Fundamentos de Análisis de Datos
Trabajo Práctico 2

Antes de comenzar a realizar modelos estadísticos se debe primero hacer un análisis exploratorio de los datos con el objetivo de asegurarse la presencia y consistencia de los valores de las variables. En este caso el dataset consta de 507 observaciones y 24 variables, todas ellas son del tipo numeric y no existen valores faltantes, como se muestra en la Fig 1.

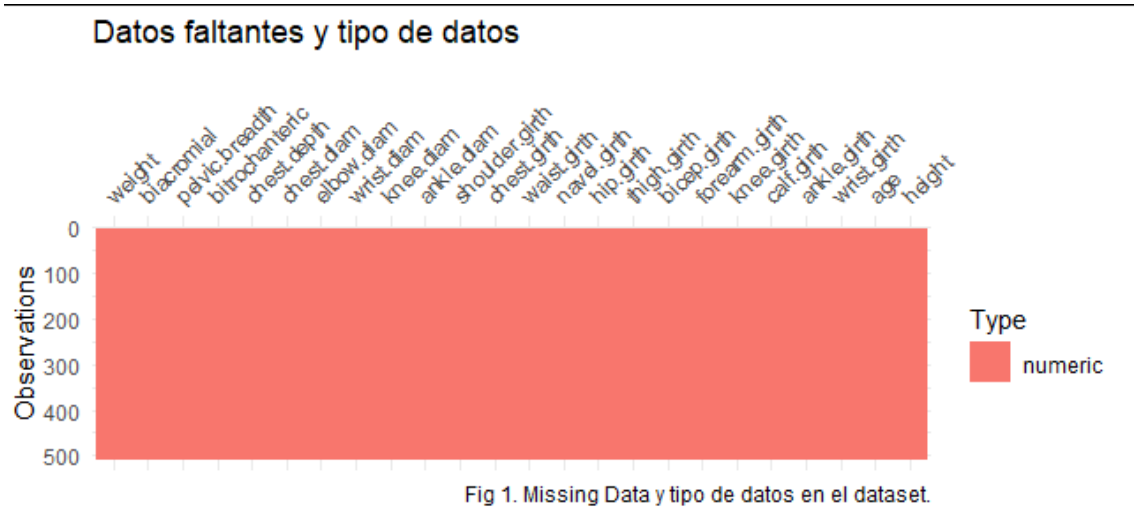

Fig 1. Missing Data y tipo de datos en el dataset.

Sin embargo, es necesario también saber la distribución de los mismos ya que la presencia de un valor no asegura la consistencia de los mismos, es decir podríamos tener valor negativo en la variable edad y no ser conscientes de ello. Entonces, para asegurarnos la consistencia de los datos debemos realizar un boxplot para cada variable, y analizar los valores atípicos de las mismas. La Fig 2 muestra la distribución de cada variable, y si contiene o no valores atípicos.

