Estimación de porcentaje de grasa corporal por medio de regresión lineal múltiple

Juan David Bonilla Sotelo[1], Luisa Fernanda Guantiva Vargas[2], Andrés Fernando Niño Viancha[3],

21 de agosto de 2024

Resumen: El conocimiento del porcentaje de grasa corporal en una persona puede contribuir a detectar a tiempo enfermedades como la obesidad, la cual aumenta el riesgo de padecer otras enfermedades cronicas no transmitibles con un alto riesgo de muerte, por lo cual por medio de regresión lineal múltiple y diferentes análisis descriptivos y exploratorios analizaremos los factores que más influyen en el calculo del porcentaje de grasa corporal en hombres.
Palabras clave: Índice de grasa Corporal, obesidad, salud, peso, regresión, correlación, modelos, sobrepeso.

1. Introducción

En la actualidad la obesidad es una de las enfermedades más comunes y complejas, de acuerdo con la Organización Mundial de la Salud (OMS), desde el año 1975, la obesidad prácticamente se ha triplicado en todo el mundo, además es catalogada como una epidemia global que afecta a niños, jóvenes y adultos. Las autoridades de la salud han manifestado preocupación por el incremento de casos en todos los países ya que para el año 2016, las cifras indican que más de 1900 millones de adultos, mayores de 18 años, tenían sobrepeso, y de estos, más de 650 millones, que equivalen al 34,21 %, eran personas obesas. Es necesario señalar, que la mayor parte de la población mundial (65 %), vive en países en los cuales la obesidad y el sobrepeso, son causa de defunciones, mucho más que la insuficiencia ponderal debido a la desnutrición.

Un elevado porcentaje de grasa corporal es considerado un factor de riesgo que desencadena múltiples enfermedades crónicas no transmisibles con riesgo de muerte, principalmente por agudas o crónico-degenerativas como diabetes tipo 2,cáncer, hipertensión, dislipidemia, problemas en el higado y vesicula, apnea, problemas ginecologicos, accidentes cardiovasculares e incluso según científicos de la Universidad de Aarhus , en Dinamarca, concluyen que la grasa corporal de más incrementa la posibilidad de contraer esta enfermedad hasta un 15 %. Los investigadores sugieren que esta relación causa-efecto se debe a causas psicológicas, derivadas de la estigmatización social causada por los kilos de más que hoy en día afecta desde niños hasta adultos, que como efecto generan transtornos alimentarios que con el tiempo pueden ser nocivos. Además actualmente se han descrito múltiples factores de riesgo que se asocian a peores desenlaces y a muerte por COVID-19, siendo la Obesidad una de estas condiciones, por lo que las personas con obesidad en todo el mundo ya tienen un alto riesgo de complicaciones graves de COVID-19, en virtud del mayor riesgo de enfermedades crónicas que impulsa la obesidad.

Por todo lo mencionado anteriormente es importante conocer nuestro IMC y sobretodo nuestro porcentaje de grasa corporal (% GC). A pesar de que las diferentes escalas de sobrepeso y obesidad son determinadas por medio del IMC, sin duda el porcentaje de grasa es el componente más importante a la hora de determinar estos estándares ya que el IMC no es una medición directa de la gordura y que se calcula con base en el peso de la persona, lo cual incluye tanto músculo como grasa por lo cual no funciona adecuadamente con personas con una gran cantidad de masa magra, como deportistas o culturistas quienes tienen un índice alto en masa corporal pero no muy alto en grasa corporal, Por esta razón es importante conocer nuestro porcentaje de grasa corporal.

La medición precisa de la grasa corporal es un procedimiento difícil y caro de realizar en la práctica clínica. Se han utilizado diferentes metodologías para medir la grasa corporal. Entre los métodos utilizados se encuentran la medición de los pliegues subcutáneos en distintos puntos (bicipital, tricipital, subescapular y suprailiaco), cuya suma se considera un indicador de la grasa subcutánea. Sin embargo,

[1]jubonillas@unal.edu.co
[2]lfguantivav@unal.edu.co
[3]afninov@unal.edu

es un método de alta variabilidad interobservador y difícil de realizar en pacientes obesos con pliegues cutáneos muy grandes. El Dr. A. Garth Fisher realizó un trabajo de medición para 252 hombres generó una base de datos de diferentes variables en donde enumera las estimaciones del porcentaje de grasa corporal determinado por el pesaje bajo el agua y varias medidas de circunferencia corporal. Concedio el permiso para distribuir libremente los datos y utilizarlos con fines no comerciales.

2. Análisis Descriptivo

Para realizar el análisis de este proyecto contamos con una base de 252 registros sin datos faltantes en las variables, junto con 15 columnas en total (Density, BodyFat, Age, Weight, Height, Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, Wrist). Donde la variable respuesta será "$BodyFat$".

- En la variable respuesta, se obtiene que el mínimo es de 0.0 (Que se le asigna a personas sin porcentaje de grasa), el máximo de 47.5 y una media de 19.0368 para la variable "$BodyFat$" en los 252 individuos.
- Hay un total de 252 hombres a los cuales se les tomaron 15 medidas, pero en nuestro proyecto utilizaremos 14 de estas 15 medidas, excluyendo "$Density$" = Densidad determinada por pesaje bajo el agua"por ser una medida costosa y que debe ser tomada por un especialista.
- "$BodyFat$" = Porcentaje de grasa corporal de la ecuación de Siri (1956)
- "Age"= Años (de edad)
- "$Weight$"= Peso (libras)
- "$Height$" = Altura (pulgadas)
- "$Neck$" = Circunferencia del cuello (cm)
- "$Chest$" = Circunferencia del pecho (cm)
- "$Abdomen$" = Circunferencia del abdomen (cm)
- "Hip" = Circunferencia de la cadera (cm)
- "$Thigh$" = Circunferencia del muslo (cm)
- "$Knee$" = Circunferencia de la rodilla (cm)
- "$Ankle$" = Circunferencia del tobillo (cm)
- "$Biceps$" = Circunferencia del bíceps (extendido) (cm)
- "$Forearm$" = Circunferencia del antebrazo (cm)
- "$Wrist$" = Circunferencia de muñeca (cm)

En este orden de ideas nuestras variables explicativas son (Age, Weight, Height, Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, Wrist) todas continuas.

