	111.9	74.4	40.6	34.3
113.4	115.8	109.8	65.6	46.0	32.6
115.6	117.0	116.1	71.2	43.3	32.3
94.7	128.3	93.8	54.8	39.4	22.2
102.9	108.3	114.4	72.9	43.5	19.6

Para visualizar esta relación se construyó un gráfico de dispersión entre `abdomen` y `grc`, destacando que:

- La mayoría de estos individuos presentan altos valores de `abdomen` y `grc`, lo que respalda su perfil como personas con obesidad severa.
- Algunos casos, si bien son atípicos en otras medidas (por ejemplo, `pecho` o `muslo`), tienen un % de grasa corporal moderado, lo cual puede deberse a contexturas particulares (personas con alta masa muscular o distribución atípica de grasa).
- Esto refuerza la idea de que no todos los outliers son errores: en algunos casos reflejan variabilidad real de la población.

El siguiente gráfico ilustra esta relación:

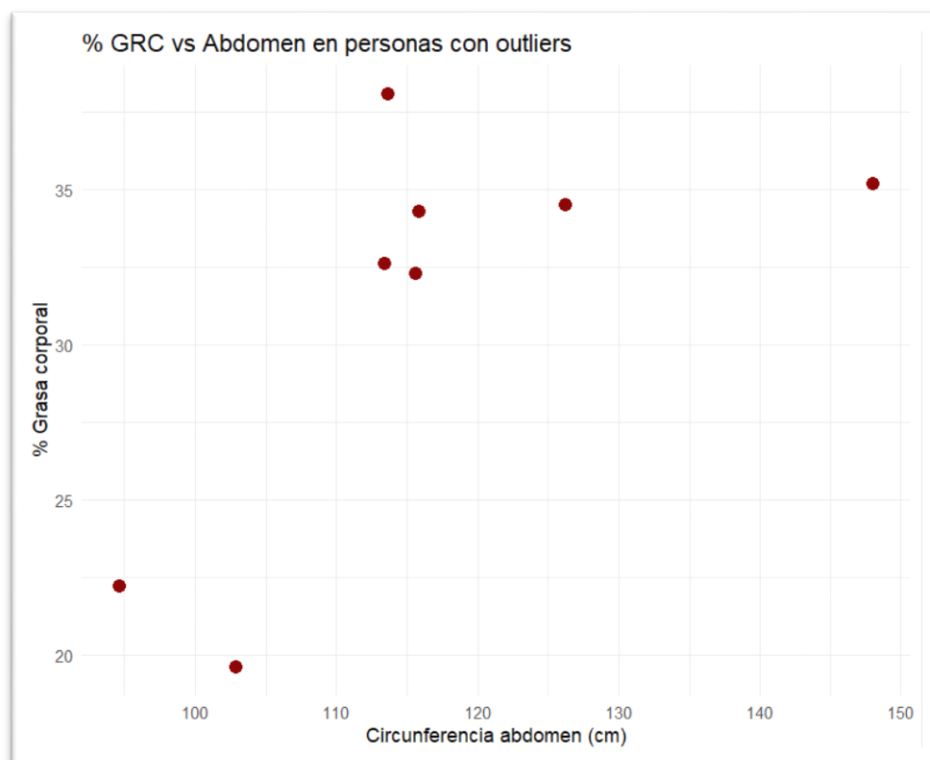


Ilustración 3 - Porcentaje GRC vs Abdomen con outliers

Análisis del impacto de los outliers en abdomen sobre el modelo de regresión

Para estudiar la influencia de los valores atípicos en la variable `abdomen`, se aplicó nuevamente la regla de Tukey. Se identificaron como outliers aquellos casos cuya circunferencia abdominal se encontraba fuera del rango:

$$[Q1 - 1.5 \times IQR ; Q3 + 1.5 \times IQR]$$

Se detectaron **3 casos atípicos** con valores elevados de circunferencia abdominal (mayores a 122 cm). Estos registros presentan, además, altos porcentajes de grasa corporal y peso, por lo que podrían corresponder a individuos con obesidad severa.

Casos con valores atípicos en la variable abdomen								
grc	peso	altura	cuello	pecho	abdomen	cadera	muslo	rodilla
35.2	363.1	180.6	51.2	136.2	148.1	147.7	87.3	49.1
34.5	262.8	171.9	43.2	128.3	126.2	125.6	72.5	39.6
47.5	219.0	160.0	41.2	119.8	122.1	112.8	62.5	36.9

Comparación de modelos lineales

Se ajustaron dos modelos lineales simples, *con y sin outliers*, para predecir el porcentaje de grasa corporal en función del `abdomen`:

- **Modelo 1:** utilizando todos los datos (incluye outliers).
- **Modelo 2:** excluyendo los outliers detectados en `abdomen`.

La siguiente tabla resume los principales indicadores de ajuste para ambos modelos de regresión lineal:

MODELO	INTERCEPTO	COEF. ABDOMEN	R ² AJUSTADO	ERROR ESTÁNDAR RESIDUAL
CON OUTLIERS	-38.56	0.623	0.6437	4.93
SIN OUTLIERS EN ABDOMEN	-42.10	0.662	0.6462	4.75

El gráfico a continuación muestra la dispersión de los datos y las líneas de ajuste lineal correspondientes a los modelos con (azul) y sin outliers (rojo punteado) en la variable `abdomen`.

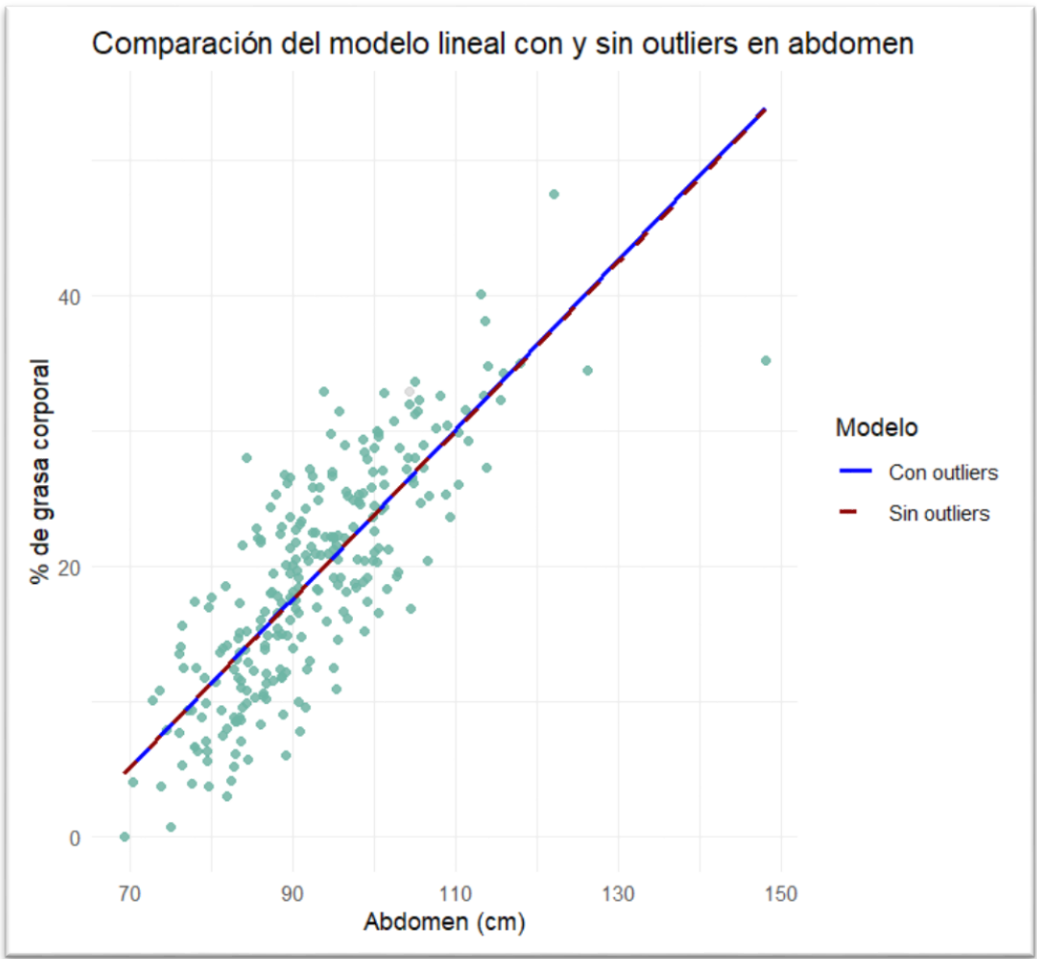


Ilustración 4 - Comparación del modelo lineal con y sin outliers en abdomen

Interpretación de los resultados:

- La pendiente se incrementa ligeramente al eliminar los outliers, indicando que la relación abdomen–grasa corporal se vuelve algo más pronunciada.
- El R^2 ajustado apenas varía, lo que sugiere que los outliers no distorsionan sustancialmente el modelo.
- El error estándar residual disminuye al quitar los outliers, lo cual indica una leve mejora en la precisión de las predicciones.
- Visualmente, se observa que ambos modelos son similares, aunque el modelo sin outliers ajusta mejor los valores intermedios.