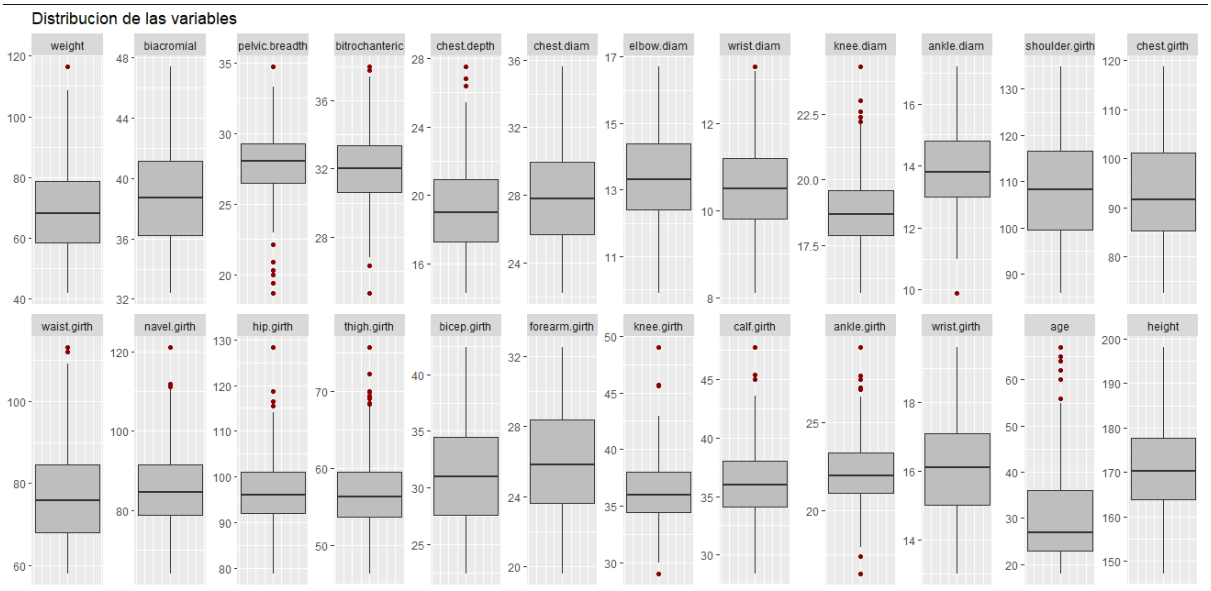

Fig 2. Boxplot de las variables

Dada la dimensionalidad del dataset, para saber si una observación es atípica en relación a las demás, no sólo con respecto a valores de esa misma variable; se necesita contar con un método para medir qué tan distante se ubica la observación de la media de la distribución. Para eso

utilizaremos la distancia de Mahalanobis, que luego analizando su distribución podremos clasificar las observaciones en valores atípicos, la Fig 3 muestra un boxplot con la distribución correspondiente.

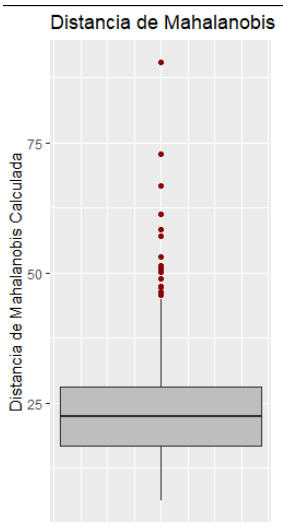

Fig 3. Boxplot Mahalanobis

La cantidad de valores atípicos según la distribución de distancia de mahalanobis asciende a 19, según la Fig 3, cabe aclarar que el criterio utilizado para considerar a un valor como outlier es si este valor es mayor o menor del rango IQR * 1.5 desde el cuartil 25 o 75 según corresponda.
Dada la baja cantidad de outliers en términos relativos con la muestra se decide removerlos de la misma, y así mejorar la capacidad predictiva de los modelos.

Para construir los modelos, realizar predicciones y evaluar los resultados se dividirá la muestra aleatoriamente en dos subgrupos. El primero, el grupo de entrenamiento, consta del 80% de los datos de la muestra y se utilizará para calcular los parámetros de los modelos. Luego el segundo, el grupo de testeo, compuesto por el 20% de los datos restantes, y será utilizado para realizar las predicciones y evaluar los resultados.

Antes de construir cualquier modelo es necesario entender la teoría detrás de los mismos. En este caso, todas las predicciones se realizan utilizando un modelo de regresión lineal múltiple. Este modelo define una variable dependiente o explicada, en función de varias variables independientes o explicativas.
De forma matricial se tiene:

$$Y = X * \beta + \varepsilon$$

Donde Y representa un vector de nx1 observaciones, X es una matriz de nxk, β es un vector de dimensión kx1 y luego el vector ε nx1 residuos. Los residuos o errores tienen la propiedad de que el valor esperado de E[ε] = 0, dada esta propiedad podemos afirmar que los cambios en la variable dependiente se corresponden con cambios en las variables independientes, es decir no son aleatorios.
Sabiendo esto, se tiene que el valor esperado de Y es:

$$E[Y] = X * \beta$$

El objetivo del modelo es encontrar un $\hat{\beta}$ tal que minimice el error o residuos:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

Para calcular los estimadores de β se parte de:

$$X * \widehat{\beta} = Y$$

Dado que la matriz X es de n x (k+1), siendo k la cantidad de variables explicativas, no es invertible, se puede realizar una transformación de la misma y demostrar que:

$$X^t * X * \widehat{\beta} = X^t Y$$
$$\widehat{\beta} = (X^t X)^{-1} * X^t * Y$$

Para poder afirmar que el vector $\widehat{\beta}$ es el mejor estimador lineal insesgado se deben realizar los siguientes supuestos,