- En la realización de un modelo de regresión lineal múltiple es importante para el estadístico analizar la relación que existe entre las variables, con el fin de identificar las mejores variables regresoras para el modelo. Adémas de observar una aproximación de la distribución de las variables y posible identificación a primer vista de datos atipicos mediante histogramas como se muestra a continuación.

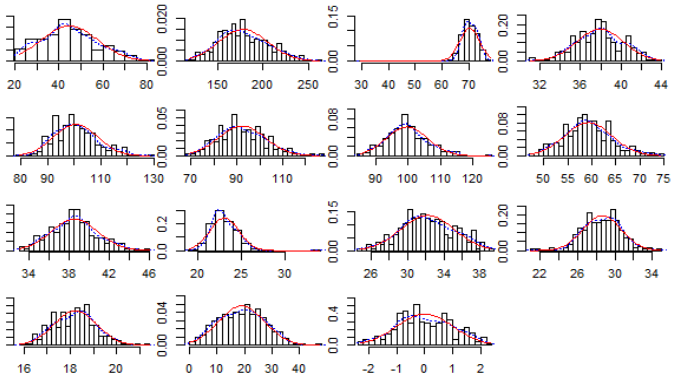

Figura 1

Analizando la correlación con el siguiente gráfico, se tienen los valores de correlación y la distribución de las variables. Asi, se observa la relación que existe entre las variables para poder identificar cuales pueden ser nuestras mejores variables predictoras en el modelo, qué variables no presentan relaciones lineales(para no incluirlas) e identificar colinialidad entre predictoras.

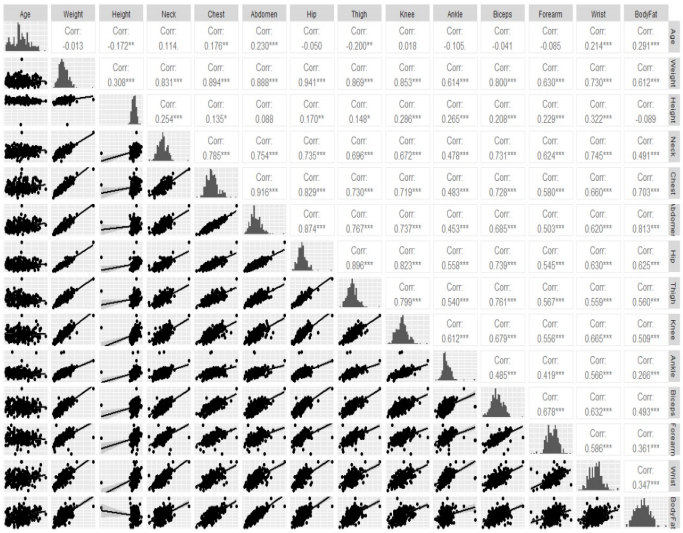

Figura 2

Es de notar que el grafico evidencia que la variables que mas se relacionan con "BodyFat"son, Abdomen con 0.82, Chest con 0.70, Hip con 0.63 y

Weight con 0.62.
De lo anteior es importante resaltar que Weight tiene una alta correlación con las demás variables lo que posiblemente nos genere algún tipo de porblema, dado que puede estar recogiendo el efecto de las demás variables.

3. Planteamiento del problema

Dado el conjunto de datos, se plantea encontrar el mejor modelo de regresión que estime con mayor presición el *Porcentaje de Grasa Corporal* (BodyFat), por medio de medidas corporales de una persona, tales como el peso, la estatura y medidas de perimetro de distintas partes del cuerpo. Además se busca que la estimación del *Porcentaje de Grasa Corporal* se explique empleando el menor número de predictores o variables regresoras.

3.0.1. Hipótesis

El modelo de regresión lineal múltiple se sigue como:

$$\mathbf{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots \beta_{12} \mathrm{X}_{12} + \beta_{13} \mathrm{X}_{13} + \epsilon$$

Bajos los supuestos a priori:
i. $E\left(e_i\right) = 0$ para todo i.
ii. $E\left(e_i^2\right) = \sigma^2$ para todo i (Homocedasticidad).
iii. $E\left(e_i e_j\right) = 0$ para todo i,j (correlación).
iv. $e_i \overset{\text{iid}}{\sim}$ N(0,σ^2).
v. Cov$(e_i x_{ij}) = 0$ para todo i,j
vi. X_i y x_j son linealmente independientes.

Donde la variable dependiente Y representa el porcentaje de grasa muscular, y las variables regresoras representan : Age (X_1), Weight (X_2), Height (X_3), Neck (X_4), Chest (X_5), Abdomen (X_6), Hip (X_7), Thigh (X_8), Knee (X_9), Ankle (X_{10}), Biceps (X_{11}), Forearm (X_{12}), Wrist (X_{13}). Con el modelo se puede hacer el planteamiento de la prueba de hipótesis de interés como:

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_{12} = \beta_{13} = 0$$

.versus

$$\mathrm{H}_a : \text{Al menos un } \beta_j \neq 0 \text{ con j} = 1, \ldots, 13$$

Con la cual se prueba estadísticamente si las variables regresoras son relevantes para el modelo.