Si bien los outliers en `abdomen` representan casos extremos, su influencia sobre el modelo es moderada. Por tanto, son conservados en el análisis final, ya que corresponden a posibles perfiles reales y ayudan a capturar la variabilidad de la población con obesidad.

Decisiones sobre el tratamiento e outliers

- **Altura = 73.75 cm:** este valor resulta notablemente inferior a lo esperable en adultos y se considera un error de medición o carga. Se decidió **eliminar esta observación** del conjunto de datos (`datos_sin_outliers`) por no ser confiable.
- **Peso > 260 kg:** aunque poco frecuentes, estos valores podrían corresponder a casos clínicos de obesidad extrema. Se optó por **conservarlos**, ya que se encuentran dentro del rango fisiológicamente posible.
- **Abdomen, cadera y pecho:** si bien presentan valores altos, estos corresponden a casos plausibles en personas con obesidad severa, por lo tanto, también se **decidió conservarlos**.
- **Cuello, rodilla y muslo:** los valores extremos observados están dentro de los márgenes anatómicos, aunque sean poco frecuentes. Se **mantuvieron** para preservar la variabilidad real del fenómeno.

En resumen, **solo se eliminó el caso con altura = 73.75 cm**, mientras que los demás valores atípicos fueron **conservados** para no perder información relevante en el análisis. Con esta limpieza, se conformó un nuevo objeto con **257 observaciones completas**, sobre el cual se desarrollarán los análisis siguientes.

2.5. Análisis bivariado y correlaciones

Se calculó la matriz de **correlaciones de Pearson** entre las variables numéricas, con el objetivo de explorar las relaciones lineales entre pares de variables del conjunto de datos.

A continuación, se presenta el mapa de calor correspondiente:

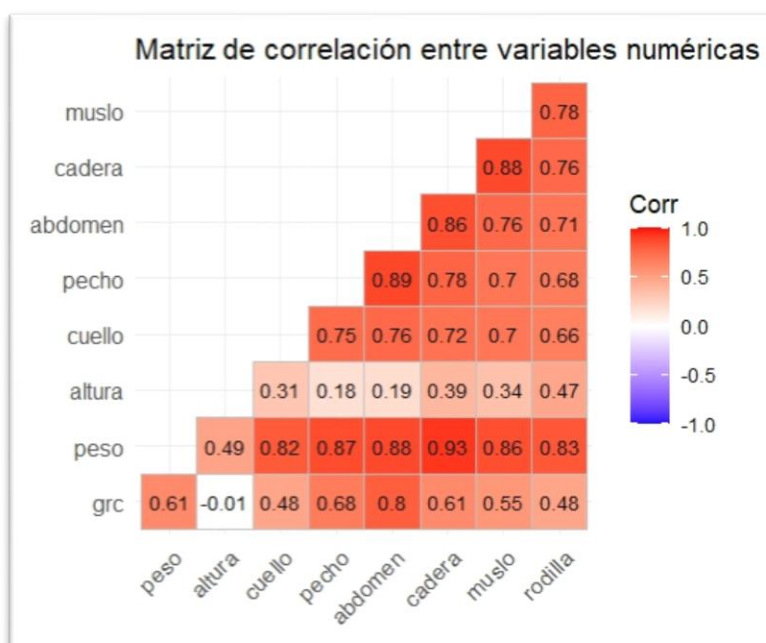


Ilustración 5 - Matriz de correlación entre variables numéricas

Principales observaciones:

Relación entre variables antropométricas:

Se observa una alta correlación positiva entre muchas medidas corporales, lo cual es esperable desde un punto de vista anatómico. Por ejemplo:

VARIABLES	R
PESO Y CADERA:	0.93
PECHO Y ABDOMEN:	0.89
PESO Y ABDOMEN:	0.88
MUSLO Y CADERA:	0.88

Estas fuertes asociaciones indican que un mayor peso corporal suele estar acompañado por mayores circunferencias corporales en distintas regiones.

Relación con el % de grasa corporal (grc):

- La variable más correlacionada con grc es abdomen ($r = 0.8$), seguida por pecho (0.68), peso/cadera (0.61), muslo (0.55) y cuello/rodilla (0.48).
- Esto sugiere que la circunferencia abdominal es una excelente predictora del porcentaje de grasa corporal, lo cual respalda su uso como variable explicativa en modelos de regresión.

Altura:

Presenta muy baja correlación con el resto de las variables, incluyendo `grc` ($r = -0.01$), lo que indica que la estatura no es un buen predictor del % de grasa corporal ni se relaciona fuertemente con otras medidas corporales en este conjunto de datos.

El análisis de correlación evidencia que las medidas corporales están altamente asociadas entre sí, especialmente aquellas vinculadas al perímetro (`abdomen`, `cadera`, `pecho`, `muslo`).

La variable `abdomen` se destaca como el **mejor indicador del porcentaje de grasa corporal**, lo cual justifica su protagonismo en los modelos de predicción que se desarrollarán a continuación.

Exploración bivariada: relación entre `grc` y variables más correlacionadas

Para profundizar el análisis de correlación, se seleccionaron las tres variables que mostraron mayor asociación lineal con el porcentaje de grasa corporal (`grc`): `abdomen`, `pecho` y `cadera`:

Correlación con % de grasa corporal (<code>grc</code>)	
Variable	Correlación
<code>grc</code>	1.000
<code>abdomen</code>	0.803
<code>pecho</code>	0.675
<code>cadera</code>	0.614
<code>peso</code>	0.612
<code>muslo</code>	0.547
<code>rodilla</code>	0.482
<code>cuello</code>	0.476
<code>altura</code>	-0.010

Se construyeron gráficos de dispersión con línea de regresión lineal ajustada para visualizar la relación entre `grc` y cada una de estas variables predictoras:

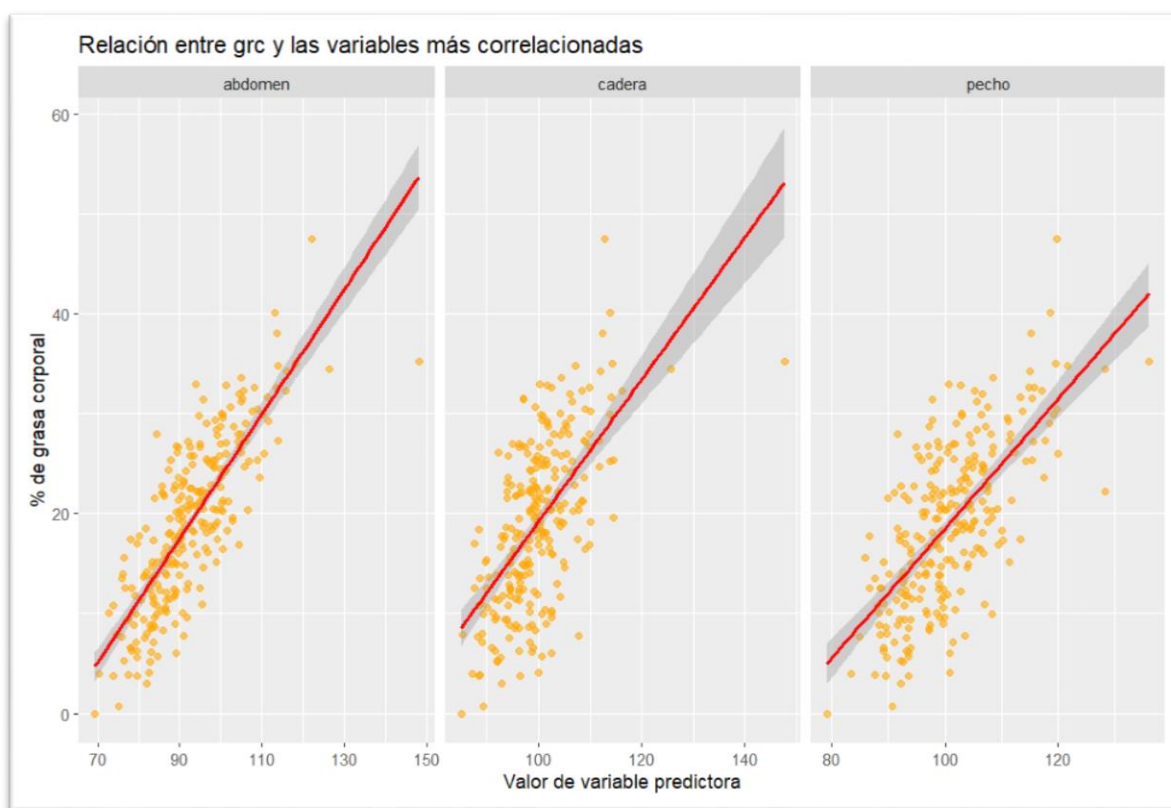


Ilustración 6 - Relación entre grc y las variables más correlacionadas

Interpretación de los resultados:

- **Abdomen:** se confirma visualmente que la relación es fuertemente lineal. La nube de puntos es compacta y la pendiente es bien definida. Es la mejor candidata como predictor del porcentaje de grasa corporal.
- **Cadera:** también muestra una tendencia creciente clara, aunque con mayor dispersión respecto al modelo ajustado.
- **Pecho:** presenta una relación positiva más moderada y con mayor variabilidad, aunque sigue siendo relevante.