1. $E[\varepsilon_i] = 0, \ \forall i. \ Equivale\ a\ decir\ E[Y_i] = \beta_o + \beta_1 x_1 + \ldots + \beta_k x_k$
2. $var(\varepsilon_i) = \sigma^2, \ \forall i. \ \ Equivale\ a\ decir\ var(Y_i) = \sigma^2$
3. $cov(\varepsilon_i, \varepsilon_j) = 0, \ \forall\ i \neq j. \ Equivale\ a\ decir\ cov(Y_i, Y_j) = 0$
4. $\varepsilon \sim N(0, \sigma^2)$

El supuesto 1 establece que Yi depende únicamente de las variables X de i hasta n. El supuesto 2 establece que la varianza del error o de la variable dependiente es constante y no depende de X. El supuesto 3 establece que los errores no están correlacionados entre sí. Y por último, el supuesto 4 establece que los residuos siguen una distribución normal con media 0 y varianza σ^2.

Una vez analizado el modelo general de regresión lineal y su supuestos correspondientes, es posible realizar un modelo de regresión lineal múltiple utilizando las variables de nuestro dataset.
Se desea estudiar como la variable weight depende de las otras variables. Por ende, weight será la variable dependiente y las demás serán explicativas o independientes.
Para ello construimos un modelo lineal utilizando la función lm del paquete stats, y suministrando para la construcción el dataset correspondiente para el entrenamiento.
En el coefficient plot de la Fig 4, se observan los valores de los Betas estimados para cada variable, sumado a si son significativos al 5%.

A su vez, el comando summary sobre el modelo arroja:

Residual standard error: 2.149 on 366 degrees of freedom
Multiple R-squared: 0.9755, Adjusted R-squared: 0.9739
F-statistic: 632.3 on 23 and 366 DF, p-value: < 2.2e-16

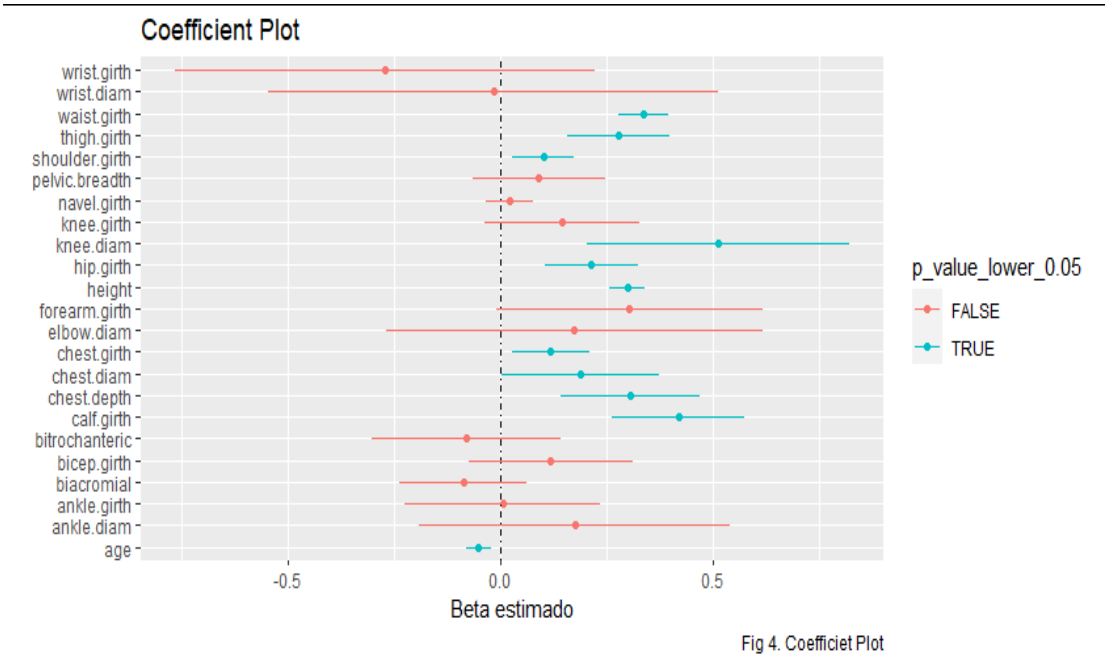

Fig 4. Coefficiet Plot

En principio el modelo parece muy bueno, con un R^2 excelente, un error estándar de los residuos bajo y un estadístico F con un valor extremadamente chico indicando que los valores de los Betas son distintos de 0.