4. Aplicación

4.0.1. Selección del modelo

En la selección del modelo se utiliza RLM con los métodos de selección de variables (Forward,Backward y Stepwise). Para la realización de esto método se genera un modelo de regresión sin variables explicativas "$RLM.vacio$" y uno con todas las variables explicativas "$RLM.completo$", con la ayuda del software R y utilizando el modelo $RLM - STEPWISE$ para escoger el mejor modelo obtuvimos:

$$\begin{aligned}\text{BodyFat} = \beta_0 &+ \text{Abdomen} + \text{Weight} + \text{Wrist} + \text{Forearm} \\ &+ \text{Neck} + \text{Age} + \text{Thigh} + \text{Hip} + \text{error}\end{aligned}$$

	Estimate	Std.Error	t value	P(> |t|)
Intercept	-22.656	11.714	-1.93	0.054
Abdomen	0.945	0.072	13.13	0.000
Weight	-0.089	0.040	-2.25	0.025
Wrist	-1.537	0.509	-3.02	0.003
Forearm	0.516	0.186	2.77	0.006
Neck	-0.467	0.225	-2.08	0.039
Age	0.066	0.031	2.14	0.034
Thigh	0.302	0.129	2.34	0.020
Hip	-0.195	0.138	-1.41	0.159

```
Residual standard error: 4.282 on 243
    degrees of freedom
Multiple R-squared:  0.7466,
Adjusted R-squared:  0.7382
F-statistic: 89.47 on 8 and 243 DF,
p-value: < 2.2e-16
```

Despues de la ejecución del paso *STEPWISE*, se obtiene que la variable Hip tiene un Pr(> |t|) igual a 0.1594 por tanto el modelo $RLM - STEPWISE$ tiene una variable no significativa, por lo cual se plantea el siguiente modelo omitiendo variable no significativa.

$$\begin{aligned}\text{BodyFat} = \beta_0 &+ \text{Abdomen} + \text{Weight} + \text{Wrist} + \text{Forearm} \\ &+ \text{Neck} + \text{Age} + \text{Thigh} + \text{error}\end{aligned}$$

Con la ejecución del modelo lineal multiple se obtiene los resultados

```
Residual standard error: 4.291 on 244
    degrees of freedom
Multiple R-squared:  0.7445,
Adjusted R-squared:  0.7371
F-statistic: 101.6 on 7 and 244 DF,
p-value: < 2.2e-16
```

Note que R considera que las variables 'Thigh' y 'Neck' son significativas porque utilizan la estadística de

	Estimate	Std.Error	t value	Pr(> |t|)
Intercept	-33.258	9.007	-3.69	0.0003*
Abdomen	0.918	0.069	13.21	0.0000*
Weight	-0.119	0.034	-3.51	0.0005*
Age	0.068	0.031	2.21	0.0278*
Wrist	-1.532	0.510	-3.00	0.0030*
Thigh	0.222	0.116	1.91	0.0569
Forearm	0.553	0.185	2.99	0.0030*
Neck	-0.404	0.220	-1.83	0.0684*

prueba AIC, pero la estadística de prueba F-statistic no la considera significativa Para términos del ejercicio nos guiaremos por la prueba de R es decir, la AIC. Posteriormente se hace una identificación de posibles valores atípicos o influyentes.

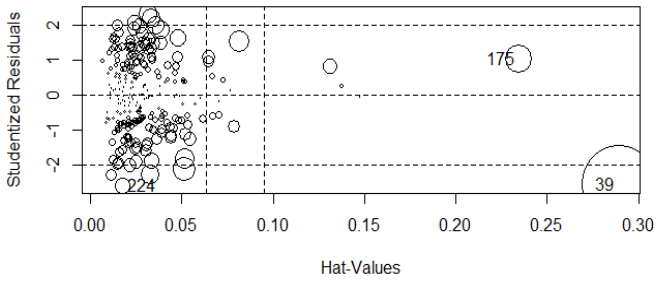

Figura 3

Se encuentra que la observacion número 39 es influyente, por lo que se opta por evaluar nuevamente el modelo sin la observación y determinar si esta influye en el modelo. Siguiendo lo lineamientos anteriores se utiliza RLM: Métodos de selección de variables (Forward, Backward y Stepwise) por tanto para el nuevo modelo $RLMSTEPWISE$ tenemos:

$$\text{BodyFat} = \beta_0 + \text{Abdomen} + \text{Weight} + \text{Wrist} + \text{Biceps} + \text{Age} + \text{Thigh} + \text{error}$$

	Estimate	Std.Error	t value	Pr(> |t|)
Intercept	-32.462	8.348	-3.89	0.0001*
Abdomen	0.882	0.069	12.67	0.0000*
Weight	-0.106	0.035	-3.08	0.0023*
Wrist	-1.799	0.481	-3.74	0.0002*
Biceps	0.238	0.156	1.53	0.1271
Age	0.064	0.030	2.09	0.0380*
Thigh	0.190	0.119	1.60	0.1100

```
Residual standard error: 4.255 on 244
    degrees of freedom
Multiple R-squared:  0.7449,
Adjusted R-squared:  0.7386
F-statistic: 118.8 on 6 and 244 DF,
p-value: < 2.2e-16
```

Despues de la ejecución del segundo paso *STEPWISE*, es de notar que las variables Biceps y Thigh tienen un valor Pr(> |t|) igual a 0.1271 y 0.1100 respectivamente por tanto el modelo $RLM - STEPWISE$ tiene dos variables no significativas, en este orden de ideas procedemos a plantear el siguiente modelo sin estas variables.

$$\text{BodyFat} = \beta_0 + \text{Abdomen} + \text{Weight} + \text{Wrist} + \text{Age} + \text{error}$$