Estos gráficos refuerzan la evidencia previa obtenida con la matriz de correlación. En particular, la circunferencia abdominal destaca nuevamente como **la variable más informativa** para explicar la variabilidad del porcentaje de grasa corporal en esta población.

Relaciones entre el % de grasa corporal y las principales variables predictoras

Con el objetivo de visualizar en conjunto la relación entre el porcentaje de grasa corporal (*grc*) y las variables con mayor correlación (*abdomen*, *pecho* y *cadera*), se generó un gráfico de pares coloreado según nivel de grasa corporal. Para ello, se clasificaron los individuos en tres grupos según los **terciles del % de grasa corporal**:

- **Bajo:** menor al primer tercil (color rojo),
- **Medio:** entre primer y segundo tercil (color verde),
- **Alto:** mayor al segundo tercil (color azul).

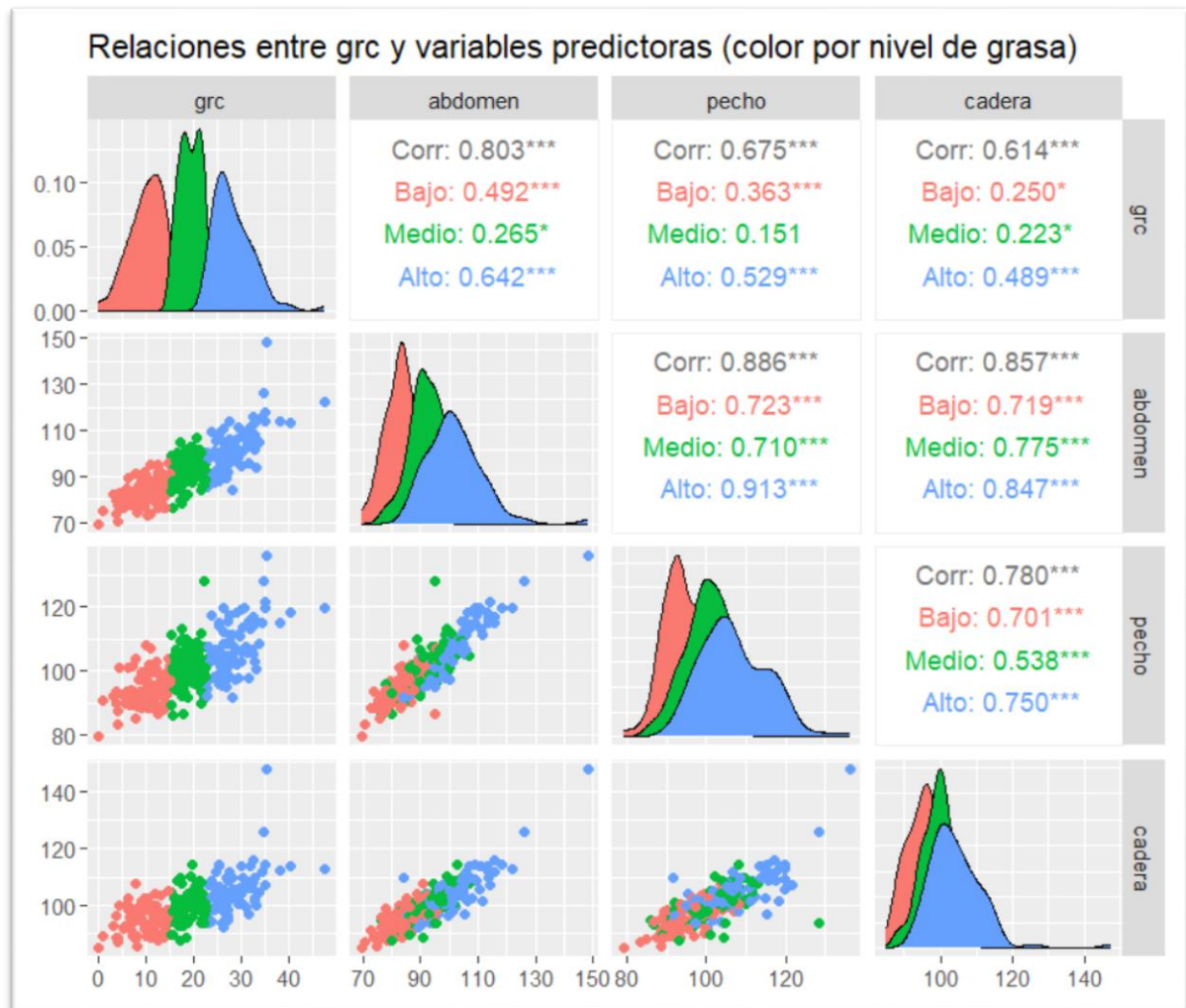


Ilustración 7 - Relaciones entre grc y variables predictoras

Este gráfico permite observar de forma simultánea:

- **Distribuciones** (en la diagonal): los grupos de mayor grasa corporal presentan distribuciones claramente desplazadas hacia valores más altos de abdomen, pecho y cadera.
- **Correlaciones** (en la parte superior): se muestran los coeficientes de correlación de Pearson generales y también por subgrupo. Se destaca que:
 - La relación entre grc y abdomen es la más fuerte ($r = 0.80$), y se mantiene alta incluso dentro de los grupos (ej. $r = 0.64$ en el grupo alto).
 - Las correlaciones con pecho y cadera también son moderadas, aunque disminuyen en el grupo medio.
- **Dispersión** (parte inferior): se visualiza una clara tendencia creciente, especialmente marcada en el grupo de mayor grasa corporal.

Esta visualización evidencia cómo la relación entre variables puede variar según el nivel de grasa corporal, y refuerza la importancia de **abdomen** como principal predictor.

2.6. Distribución del porcentaje de grasa corporal

Antes de ajustar modelos predictivos, se exploró la distribución del porcentaje de grasa corporal para evaluar su comportamiento general. A continuación, se muestra el histograma con su curva de densidad estimada.

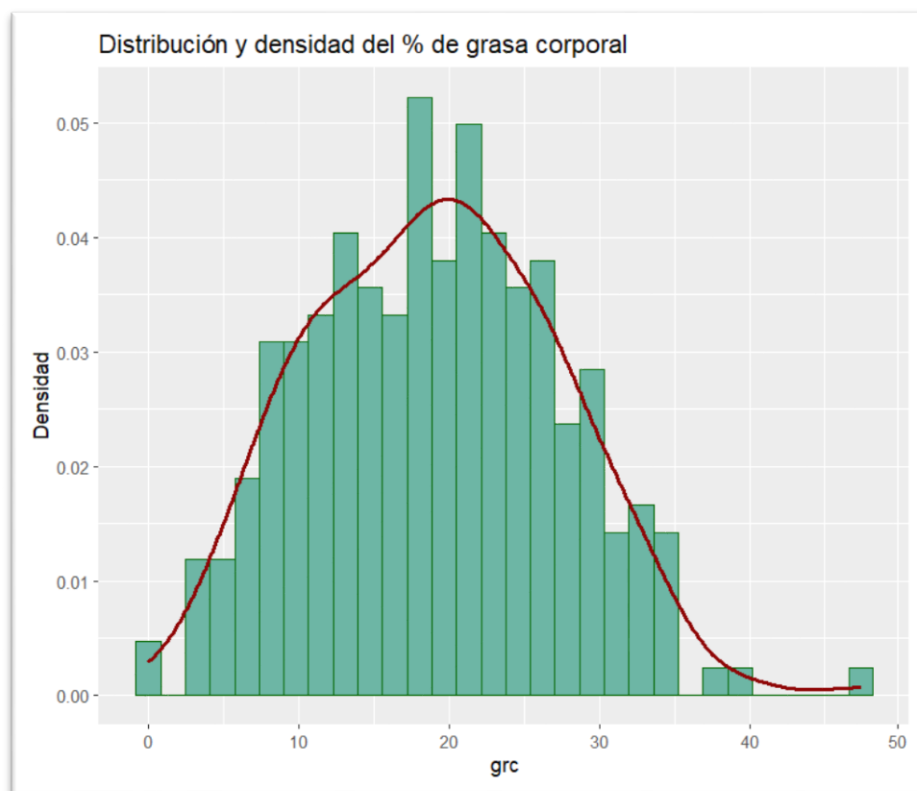


Ilustración 8 - Distribución y densidad del grc

En la figura se presenta la distribución del porcentaje de grasa corporal (*grc*) utilizando un histograma con superposición de curva de densidad. Se observa una distribución aproximadamente simétrica y levemente sesgada a la derecha, con una mayor concentración de individuos en torno al rango 15–25% de grasa corporal.