Pero para poder utilizar este modelo es necesario validar el mismo, es decir verificar si se cumplen los supuestos planteados anteriormente.

La Fig 5 se utiliza para validar el modelo, y está compuesta por dos gráficos en los que se plotean los residuos y las predicciones; para el caso del titulado Residuals vs Fitted, los residuos no están estandarizados; y para el caso de Scale - Location, los residuos están estandarizados. Además, se puede observar que un QQ plot para verificar la normalidad de los residuos.

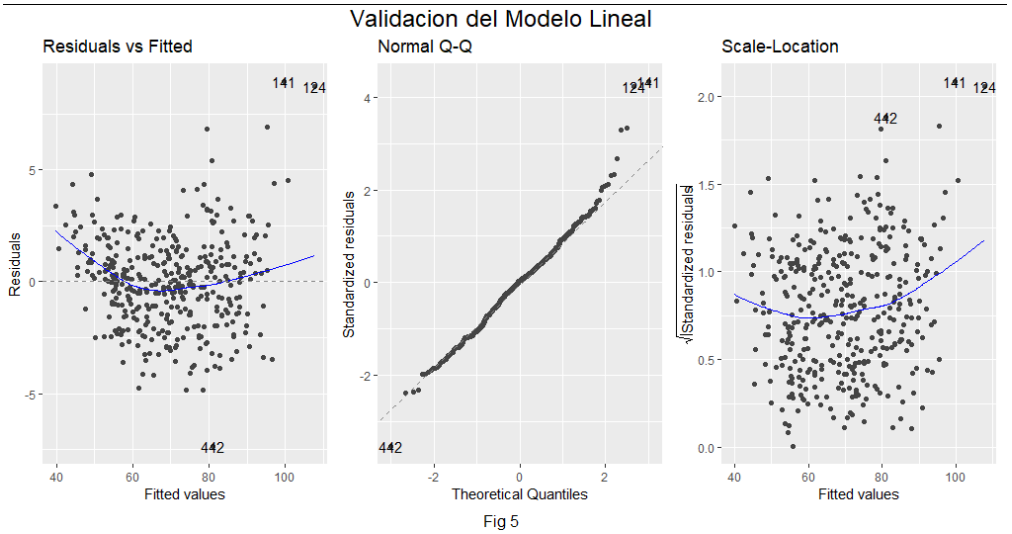

Fig 5

Un primer análisis de la varianza de los residuos muestra una distribución poco uniforme, tanto en los residuos estandarizados como en los residuos simples. Para poder tener certeza si el supuesto de homocedasticidad (varianza constante) se cumple, y por ende saber si los Betas son los mejores estimadores lineales insesgados para modelar esta regresión, podemos realizar un test de

homocedasticidad conocido como *Breusch-Pagan Test,* es un test de hipótesis con hipótesis nula de varianza constante. Como todo test de hipótesis si el p valor es chico se rechaza la hipótesis nula.
La función ols_test_breusch_pagan del paquete olsrr nos permite realizar el test. Este arroja como resultado un p valor pequeño, por ende se rechaza la hipótesis nula y existe evidencia para decir que la muestra posee heteroscedasticidad, es decir la varianza no es uniforme.

```
Breusch Pagan Test for Heteroskedasticity
-----------------------------------------
Ho: the variance is constant
Ha: the variance is not constant

            Data
-----------------------------------
Response : weight
Variables: fitted values of weight

            Test Summary
-------------------------------
DF            =        1
Chi2          =        35.07023
Prob > Chi2   =        3.180253e-09
```

La solución propuesta es aplicar logaritmo a la variable explicada (weight), crear nuevamente el modelo de regresión lineal múltiple y constatar si los residuos siguen una distribución normal, con varianza homogénea. En la Fig 6 se muestra nuevamente los gráficos para la validación del modelo, en este caso se observa que la distribución de los residuos es más uniforme, tanto los residuos simples como los estandarizados. Además, vemos que el QQ plot ajusta de buena manera a la nueva distribución de residuos. Para corroborar la presencia de homocedasticidad se realiza el Breusch-Pagan test, obteniendo el siguiente resultado:

```
Breusch Pagan Test for Heteroskedasticity
-----------------------------------------
Ho: the variance is constant
Ha: the variance is not constant

            Data
---------------------------------------
Response : log(weight)
Variables: fitted values of log(weight)

            Test Summary
----------------------------
DF            =        1
Chi2          =        0.1377462
Prob > Chi2   =        0.7105319
```