Con la ejecución del modelo lineal multiple se obtiene los resultados

	Estimate	Std.Error	t value	Pr(> |t|)
Intercept	-22.710	7.34	-3.09	0.0022*
Abdomen	0.901	0.068	13.14	0.0000*
Weight	-0.064	0.031	-2.09	0.0373*
Wrist	-1.748	0.482	-3.63	0.0003*
Age	0.042	0.028	1.49	0.1386

```
Residual standard error: 4.295 on 246
    degrees of freedom
Multiple R-squared:  0.738,
Adjusted R-squared:  0.7338
F-statistic: 173.3 on 4 and 246 DF,
p-value: < 2.2e-16
```

Note que la variable Age tiene un valor Pr(> |t|) igual a 0.1386 por tanto el modelo $RLM - STEPWISE$ tiene una variable no significativa, en este orden de ideas procedemos a plantear el siguiente modelo sin esta variable.

$$\text{BodyFat} = \beta_0 + \text{Abdomen} + \text{Weight} + \text{Wrist} + \text{error}$$

Con la ejecución del modelo lineal múltiple obtenemos los siguientes resultados:

	Estimate	Std. Error	t value	Pr(> |t|)
Intercept	-27.275	6.684	-4.08	0.0001*
Abdomen	0.962	0.055	17.43	0.0000*
Weight	-0.092	0.024	-3.83	0.0002*
Wrist	-1.425	0.431	-3.31	0.0011*

```
Residual standard error: 4.305 on 247
    degrees of freedom
Multiple R-squared:  0.7357,
Adjusted R-squared:  0.7325
F-statistic: 229.2 on 3 and 247 DF,
p-value: < 2.2e-16
```

Posteriormente se hace una identificación de posibles valores atípicos o influyente.

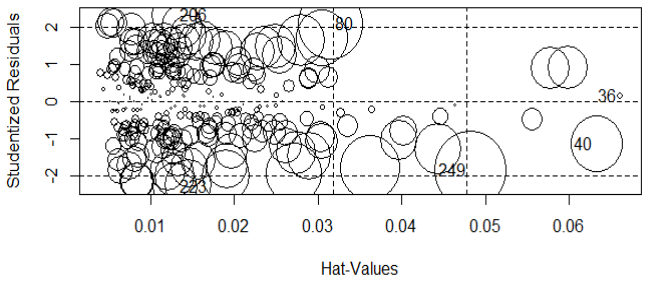

Figura 4

Se encuentra que la observacion número 40 que a pesar de que 36 esta más lejana es más influyente , por lo que se opta por evaluar nuevamente el modelo sin la observación y determinar si esta influye en el modelo. Al eliminar la observación 40 se obtiene:

```
Residual standard error: 4.302 on 246
    degrees of freedom
Multiple R-squared:  0.7334,
Adjusted R-squared:  0.7302
F-statistic: 225.6 on 3 and 246 DF,
p-value: < 2.2e-16
```

Posteriormente se hace una identificación de posibles valores atípicos o influyentes.

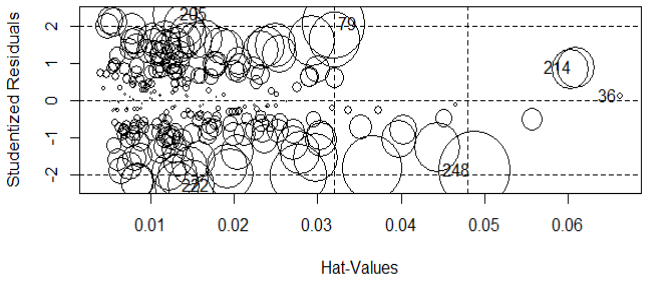

Figura 5

Al observar los resultados, identificamos que las observaciones 36, 214 y 248 son valores influyentes, a diferencia de los anteriores pasos, omitir estos valores y comparar su influencia en el modelo se encuentra que disminuye considerablemente el R^2_{ajus} y afectar pruebas de supuestos como correlación, lo cual nos sugiere que estas observaciones necesarias para el modelo, asi que se recurre a darles un peso a estas observaciones y dejarlas en el modelo.

Al cambiarle su peso a la mitad podemos observar que el error estandar del residual disminuye de 4.302 a 4.284, además su $R^2 ajustado$ pasó de 0.7302 a 0.7246.

```
Residual standard error: 4.284 on 246
    degrees of freedom
Multiple R-squared:  0.7246,  Adjusted R-
    squared:  0.7213
F-statistic: 215.8 on 3 and 246 DF,  p-
    value: < 2.2e-16
```

Posteriormente se hace de nuevo una identificación de posibles valores atípicos o influyentes. Se encuentra que los siguientes pasos estan guiados por el criterio anterior, donde los valores influyentes son necesarios para el modelo y solo se les corregira su peso. A continuacion se muestran las gráficas y sus respectivos resultados.

Evaluación observaciones 250 y 203

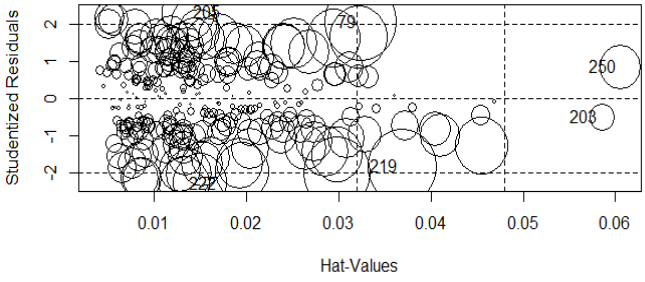

Figura 6

```
Residual standard error: 4.28 on 246
    degrees of freedom
Multiple R-squared:  0.7216,  Adjusted R-
    squared:  0.7182
F-statistic: 212.5 on 3 and 246 DF,  p-
    value: < 2.2e-16
```

Al cambiarle su peso a la mitad podemos observar que el error estandar del residual queda igual al anterior con un valor de 4.28, además su $R^2 ajustado$ pasó de 0.7246. a 0.7182.