Aunque la distribución no es perfectamente simétrica, tiene una forma bastante regular, con una única concentración principal de valores. Esto indica que el porcentaje de grasa corporal varía de manera continua entre los individuos y que no hay grupos muy distintos o casos extremadamente alejados del resto.

La curva de densidad (en rojo) refuerza esta idea, mostrando que la mayoría de los valores se concentran en una zona intermedia y que no hay saltos bruscos ni agrupamientos poco frecuentes.

3. Modelo 1: Regresión Simple

El objetivo de esta sección es construir un modelo de regresión lineal simple que permita explicar el porcentaje de grasa corporal (*grc*) a partir de una única variable cuantitativa, la cual esté fuertemente asociada con ella. Para esto, se desarrolló un análisis exploratorio y se aplicaron herramientas estadísticas con el fin de seleccionar la mejor variable predictora y hacer el ajuste del modelo correspondiente.

3.1. Selección de la variable predictora

Para seleccionar la variable predictora, se tuvo en cuenta lo visto anteriormente en el análisis univariado y bivariado. Este análisis permitió identificar las variables que presentan mayor asociación lineal con la variable objetivo.

Basados en la matriz de correlación, la variable con mayor correlación positiva es *abdomen*, con un coeficiente cercano a 0.80, lo que evidencia una relación lineal fuerte y directa con el porcentaje de grasa corporal.

Según la teoría de la regresión lineal simple, una alta correlación entre la variable dependiente y una potencial variable independiente es un buen punto de partida para construir un modelo predictivo. En función de este resultado, se seleccionó la variable *abdomen* como la predictora principal para el modelo:

$$grc \sim abdomen$$

3.2. Ajuste del modelo lineal simple

Para el ajuste del modelo de regresión lineal simple se utilizó la función `lm()` del lenguaje R, que estima los coeficientes mediante el método de mínimos cuadrados ordinarios (OLS).

INTERCEPTO	COEF. ABDOMEN	R ² AJUSTADO	ERROR ESTÁNDAR RESIDUAL
-38.56	0.623	0.6437	4.93

El modelo obtenido fue el siguiente:

$$\widehat{grc} = -38.56 + 0.62 \times abdomen$$

Este modelo indica que existe una relación positiva entre el perímetro abdominal y el porcentaje de grasa corporal: a mayor perímetro, mayor es el valor estimado de grasa.

3.3. Interpretación del modelo

Variable dependiente: % de grasa corporal (*grc*)

Variable predictora: circunferencia de abdomen (*abdomen*)

TÉRMINO	ESTIMACIÓN	ERROR ESTÁNDAR	VALOR T	VALOR P	SIGNIFICANCIA
(INTERCEPTO)	-38.562	2.692	-14.32	< 0.001	***
ABDOMEN	0.623	0.029	21.53	< 0.001	***

Estadísticos del modelo: a continuación, se presentan los principales indicadores del desempeño del modelo.

- Error estándar residual: **4.929**
- Grados de libertad: **255**
- R^2 múltiple: **0.645**
- R^2 ajustado: **0.644**
- Estadístico F: **463.5** (gl = 1 y 255)
- Valor p del modelo: **< 0.001**

El resumen estadístico del modelo proporciona los siguientes resultados clave:

- El coeficiente de la variable `abdomen` es estadísticamente significativo ($p < 2.2e-16$), lo que confirma su influencia sobre la variable dependiente.
- El modelo es globalmente significativo ($F = 463.5$, $p < 2.2e-16$), es decir, explica una proporción considerable de la variabilidad de la respuesta.
- $R^2 = 0.645$: El modelo explica el 64.5% de la variabilidad total en `grc` con la variable `abdomen`. Es un valor moderadamente alto, indicando un buen ajuste.
- R^2 Ajustado = 0.6437: Ajustado por la cantidad de predictores. En este caso con un valor muy similar al R^2 , dado que el modelo es simple (contempla una sola variable, la más correlacionada).

Interpretación del coeficiente:

El valor de 0.62 indica que, por cada centímetro adicional en el perímetro abdominal, se estima un incremento promedio del 0.62% en el porcentaje de grasa corporal.

3.4. Visualización del ajuste

A continuación, se presenta el gráfico del modelo ajustado:

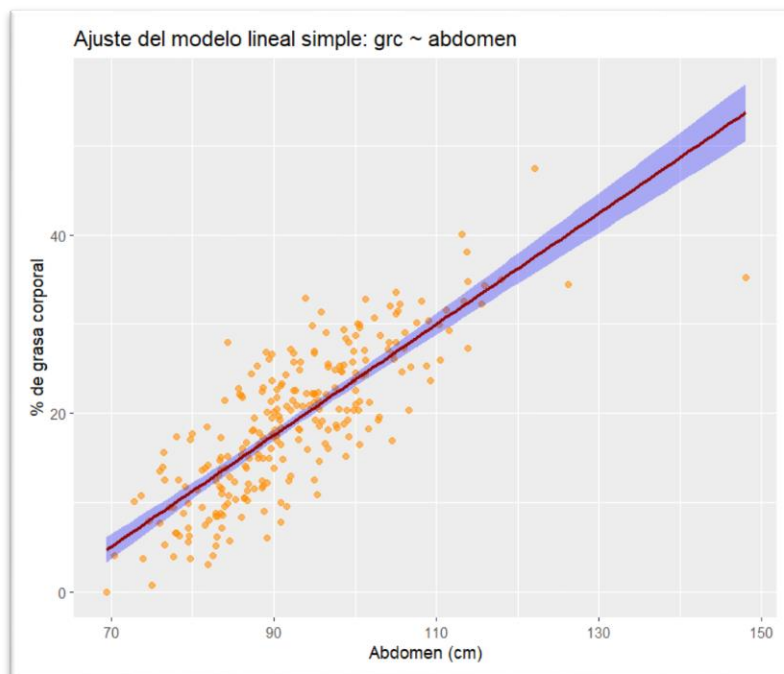


Ilustración 9 - Ajuste del modelo lineal simple

Este gráfico incluye:

- Los puntos de dispersión que representan la relación empírica entre `abdomen` y `grc`.
- La recta de regresión ajustada.
- La banda de confianza del 95% para la media estimada, representada como una zona sombreada alrededor de la recta, que refleja la incertidumbre del modelo en la estimación del valor promedio de `grc` para cada nivel de `abdomen`.

4. Modelo 2: Regresión Múltiple con todas las variables numéricas

4.1. Ajuste del modelo múltiple

A continuación, se desarrolla un modelo de regresión lineal múltiple que relaciona el porcentaje de grasa corporal (`grc`) con todas las variables numéricas disponibles en el conjunto de datos: `peso`, `altura`, `cuello`, `pecho`, `abdomen`, `cadera`, `muslo` y `rodilla`.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$$

Donde:

- y = porcentaje de grasa corporal (grc), variable dependiente
- x_1 = peso (kg)
- x_2 = altura (cm)
- x_3 = cuello (cm)
- x_4 = pecho (cm)
- x_5 = abdomen (cm)
- x_6 = cadera (cm)
- x_7 = muslo (cm)
- x_8 = rodilla (cm)
- β_0 = intercepto del modelo
- β_i = coeficientes

4.2. Evaluación global del modelo

Para evaluar el modelo de regresión lineal múltiple y saber si las variables explicativas tienen un efecto significativo sobre la variable dependiente, porcentaje de grasa corporal, usamos el test F. Planteamos las hipótesis.

- H_0 = todos los coeficientes de las variables predictoras son iguales a cero.
- H_1 = al menos uno de los coeficientes es distinto de cero

Valor del estadístico F en el modelo:

- Estadístico F: 74,29 gl 8 y 248
- p-value: $< 2,2e-16$

Dado que el valor p del test F es chico (menor a 0,05) \rightarrow se rechaza la hipótesis nula. Lo que indica que el modelo es significativo, al menos una de las variables explica el porcentaje de grasa corporal.