El mismo test de hipótesis muestra un p valor mucho más grande que el anterior, por ende no es posible rechazar la hipótesis nula y se verifica ahora si homocedasticidad. En este caso se puede afirmar que todos los supuestos del modelo de regresión se cumplen, y los Betas son los mejores estimadores lineales insesgados.

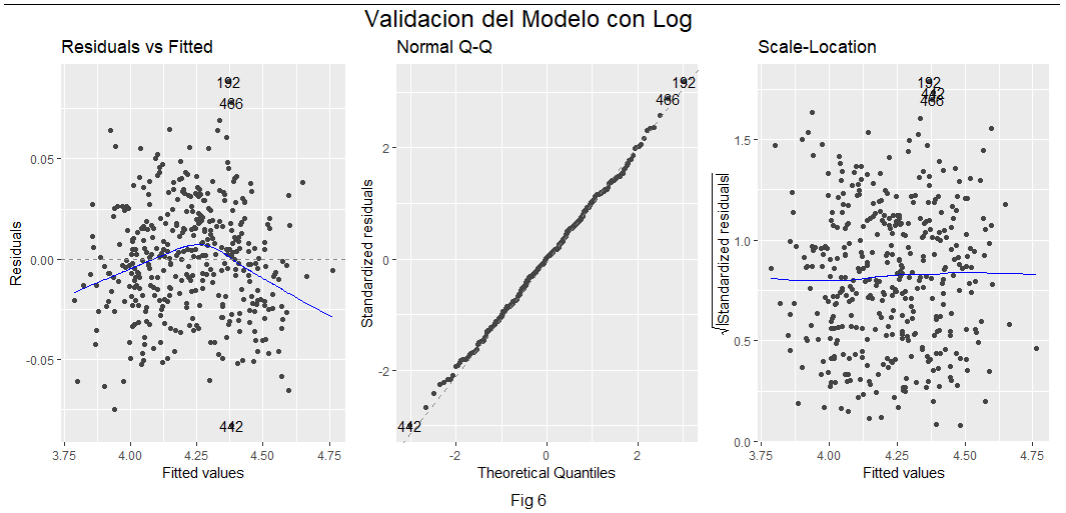

Fig 6

A posterior en la Fig 7 vemos un coefficient plot de los betas estimados, coloreados según si tienen un p valor menor a 0.05.