Evaluación observaciones 12 y 176

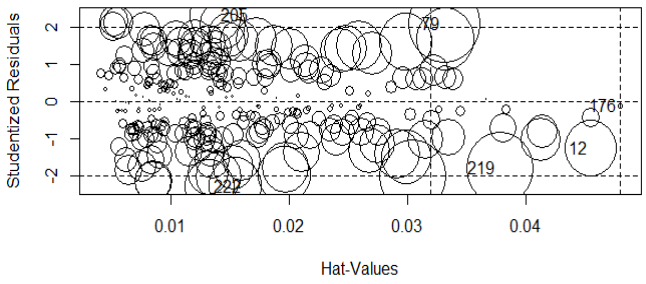

Figura 7

Al cambiarle su peso a la mitad podemos observar que el error estandar del residual cambia y disminuye de 4.284 a 4.274, además su $R^2 ajustado$ pasó de 0.7182 a 0.7169 .

```
Residual standard error: 4.274 on 246
    degrees of freedom
Multiple R-squared:  0.7203,  Adjusted R-
    squared:  0.7169
F-statistic: 211.2 on 3 and 246 DF,  p-
    value: < 2.2e-16
```

Evaluación observaciones 9 y 150

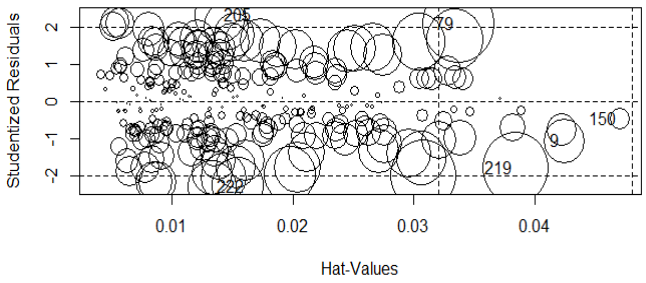

Figura 8

```
Residual standard error: 4.268 on 246
    degrees of freedom
Multiple R-squared:  0.7192,  Adjusted R-
    squared:  0.7157
F-statistic:   210 on 3 and 246 DF,  p-
    value: < 2.2e-16
```

Al cambiarle su peso a la mitad podemos observar que el error estandar del residual cambia y disminuye de 4.274 a 4.268, además su $R^2 ajustado$ pasó de 0.7182 a 0.7157. **Evaluación observacion 246**

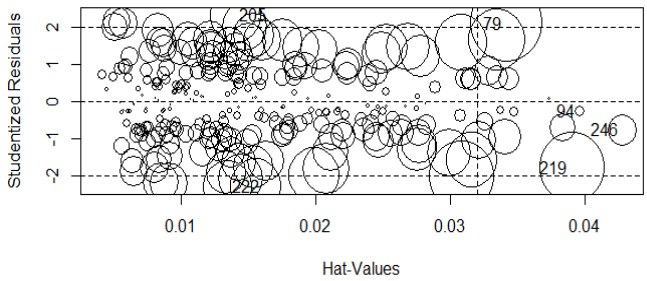

Figura 9

Al cambiarle su peso a la mitad podemos observar que el error estandar del residual cambia y disminuye de 4.268 a 4.265, además su $R^2 ajustado$ pasó de 0.7157 a 0.7155.

```
 Residual standard error: 4.265 on 246
     degrees of freedom
Multiple R-squared:  0.719, Adjusted R-
    squared:  0.7155
F-statistic: 209.8 on 3 and 246 DF,  p-
    value: < 2.2e-16
```

Finalmente

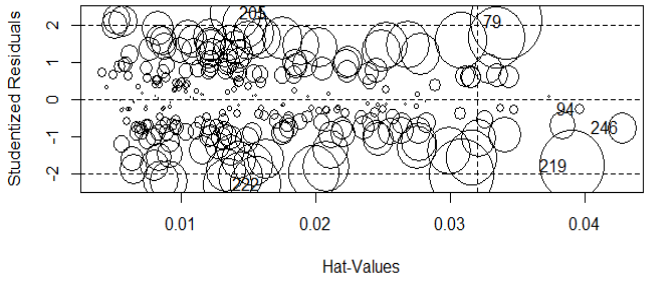

Figura 10

Finalmente observamos que ya no hay datos influyentes. Llegando al siguiente modelo de tres variables:

$$\text{BodyFat} = \text{Abdomen} + \text{Weight} + \text{Wrist} + \text{error}$$

5. Validación de Supuestos y Condiciones del Modelo

5.0.1. Prueba de Normalidad

Se realizan los test de normalidad de los residuales para el modelo escogido, arrojando los siguientes resultados:

```
Shapiro-Wilk normality test

data:  modelo8$residuals
W = 0.9889, p-value = 0.05167

Lilliefors (Kolmogorov-Smirnov) normality
     test

data:  modelo8$residuals
D = 0.052351, p-value = 0.09392
```

Tanto Shapiro-Wilk y Lilliefors nos arrojan un P-Valor mayor a un $\alpha = 0{,}05$, por lo tanto se acepta la normalidad en los datos.
Tambien se muestra gráficamente la distribución normal de los residuos mediante el Q-Q plot.