Por otro lado, el R^2 ajustado es de 0,6961, el modelo explica alrededor del 70% de la variabilidad en el porcentaje de grasa corporal.

4.3. Ecuación del modelo ajustado

$$\widehat{grc} = 2,74548 - 0,02103 \text{ peso} - 0,09432 \text{ altura} - 0,79179 \text{ cuello} - 0,03021 \text{ pecho} \\ + 0,91482 \text{ abdomen} - 0,19009 \text{ cadera} + 0,16202 \text{ muslo} - 0,14177 \text{ rodilla}$$

Interpretación de los coeficientes:

Cada coeficiente β_i indica el cambio esperado en el porcentaje de grasa corporal ante una unidad de aumento en esa variable, manteniendo las demás constantes. Por ejemplo, un aumento de 1 cm en la circunferencia del abdomen se asocia con un aumento promedio de 0.91 puntos en grc, si todo lo demás permanece constante.

4.4. Análisis de significancia individual de variables

Para evaluar si cada variable del modelo de regresión múltiple tiene un efecto significativo sobre el porcentaje de grasa corporal (grc), llevamos a cabo las pruebas t individuales asociadas a cada coeficiente del modelo. Para evaluar si cada variable independiente tiene un efecto significativo sobre el porcentaje de grasa corporal (grc), se realiza una prueba de hipótesis para cada coeficiente β_i (excepto el intercepto):

- $H_0 = \beta_i = 0 \rightarrow$ la variable no tiene efecto sobre grc.
- $H_1 = \beta_i$ distinto de 0 \rightarrow la variable si tiene efecto sobre grc.

Se utiliza un nivel de significancia $\alpha = 0,05$

Si el p-valor $< 0,05$, se rechaza H_0 , lo que indica que la variable correspondiente es significativa.

TÉRMINO	ESTIMACIÓN	ERROR ESTÁNDAR	VALOR T	Pr(> t)	
PESO	-0.0210	0.0548	-0.3837	7.015040e-01	> 0.05
ALTURA	-0.0943	0.0698	-1.3517	1.777023e-01	> 0.05
CUELLO	-0.7918	0.2134	-3.7104	2.554726e-04	< 0.05
PECHO	-0.0302	0.0919	-0.3286	7.427344e-01	> 0.05
ABDOMEN	0.9148	0.0752	12.1646	5.199799e-27	< 0.05
CADERA	-0.1901	0.1360	-1.3981	1.633208e-01	> 0.05
MUSLO	0.1620	0.1309	1.2376	2.170179e-01	> 0.05
RODILLA	-0.1418	0.2223	-0.6378	5.241975e-01	> 0.05

Tal como se puede observar en la tabla anterior, las únicas variables significativas al 5% son cuello ($2.554726e-04 < 0.05$) y abdomen ($5.199799e-27 < 0.05$).

4.5. Diagnóstico de multicolinealidad

La multicolinealidad ocurre cuando dos o más variables explicativas están fuertemente relacionadas entre sí. Esto dificulta saber cuál de ellas realmente está afectando a la variable que queremos predecir, y puede hacer que los coeficientes del modelo sean inestables o poco confiables.

Para verificar la presencia de multicolinealidad entre las variables explicativas, se analiza el Factor de Inflación de la Varianza (VIF) para cada variable independiente.

Como se puede observar, las variables peso y cadera tienen un VIF > 10 , lo que indica un grado alto de multicolinealidad. Mientras que, las variables pecho, abdomen y muslo tienen un VIF > 5 .

Esto indica que algunas variables están fuertemente correlacionadas entre sí, lo que puede dificultar la interpretación de sus efectos individuales en el modelo. Sin embargo, el modelo en conjunto sigue siendo significativo (Estadístico F = 74.29, $p < 0.001$), aunque sería recomendable considerar un modelo ajustado con selección de variables.

TÉRMINO	VIF
PESO	31.485744
ALTURA	2.629400
CUELLO	3.292974
PECHO	7.675406
ABDOMEN	7.913028
CADERA	11.577700
MUSLO	5.670690
RODILLA	3.550431

5. Comparación y Selección de Modelos de Regresión

5.1. Modelo con selección automática de variables

Se aplicó la técnica de selección automática de variables mediante el criterio de información AIC, utilizando el método stepwise. Este método empieza con un modelo de regresión simple y en cada paso añade una variable, pero verifica si alguna de las variables presentes en el modelo puede ser eliminada. Este proceso identificó un modelo parsimonioso que incluye cuatro predictores: *altura*, *cuello*, *abdomen* y *cadera*.

Ecuación ajustada:

$$\widehat{grc} = 7.44779 - 0.12592 \text{ altura} - 0.83080 \text{ cuello} + 0.86565 \text{ abdomen} - 0.14782 \text{ cadera}$$

5.2. Modelo con selección manual

Con el objetivo de obtener un modelo estadísticamente más robusto, se construyó un modelo de regresión lineal múltiple seleccionando manualmente las variables predictoras, teniendo en cuenta los siguientes criterios:

- Correlación con la variable objetivo (*grc*): se seleccionaron las variables más correlacionadas con el porcentaje de grasa corporal según la matriz de correlación.
- Significancia estadística individual de las variables explicativas: se priorizaron las variables que mostraron un valor $p < 0,05$ en el modelo completo.
- Multicolinealidad: se descartaron variables con un Factor de Inflación de la Varianza (VIF) alto para evitar colinealidad entre predictoras.

Siguiendo este criterio se seleccionaron las variables *cuello* y *abdomen*.

Ecuación:

$$\widehat{grc} = \beta_0 + \beta_1 \text{ abdomen} + \beta_2 \text{ cuello}$$

$$\widehat{grc} = -15.68963 + 0.80067 \text{ abdomen} - 1.03392 \text{ cuello}$$

5.3. Modelo completo (todas las variables)

En este apartado se presenta el modelo de regresión con todas las variables predictoras del conjunto de datos realizado en el apartado 4.

Ecuación ajustada:

$$\widehat{grc} = 2,74548 - 0,02103 \times \text{peso} - 0,09432 \times \text{altura} - 0,79179 \times \text{cuello} - 0,03021 \times \text{pecho} + 0,91482 \times \text{abdomen} - 0,19009 \times \text{cadera} + 0,16202 \times \text{muslo} - 0,14177 \times \text{rodilla}$$

5.4. Comparación de modelos y selección final

A continuación, se presenta una tabla comparativa donde se presentan los coeficientes estimados, error estándar, valores t y p- valores de cada variable incluida en los tres modelos presentados en los puntos anteriores.

- En el modelo automático, las variables *altura*, *cuello* y *abdomen* resultan significativas ($p < 0.05$), mientras que *cadera* no lo es.
- En el modelo manual, se seleccionaron solo dos variables: *abdomen* y *cuello*. Ambas son altamente significativas.
- En el modelo completo, que incluye todas las variables, solo *abdomen* y *cuello* resultan significativas. El resto de las variables, como *peso*, *pecho*, *muslo* y *rodilla*, presentan p- valores elevados.

TÉRMINO	ESTIMACIÓN	ERROR ESTÁNDAR	VALOR T	Pr(> t)
MODELO SELECCIÓN AUTOMÁTICA VARIABLES				
INTERCEPTO	7.44779	7.92615	0.940	0.3483
ALTURA	-0.12592	0.04957	-2.540	0.0117 < 0.05
CUELLO	-0.83080	0.18602	-4.466	1.2E-05 < 0.05
ABDOMEN	0.86565	0.05957	14.532	< 2E-16 < 0.05
CADERA	-0.14782	0.08684	-1.702	0.0899 > 0.05
MODELO CON SELECCIÓN MANUAL				
INTERCEPTO	-15.68963	4.78967	-3.276	0.0012
ABDOMEN	0.80067	0.04171	19.196	2E-16 < 0.05
CUELLO	-1.03392	0.18348	-5.635	4.64E-08 < 0.05
MODELO COMPLETO (TODAS LAS VARIABLES)				
INTERCEPTO	2.74548	20.3199	0.1351	8.926320E-01
PESO	-0.0210	0.0548	-0.3837	7.015040e-01 > 0.05
ALTURA	-0.0943	0.0698	-1.3517	1.777023e-01 > 0.05
CUELLO	-0.7918	0.2134	-3.7104	2.554726e-04 < 0.05
PECHO	-0.0302	0.0919	-0.3286	7.427344e-01 > 0.05
ABDOMEN	0.9148	0.0752	12.1646	5.199799e-27 < 0.05
CADERA	-0.1901	0.1360	-1.3981	1.633208e-01 > 0.05
MUSLO	0.1620	0.1309	1.2376	2.170179e-01 > 0.05
RODILLA	-0.1418	0.2223	-0.6378	5.241975e-01 > 0.05

La siguiente tabla compara:

- El error estándar de los residuos, que representa la desviación típica entre los valores observados y los predichos por el modelo.
- El coeficiente de determinación (R^2), que indica el porcentaje de variabilidad explicada.
- El R^2 ajustado, que indica el porcentaje de variabilidad explicada por el modelo, teniendo en cuenta el número de variables predictoras.
- El estadístico F y su p-valor global, que evalúa la hipótesis nula de que todos los coeficientes sean iguales a cero.