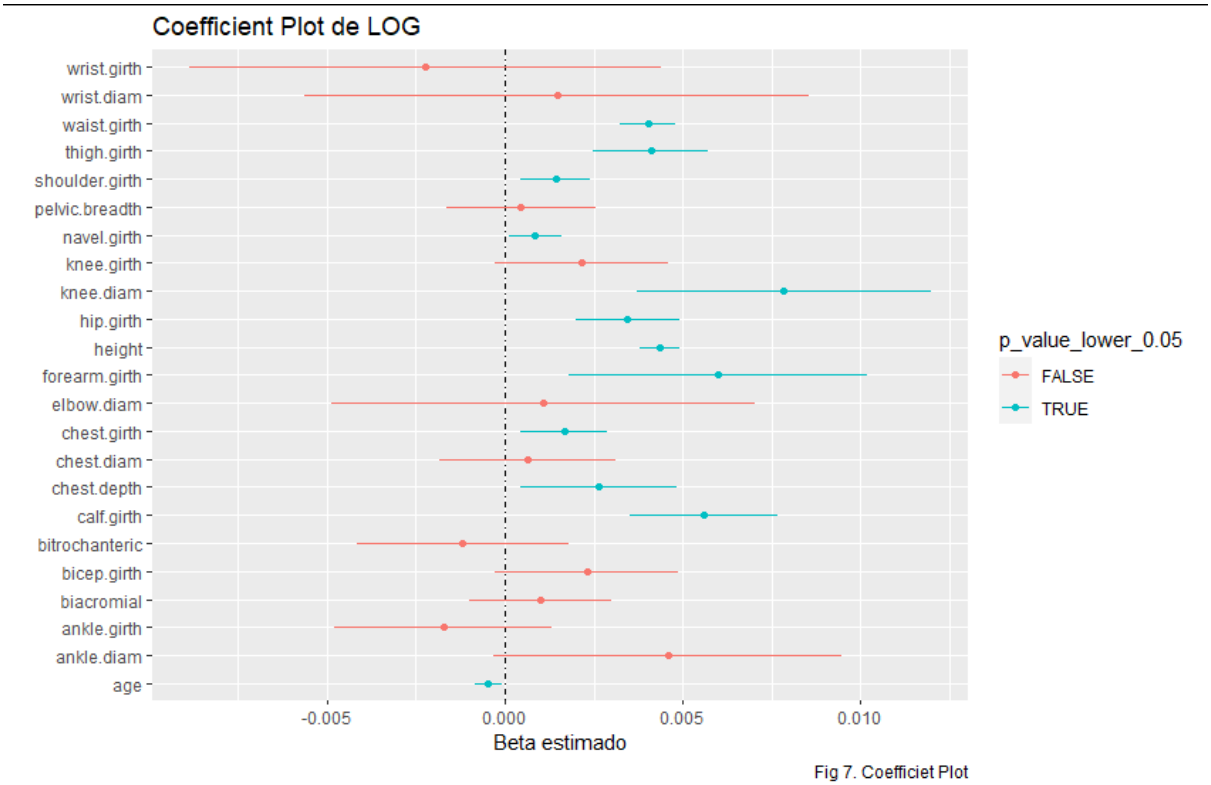

Fig 7. Coefficiet Plot

Ya validado el modelo podemos utilizar el comando summary en el nuevo modelo con la variable explicada transformada, y así obtenemos:

Residual standard error: 0.02871 on 366 degrees of freedom

Multiple R-squared: 0.9789, Adjusted R-squared: 0.9776

F-statistic: 738.9 on 23 and 366 DF, p-value: < 2.2e-16

Este nuevo modelo tiene una buena capacidad predictiva, pero siempre teniendo en cuenta que se utiliza el dataset de entrenamiento. Luego todos los modelos desarrollados en el trabajo serán evaluados con el dataset de prueba, y de esta manera poder sacar conclusiones.

Algo importante a tener en cuenta a la hora de realizar inferencias sobre los estimadores es tener en cuenta la correlación entre las variables. Hasta ahora se utilizaron todas las variables del dataset para construir el modelo. La Fig 8 muestra un corrplot de todas las variables, en este gráfico observamos que algunas variables están correlacionadas positivamente.