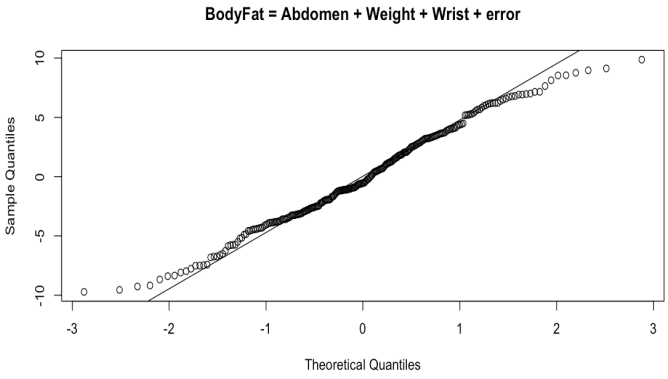

Figura 11

5.0.2. Prueba de Homocedasticidad

Se realiza un test de homocedasticidad arrojando los resultados:

```
studentized Breusch-Pagan test

data:  modelo8
BP = 3.8167, df = 3, p-value = 0.282
```

Dado que el test nos da un P-Valor =0.282 y este es mayor a un $\alpha = 0,05$ no se encuentra evidencia de homocedasticidad en los datos del modelo planteado.
De igual forma se representan en la siguiente figura los residuos del modelo frente a los valores ajustados de este mismo, donde no se observa algún patron en especifico solo dispersión aleatoria de los residuos alrededor de cero.

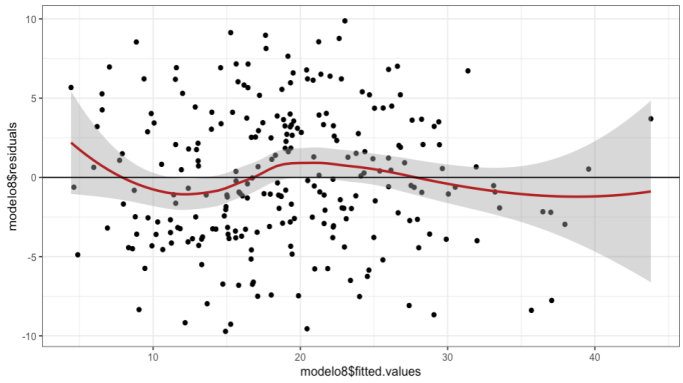

Figura 12

5.0.3. Autocorrelación

Se realiza un test de autocorrelación

```
slag Autocorrelation D-WStatistic p-value
  1       0.1070679       1.779889   0.076
 Alternative hypothesis: rho != 0
```

El Test de Durbin-Watson nos arroja un P-Valor = 0.076, por lo tanto para un $\alpha = 0{,}05$ el P-valor es mayor lo que sugiere que no hay autocorrelación, aunque cabe destacar que el P-valor esta cercano a α por lo que es prudente analizar la siguiente grafica de correlacion entre las variables.

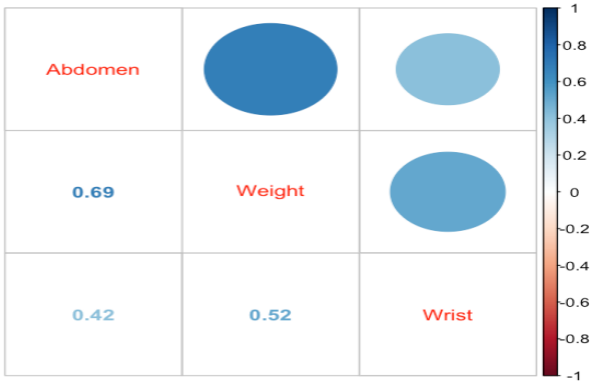

Figura 13

Se observa que se tiene una correlación que puede ser alta entre Abdomen y Weight, lo que haria que pueda presentarse colinealidad.

5.0.4. Analisis Múlticolinealidad

Para encontrar si existe múlticolinealidad se utiliza el Factor de Inflación de Varianza, al evaluarlo en el modelo planteado se obtiene los siguientes resultados:

```
> vif(modelo8)
Abdomen    Weight     Wrist
4.379681 5.852569 2.106778
```

Podemos ver que tanto como Abdomen y Wrist estan en el rango de indice bajo (1 a 5), pero Weight puede presentar problemas de multicolinealidad, dado que tambien se habia observado que existia una correlación relativamente alta de Weight con las otras dos variables, es aconsejable tener en cuenta que esto puede deberse que a que estas variables estan explicando información redundante, de la misma forma cuando se realizo el modelo estas variables siempre fueron significaticas y al omitir una de ellas el modelo perdia explicación importante de la variabilidad.

5.0.5. Datos atípicos

Con el modelo planteado se obtuvieron valores atípicos, lo cual es normal por la variabilidad de los sujetos de estudio, que son observaciones importantes a tener en cuenta pues reflejan casos especificos, como el de un sujeto que presenta indice de masa corporal cero, lo cual suena ilógico pero es un resultado valido según la ecuacíon SIRI. Los outliers se observan en la siguiente gráfica.