El modelo de selección automática presenta el mayor R^2 ajustado, el modelo explica el 69,6% de la variabilidad del porcentaje de grasa corporal.

MODELO	ERROR ESTÁNDAR	R^2	R^2 AJUSTADO	ESTADÍSTICO F	VALOR P
M. SELECCIÓN AUTOMÁTICA	4.537	0.7029	0.6982	149	< 2.2E-16
M. SELECCIÓN MANUAL	4.656	0.6845	0.682	275.6	< 2.2E-16
M. COMPLETO	4.552	0.7056	0.6961	74.29	< 2.2E-16

Esta tabla muestra los valores del Factor de Inflación de la Varianza (VIF) para cada variable, en los distintos modelos.

- Tal como se presentó anteriormente, el modelo completo presenta varios problemas de colinealidad, entre ellas *peso* y *cadera* presentan un $VIF > 10$

VIF			
TÉRMINO	M. SELECCIÓN AUTOMÁTICA	M. SELECCIÓN MANUAL	M. COMPLETO
PESO	--	--	31.485744
ALTURA	1.336283	--	2.629400
CUELLO	2.519633	2.326721	3.292974
PECHO	--	--	7.675406
ABDOMEN	4.998865	2.326721	7.913028
CADERA	4.755347	--	11.577700
MUSLO	--	--	5.670690
RODILLA	--	--	3.550431

Tras comparar los tres modelos, se selecciona el modelo con selección automática como el más adecuado. Tiene el mayor R^2 ajustado (0.6982), el menor error estándar (4.537) y baja multicolinealidad (todos los $VIF < 5$). Además, todas sus variables son estadísticamente significativas.

6. Evaluación del Modelo Seleccionado

6.1. Verificación de supuestos del modelo lineal

En este apartado se analizará el modelo seleccionado: modelo con selección automática de variables con método stepwise.

Para concluir que los resultados obtenidos a partir del modelo de regresión lineal múltiple son válidos se deben cumplir los siguientes supuestos clásicos:

- Linealidad: implica que la relación entre las variables independientes y la variable dependiente debe ser lineal.
- Media de los errores es igual a cero.
- Homocedasticidad: la varianza de los errores debe ser constante en todos los niveles de las variables predictoras.
- Normalidad: los errores se distribuyen normalmente.
- Independencia: los errores de la regresión no deben estar correlacionados entre sí.
- Las variables regresoras no son colineales (no hay multicolinealidad)

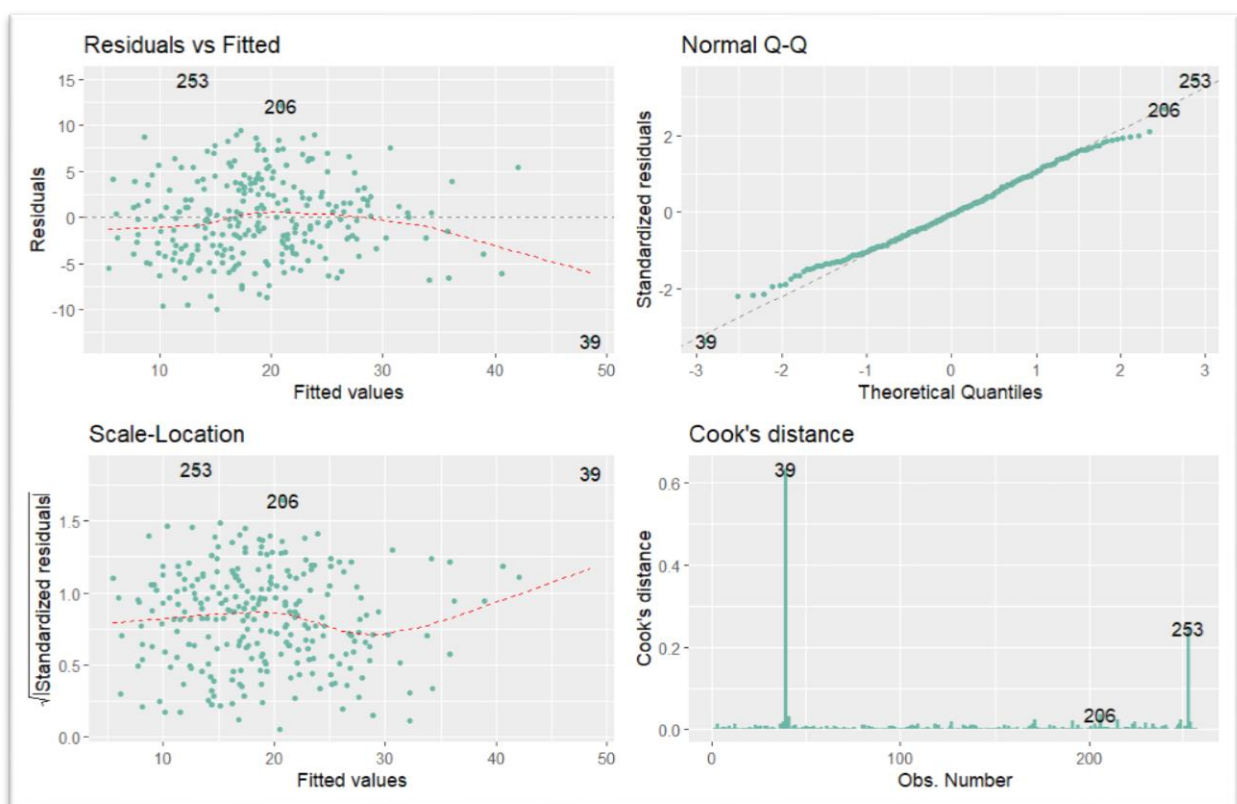


Ilustración 10 - Gráficos de diagnóstico del modelo lineal

Análisis de los gráficos

- Residuals vs Fitted: se observa una ligera curvatura y leve patrón en forma de embudo hacia la derecha, lo que podría sugerir una leve **heterocedasticidad**. Sin embargo, no se identifican patrones evidentes, podemos concluir que el supuesto de **linealidad** puede considerarse razonablemente cumplido.
- Q-Q Residuals: compara los cuantiles de nuestros datos con los cuantiles teóricos de la distribución normal estándar $N(0,1)$. La mayoría de los puntos caen cerca de la línea diagonal de referencia, con algunas desviaciones en las colas. Esto sugiere que el supuesto de

normalidad se cumple en términos generales, aunque podría haber algunas leves desviaciones en los extremos.

- Scale-Location: evalúa si la varianza de los residuos es constante. La curva roja muestra una ligera pendiente ascendente, lo que puede indicar **heterocedasticidad** leve. Sin embargo, la variación no parece crítica, por lo que la homocedasticidad se considera aceptable para este análisis.

En este trabajo, los datos analizados corresponden a un corte transversal, es decir, observaciones recolectadas en un mismo momento. Dado que no se trata de una serie temporal ni de datos con estructura secuencial, no se espera la existencia de autocorrelación entre los errores. Por lo tanto, se asume la **independencia** de los errores.

En cuanto a la presencia de **multicolinealidad**, se analizó la matriz de correlaciones entre las variables seleccionadas en el modelo. Se observó una correlación alta entre algunas de ellas, especialmente entre *abdomen* y *cadera* ($r = 0.86$), *cuello* y *abdomen* ($r = 0.72$), y *cuello* y *cadera* ($r = 0.76$). Este patrón sugiere una posible colinealidad entre variables.

Sin embargo, para evaluar si esta colinealidad representa un problema para el modelo, se calculó el factor de inflación de la varianza (VIF) presentado en el punto 5.4. Las variables incluidas en el modelo seleccionado presentan valores aceptables (todos menores a 5), por lo tanto la colinealidad no es problema en nuestros datos.