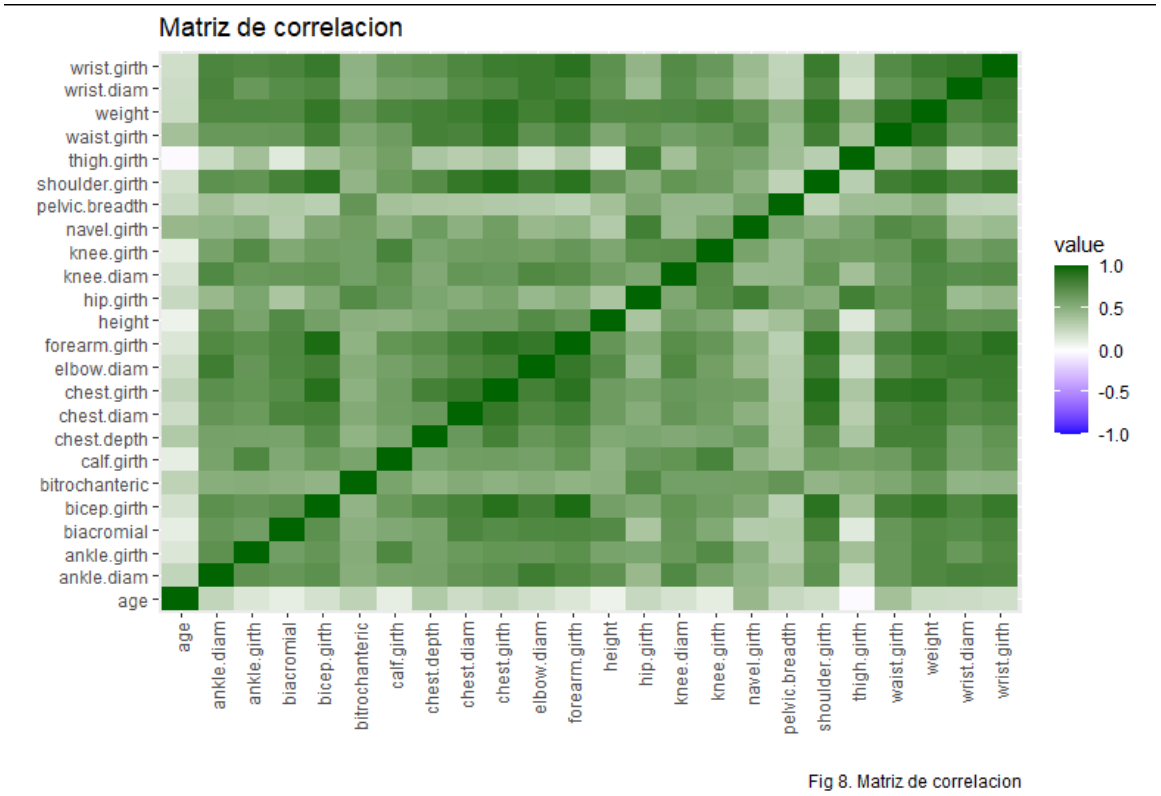

Fig 8. Matriz de correlacion

Volviendo al plot de coeficientes de la Fig 7 se observa que muchas variables tienen un p valor elevado, sumado al análisis de correlación de la Fig 8, nos indica que la interpretación de los resultados obtenidos por el modelo podrían ser cuanto menos difíciles de interpretar, o hasta llevar a sacar conclusiones erróneas.
Para mejorar la interpretación de los resultados es necesario seleccionar las mejores variables independientes a utilizar. Esto se puede lograr mediante la función de selección de variables ols_step_both_p del paquete olsrr, la cual crea un modelo de regresión incorporando o quitando potenciales variables predictoras basados en el p value correspondiente. Esto se realiza de a una variable a la vez, y finaliza cuando no quedan variables por remover o agregar.
En la Fig 9 observamos las variables con las que el modelo fue creado, sumado a que todas son significativas.

Utilizando el comando summary obtenemos la siguiente información del modelo :