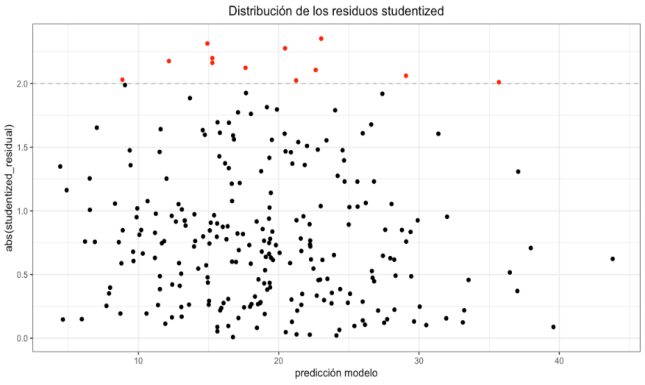

Figura 14

Se identifican los datos atípicos los cuales estan a una distancia de más de dos desvaciones estandar

```
[1]  60  79  84 126 133 138 169 202 205 2
   22 223 236
```

5.0.6. Observaciones influyentes

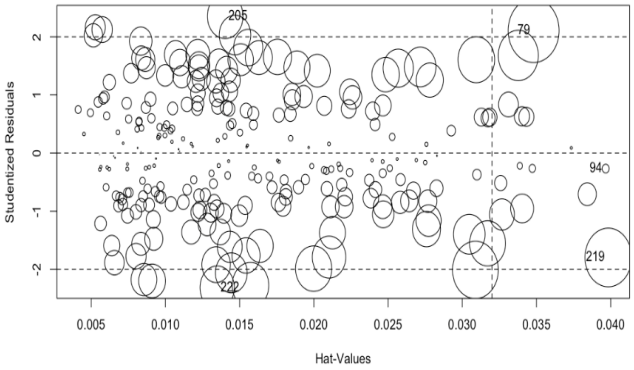

Figura 15

La anterior gráfica muestra las observaciones y el nivel de influencia que puede tener en el modelo, una forma de determinar si un dato es influyente se da si el valor HAT es mayor a la siguiente expresion, donde k es la cantidad de regresores y n es el número de observaciones :

$$2{,}5\left(\frac{k+1}{n}\right) = 0{,}04$$

por lo tanto, con los posibles datos influyentes arrojados mediante software, ninguna observación del valor Hat excede 0.04

```
        StudRes         Hat         CookD
79    2.1063097 0.03480630 0.0394459879
94   -0.2673624 0.03966303 0.0007408754
205   2.3525459 0.01402579 0.0193261674
219  -1.7960616 0.03983874 0.0331613363
222  -2.3138329 0.01348546 0.0179782409
>
```

6. Comparación de modelos

Al realizar los distintos procedimientos en busca de obtener el modelo de regresión lineal múltiple más óptimo, se obtuvieron varios modelos, comparando el

- **Modelo Inicial**

  $$\text{BodyFat} = \beta_0 + \text{Abdomen} + \text{Weight} + \text{Wrist} + \text{Forearm} + \text{Neck}$$
  $$. + \text{Age} + \text{Thigh} + \text{Hip} + \text{error}$$

  Obtenido en la primer evaluación dada por software a través del método RML el cual tiene los siguientes resultados:

```
Residual standard error: 4.282 on 243
     degrees of freedom
```

```
Multiple R-squared:  0.7466,
    Adjusted R-squared:  0.7382
F-statistic: 89.47 on 8 and 243 DF,
    p-value: < 2.2e-16
```

Se resalta el valor $R^2_{ajus} = 0{,}7382$, es decir, las medidas Abdomen, Weight, Wrist, Forearm, Neck, Age, Thigh, Hip explican un 73.82 % de la variabilidad de BodyFat.

Para el **Modelo inicial** se tiene la siguiente tabla ANOVA.

$$ANOVA \quad INCIAL$$

	Df	Sum Sq	Mean Sq	Fvalue	Pr(>F)
Abdomen	1	11631.5	11631.5	634.4	0.00 ***
Weight	1	1004.2	1004.2	54.77	0.00***
Wrist	1	157.2	157.2	8.57	0.004**
Forearm	1	127.8	127.8	6.9	0.009**
Neck	1	51.1	51.1	2.8	0.096.
Age	1	47.9	47.9	2.6	0.107
Thigh	1	67.4	67.4	3.7	0.056.
Hip	1	36.5	36.5	2.0	0.159
Residuals	243	4455.3	18.3		

Note que para este modelo, el **Modelo inicial** se cuentan con 8 variables explicativas de las cuales cuatro de ellas no son significativas según los P-Valores del ANOVA, ademas se cuenta con 252 observaciones con el mismo peso.

- **Modelo final**

$$\text{BodyFat} = \beta_p + \text{Abdomen} + \text{Weight} + \text{Wrist} + \text{error}$$

Obtenido despues de realizar eliminación de variables no significaticas a través del método RML, eliminación y asignación de pesos a observaciones influyentes, por medio de varias iteraciones para cuadrar el mejor modelo. Donde se obtienen los siguientes resultados:

```
Residual standard error: 4.265 on 246
    degrees of freedom
Multiple R-squared:  0.719, Adjusted
    R-squared:  0.7155
F-statistic: 209.8 on 3 and 246 DF,
    p-value: < 2.2e-16
```

Se resalta el valor $R^2_{ajus} = 0{,}7155$, es decir, las medidas Abdomen, Weight, Wrist, explican un 71.55 % de la variabilidad de BodyFat.

Para el **Modelo Final** se tiene la siguiente tabla ANOVA.

$$ANOVA \quad FINAL$$

	Df	Sum Sq	Mean Sq	Fvalue	Pr(>F)
Abdomen	1	10617.3	10617.3	583.6	0.00***
Weight	1	647.3	647.3	35.6	0.00***
Wrist	1	184.1	184.1	10.1	0.001**
Residuals	246	4475.4	18.2		

Note que para es modelo se cuentan con 3 variables explicativas de las cuales todas son significativas con 250 observaciones, cabe resaltar que 10 de estas tienen la mitad del peso. **comparación**.

Dado los dos modelos, es claro ver que hay una disminución en el $R^2_{ajustado}$ de 0.7382 a 0.7155 además del Error estándar residual que dismunuye de 4.282 a 4.265 sin emabargo conservan el mismo valor $P = 2{,}2e - 16$.
Finalmente se concluye que el modelo final es mas optimo por que cuenta con menos variables explicativas y el cambio entre los $R^2_{ajustados}$ y los Errores estándar residuales no son relevantes, es decir, ambos modelos difieren solo en un 2 % de la explicación de la variabilidad, pero el cambio se ve reflejado en la cantidad de variables regresoras que difieren, se pasa del Modelo Inicial 8 variables regresoras a un Modelo Final con tres variables regresoras.