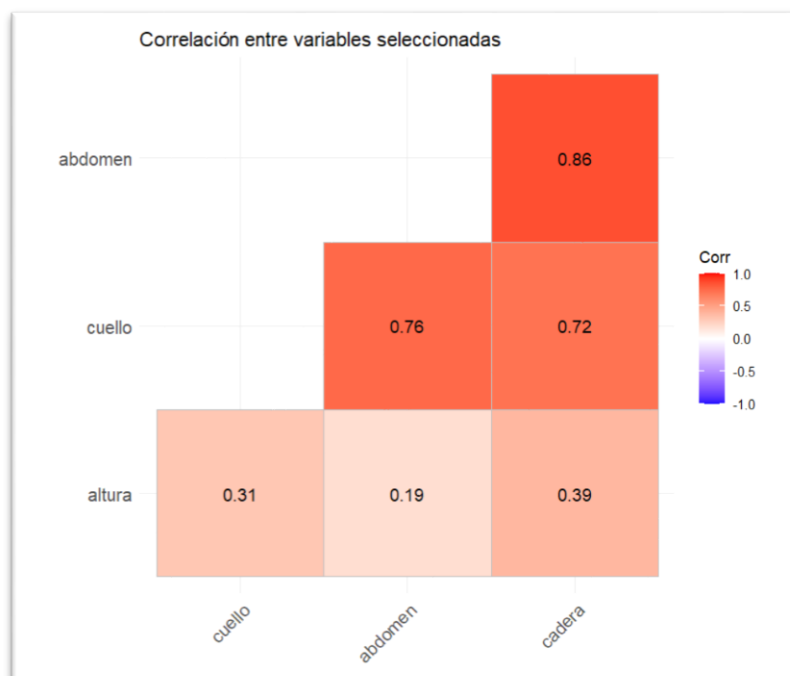


Ilustración 11 - Matriz de correlación modelo lineal

Detección de multicolinealidad: Factores de Inflación de la Varianza (VIF)

TÉRMINO	VIF
ALTURA	1.336283
CUELLO	2.519633
ABDOMEN	4.998865
CADERA	4.755347

6.2. Detección de observaciones influyentes

Sobre el modelo seleccionado por selección automática de variables (altura, cuello, abdomen, cadera), se evaluó la presencia de observaciones influyentes utilizando la Distancia de Cook y los valores de leverage.

Distancia de Cook

Se calcularon los valores de Distancia de Cook para cada observación y se identificaron **12 observaciones influyentes**.

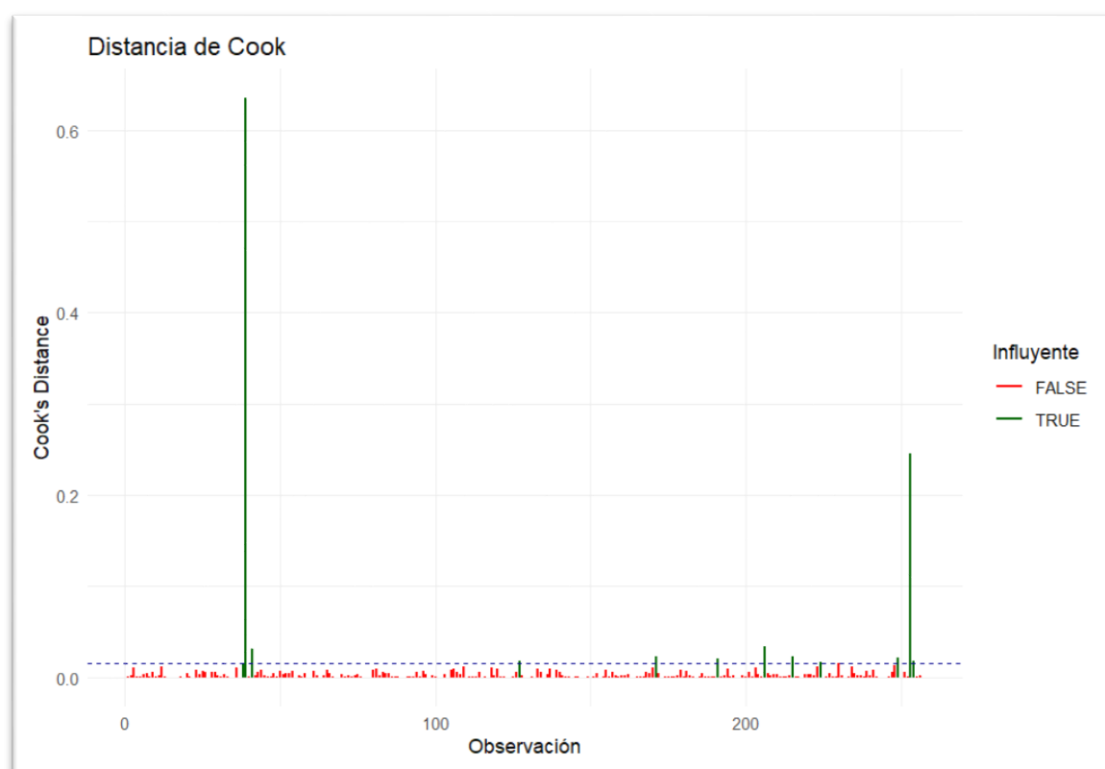


Ilustración 12 - Distancia de Cook

- El gráfico muestra que la mayoría de las observaciones tienen valores muy bajos (cerca de 0).
- Algunas pocas observaciones sobresalen significativamente (en verde en el gráfico).

- El umbral usado fue $\frac{4}{n}$, y todas las barras verdes están por encima de esa línea azul discontinua.
- Estas observaciones tienen mucha influencia en los coeficientes del modelo: si se eliminan, el modelo podría cambiar significativamente.

Leverage

El leverage (o apalancamiento) mide qué tan alejada está una observación del centroide del espacio de los predictores. Se utilizó como umbral $\frac{2p}{n}$, donde $p = 5$ (4 predictores + intercepto), detectándose **14 observaciones con leverage alto**.

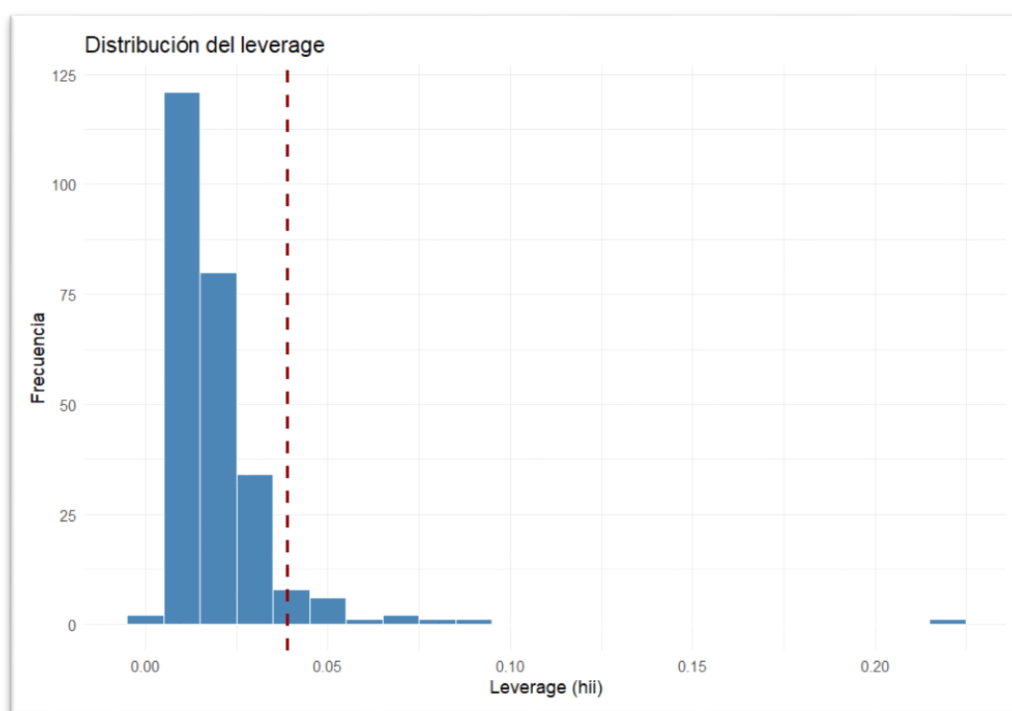


Ilustración 13 - Distribución del leverage

- El histograma del leverage muestra que la mayoría de los puntos están agrupados cerca de 0.
- Solo unas pocas observaciones tienen leverage mayor al umbral (línea roja discontinua).
- Estas observaciones tienen valores de las variables predictoras (altura, cuello, abdomen, cadera) muy distintos del promedio.

Observaciones críticas: influyentes y con leverage alto

La intersección de ambas métricas reveló **5 observaciones especialmente críticas**, listadas a continuación:

- IDs influyentes por Cook (no leverage alto): 38, 127, 171, 191, 206, 224, 254
- IDs con leverage alto (no influyentes por Cook): 10, 36, 42, 58, 105, 144, 188, 241, 251

Al cruzar ambas métricas, se identificaron **5 observaciones clave**, que son influyentes, atípicas y tienen un alto potencial de afectar el modelo.

- IDs en la intersección (Cook + leverage alto): 39, 41, 215, 249, 253

Nro Obs	% Grasa	Peso (kg)	Altura (cm)	Cuello (cm)	Pecho (cm)	Abdomen (cm)	Cadera (cm)	Muslo (cm)	Rodilla (cm)	modelo	grc_cat
39	35.2	363.1	180.6	51.2	136.2	148.1	147.7	87.3	49.1	Sin outliers	Alto
41	34.5	262.8	171.9	43.2	128.3	126.2	125.6	72.5	39.6	Sin outliers	Alto
215	47.5	219.0	160.0	41.2	119.8	122.1	112.8	62.5	36.9	Sin outliers	Alto
249	29.3	186.8	165.0	38.9	111.1	111.5	101.7	60.3	37.3	Sin outliers	Alto
253	28.0	202.2	180.6	34.2	91.6	84.3	109.8	61.3	37.3	Sin outliers	Alto

Las observaciones 39 y 41 presentan valores extremos de peso y circunferencia abdominal, lo cual justifica su influencia, siendo casos compatibles con obesidad severa.

Las observaciones 249 y 253, aunque no presentan valores extremos en forma aislada, pueden ejercer influencia por combinaciones atípicas de variables (por ejemplo bajo porcentaje de grasa con medidas moderadas).

La observación 215 también muestra influencia por su combinación específica de valores.

Decisión

Se optó por **conservar todas las observaciones** al no encontrarse errores de carga evidentes. Estos casos reflejan la variabilidad clínica esperable en una población con distintos perfiles corporales. Su exclusión podría limitar la capacidad del modelo para generalizar.

6.3. Predicción para nuevos casos

A partir del modelo ajustado mediante selección automática de variables por el **método stepwise** (función `stepAIC`), se determinaron como predictoras significativas del porcentaje de grasa corporal (`grc`) las siguientes variables cuantitativas: `altura`, `cuello`, `abdomen` y `cadera`. Estas variables fueron seleccionadas en base al criterio de información AIC, que penaliza la complejidad del modelo, priorizando un buen ajuste con la menor cantidad de variables posibles.

Construcción de nuevos casos

Con el objetivo de evaluar la capacidad predictiva del modelo, se definieron tres casos representativos, contemplando combinaciones plausibles y realistas dentro del rango observado en el conjunto de datos:

CASO	ALTURA	CUELLO	ABDOMEN	CADERA
1	165	35	85	90
2	175	40	100	100
3	185	45	115	110

Utilizando el modelo ajustado, se obtuvieron los siguientes resultados:

Intervalo de confianza del 95% para la media estimada de `grc`:

CASO	VALOR PREDICHO (GRC)	IC 95% INFERIOR	IC 95% SUPERIOR
1	17.87 %	16.62 %	19.11 %
2	23.96 %	22.95 %	24.97 %
3	30.05 %	28.11 %	32.00 %

Intervalo de predicción del 95% para un nuevo individuo:

CASO	VALOR PREDICHO (GRC)	IP 95% INFERIOR	IP 95% SUPERIOR
1	17.87 %	8.85 %	26.89 %
2	23.96 %	14.97 %	32.95 %
3	30.05 %	20.91 %	39.20 %

Interpretación

Se observa que el valor estimado de `grc` aumenta progresivamente con el incremento en las variables predictoras, lo cual es coherente con la estructura del modelo. La variable abdomen, en particular, contribuye de manera significativa, al haber mostrado una fuerte correlación con el porcentaje de grasa en etapas anteriores del análisis.

Como es esperable en modelos lineales, los intervalos de predicción son más amplios que los intervalos de confianza. Esto se debe a que los primeros incorporan tanto la incertidumbre en la estimación de la media como la variabilidad individual de los datos (error aleatorio), mientras que los IC sólo consideran la variabilidad de la media poblacional.

Por ejemplo, en el caso 2, se estima un `grc` de 23.96%, pero para un nuevo individuo con esas características, el modelo advierte que su valor real podría variar razonablemente entre 14.97% y 32.95%. Esta amplitud destaca la importancia de no interpretar el valor predicho como una certeza, sino como un estimador central dentro de un rango esperable.

El modelo ajustado logra generar estimaciones puntuales coherentes y ofrece bandas de confianza adecuadas para la media y para nuevos individuos. No obstante, la amplitud de los intervalos de

predicción sugiere que existen otros factores no incluidos en el modelo que también influyen sobre el porcentaje de grasa corporal. Esto resalta el valor predictivo del modelo, pero también sus limitaciones inherentes al tratarse de una simplificación lineal del fenómeno fisiológico.

7. Conclusiones

Análisis Exploratorio de Datos (EDA):

- Se cargó el dataset `obesidad25`, compuesto por 260 individuos y 9 variables numéricas.
- Se eliminaron 2 valores faltantes en las variables `cuello` y `rodilla`, quedando un total de 258 observaciones sin registros duplicados.
- Los gráficos univariados revelaron asimetrías y posibles valores atípicos, especialmente en las variables `altura` y `peso`.
- La mayoría de estos outliers se conservaron por ser plausibles en casos de obesidad severa, excepto un valor atípico de `altura` que fue descartado.
- El análisis bivariado mostró que la circunferencia abdominal (`abdomen`), es la variable con mayor correlación con el porcentaje de grasa corporal (`grc`), con un coeficiente de correlación de 0.8.

Modelo 1: Regresión Simple:

- Se ajustó un modelo simple para explicar el `grc` a partir de una única variable predictora, la circunferencia abdominal (`abdomen`), dada su alta correlación con el `grc`.
- El modelo resultante fue:

$$\widehat{grc} = -38.56 + 0.62 \times abdomen$$

- Este modelo es estadísticamente significativo y explica el 64.5% de la variabilidad del `grc` ($R^2 = 0.645$). Un aumento de 1 cm en la circunferencia abdominal predice, en promedio, un incremento del 0.62% en el porcentaje de grasa corporal.

Modelo 2: Regresión Múltiple con todas las variables numéricas:

- Se construyó un modelo que incluye las 8 variables numéricas para predecir el `grc`.
- El modelo global fue significativo ($p < 0.001$) y explicó cerca del 70% de la variabilidad del `grc` (R^2 ajustado = 0.6961).
- Sin embargo, al analizar la significancia individual, solo las variables `cuello` y `abdomen` resultaron estadísticamente significativas.
- Se detectaron problemas de multicolinealidad entre algunas variables, como `peso` y `cadera`, con VIF superiores a 10.

Comparación y Selección de Modelos de Regresión:

- Se compararon tres modelos: el modelo completo, uno con selección manual y otro con selección automática mediante stepwise basado en AIC.
- El modelo seleccionado automáticamente incluyó las variables `altura`, `cuello`, `abdomen` y `cadera`.
- Este modelo fue elegido por tener el mayor R^2 ajustado (0.6982) y el menor error estándar residual (4.537).
- Además, todas las variables seleccionadas, salvo `cadera`, fueron significativas y no presentó problemas de multicolinealidad (todos los VIF < 5).

Evaluación del Modelo Seleccionado:

- Se verificaron los supuestos del modelo seleccionado automáticamente.
- Los gráficos diagnósticos indicaron que la linealidad, normalidad de residuos y homocedasticidad se cumplen razonablemente, pese a pequeñas desviaciones.
- Los factores de inflación de la varianza confirmaron que no existía multicolinealidad entre las variables seleccionadas.
- Se identificaron observaciones influyentes mediante Distancia de Cook y leverage, pero se decidió conservarlas al no ser errores evidentes y reflejar variabilidad real en la muestra.
- Las predicciones para nuevos casos mostraron intervalos de predicción más amplios que los de confianza, evidenciando la variabilidad individual.

En síntesis, el análisis exploratorio y los distintos modelos de regresión permitieron identificar las variables clave que explican el porcentaje de grasa corporal, destacándose la circunferencia abdominal y el cuello como predictores relevantes.

El modelo automático optimizó la selección de variables y cumplió con los supuestos estadísticos, lo que garantiza su validez para predicciones. Estos resultados aportan una base sólida para futuras investigaciones y aplicaciones prácticas en la evaluación y monitoreo de la obesidad.

8. Anexos

Se adjuntan los archivos correspondientes al Trabajo Práctico para facilitar su revisión y replicación:

- **Código fuente en R:** Archivo `TP_Regresion_grupo10.R` que contiene todo el script utilizado para el análisis, limpieza de datos, modelado y visualización.
- **Informe en formato HTML:** Archivo `TP_Regresion_grupo10.html` generado a partir del script R Markdown, que presenta el desarrollo completo, gráficos y resultados del trabajo.

Estos materiales permiten reproducir el análisis y verificar los resultados presentados en este informe.