Por otra parte se destaca que en todas las iteraciones desde que se partio del modelo inicial hasta el final, los modelos cumplian los supuestos de normalidad, homocedasticidad, autocorrelación. Estos resultados se pueden evidenciar en el codigo de **R** anexo a este documento.
Otro resultado importante encontrado se tiene con el *Factor de Inflación de Varianza*, puesto que el resultado del VIF del modelo inicial tiene problemas de múlticolinealidad graves en dos variables y moderados en otras dos, como se muestra a continuación:

```
  Abdomen    Weight     Wrist
     Forearm      Neck       Age
     Thigh
 8.236808 18.829990  3.096051  1.9403
    09  4.081562  2.059194  6.283117
     Hip
13.471431
```

Ahora, el *Factor de Inflación de Varianza* del modelo final solo cuenta con un problema moderado de multicolinealidad en la regresora Weight.

```
 Abdomen    Weight     Wrist
4.379681 5.852569 2.106778
```

Finalmente, este último resultado nos verifica que el Modelo Final es más eficiente que el Modelo

Inicial. Por lo tanto la estimación del Porcentaje de Masa Corporal más optima esta dada por el Modelo de Regresión Lineal Múltiple :

$$\begin{aligned}\text{BodyFat} = &-27{,}72004 + 0{,}95876(\text{Abdomen}) \\ &- 0{,}08868(\text{Weight}) - 1{,}41963(\text{Wrist}) \\ &+ \text{error}\end{aligned}$$

7. Conclusiones

Se evidencia que el metodo de seleccion RLM que usa el criterio de estadistica AIC es optimo, sin embargo en casos particulares como el de este modelo vemos que que puede incluir variables no significativas, por tanto es trabajo del estadistico determinar que si es conveniete eliminar esta variable no significativa para plantear un modelo mas optimo.

Otro punto a resaltar son los valores influyentes el analisis de estos fueron fundamentales para el modelo ya que estos pueden sesgar el modelo además de algunos supuestos, como se evidenció en la parte de aplicación donde los valores influyentes hacian que ciertas variables fueran significativas.En este orden de ideas al eliminar dos variables y reducir a la mitad el peso de 10 variables se plantea un modelo sin observaciones influyentes que puedan generar este tipo de sesgos.

Modelo inicial

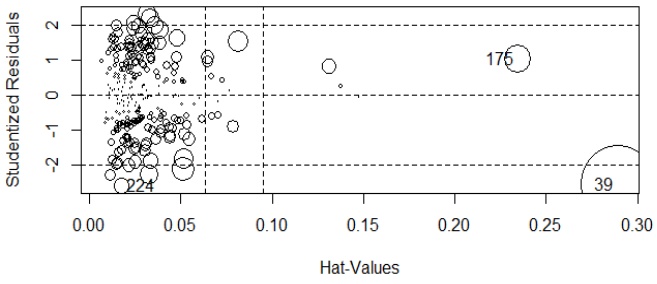

Figura 16

Para este caso se evidencian datos influyentes como el 39 que por su area se puede concluir que tiene bastante peso.

Modelo final

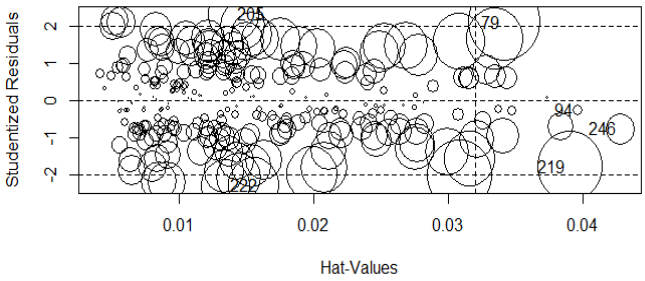

Figura 17

Ya en este modelo no hay datos influyentes que puedan sesgar el modelo.

.

Referencias

[1] Roger W. Johnson. *Body Fat prediction datasets. (s. f.)*. Recuperado de https://www.kaggle.com/fedesoriano/body-fat-prediction-dataset

[2] Melo O. (2021) "Análisis de Regresión [Notas de Clase y Código R]" *Universidad Nacional de Colombia sede Bogotá*, Bogota D.C.

[3] Melo O. (2020) "Modelos lineales [Notas de Clase y Código R]" *Universidad Nacional de Colombia sede Bogotá*, Bogota D.C.

[4] National geographic(2019) *National geographic: El exceso de grasa corporal incrementa el riesgo de padecer depresión* Recuperado de https://www.nationalgeographic.com.es/ciencia/exceso-grasa-corporal-incrementa-riesgo-padecer-depresion

[5] OMG (2019) *Obesidad. (s.f)*. .ºbesidad y sobrepeso". Recuperado de: https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight

[6] Rico. E (2020) "Obesidad y Covid-19" *Obesidad y Covid-19*. Recuperado de: https://www.archivosdemedicina.com/medicina-de-familia/obesidad-y-covid19.pdf

[7] NATIONAL INSTITUTES OF HEALTH (1998) "Obesidad y Covid-19" *Obesidad y Covid-19*. Recuperado de: https://www.nhlbi.nih.gov/files/docs/guidelines/obgdlns.pdf

[8] *Indice de masa corporal y porcentaje de grasa en adultos indigenas. (2017)*. Oleas M. Recuperado de: http://ve.scielo.org/scielo.php?script=sci_arttextpid=S0004-06222017000100006