Residual standard error: 0.02875 on 377 degrees of freedom

Multiple R-squared: 0.9782, Adjusted R-squared: 0.9775

F-statistic: 1412 on 12 and 377 DF, p-value: < 2.2e-16

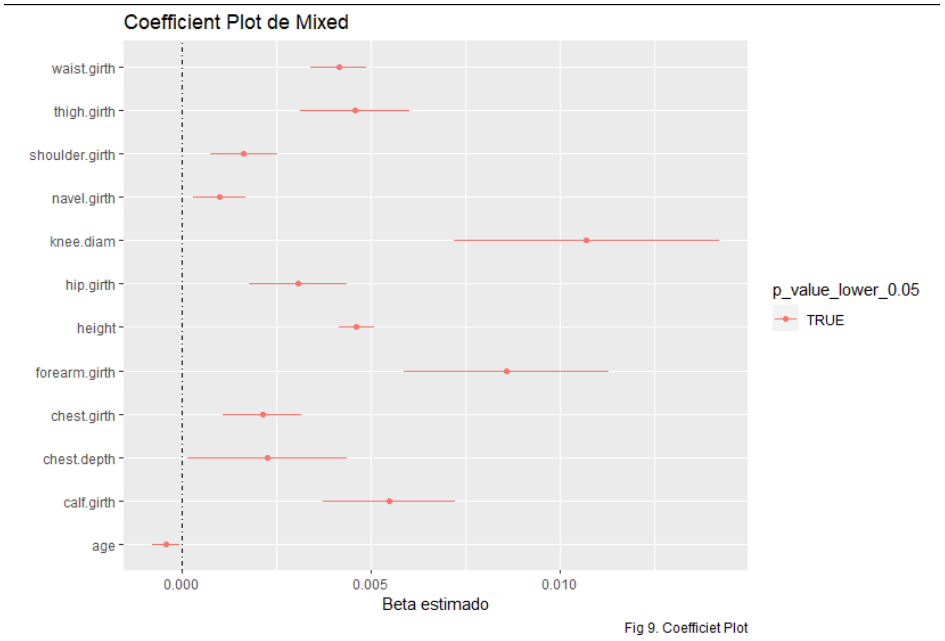

Fig 9. Coefficiet Plot

El modelo creado con menos variables tiene la misma capacidad de predicción en términos de R2, pero esta vez los resultados pueden tener una mejor interpretación.
Cabe destacar que los gráficos de validación del modelo que se observan en la Fig 10 son similares a los del modelo anterior dado que este modelo utiliza un subconjunto de las mismas variables. Al igual que en el modelo anterior los residuos tienen una distribución más homogénea, y los residuos estandarizados se ajustan correctamente en el QQ plot.

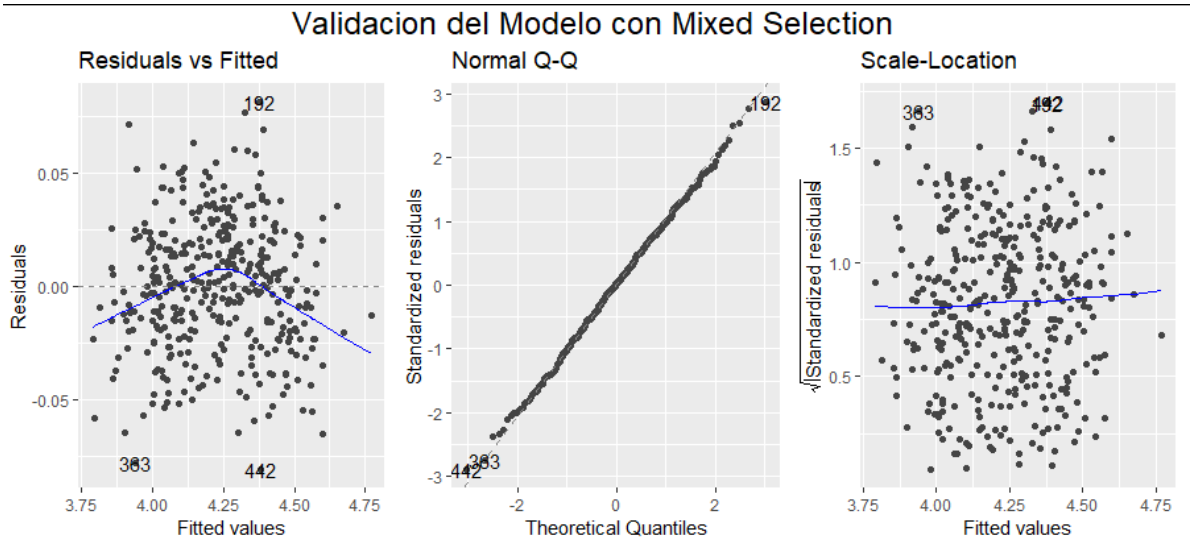

Fig 10

Por último, es interesante plantear si se puede crear un modelo de regresión lineal a partir de componentes principales de las variables explicativas del dataset.
Teniendo nuestro conjunto de variables explicativas correlacionadas entre sí, es posible calcular las componentes principales, siendo:

$$C * V = V * D$$

La matriz C de correlación de las variables explicativas, V los autovectores de esa matriz y D los autovalores asociados.
Y teniendo

$$T = B * V$$

Siendo T la matriz de componentes principales, B el conjunto de variables explicativas.
Entonces nuestro modelo de regresión lineal quedaría de la siguiente manera:

$$Y = T * \hat{\beta}_{PC}$$

Luego,

$$\hat{\beta}_{PC} = (T^t T)^{-1} * T^t * Y$$

Para realizar esto, una opción es utilizar la función pcr del paquete pls. Esta función calcula todas las componentes principales para las variables explicativas, y luego realiza un modelo de regresión lineal múltiple entre la variable explicada sin transformar y las componentes principales calculadas. La cantidad de componentes principales a utilizar viene determinada por el dataset y la cantidad de información con la que se quiere conservar.
Una vez planteado el modelo de regresión lineal es necesario validar el mismo de igual manera que validamos los anteriores. En la Fig Y observamos los residuos estandarizados vs fitted values pudiendo observar una distribución bastante homogénea de la varianza. En el segundo subplot se observa un QQ plot que muestra un buen ajuste de los residuos estandarizados a la distribución normal.

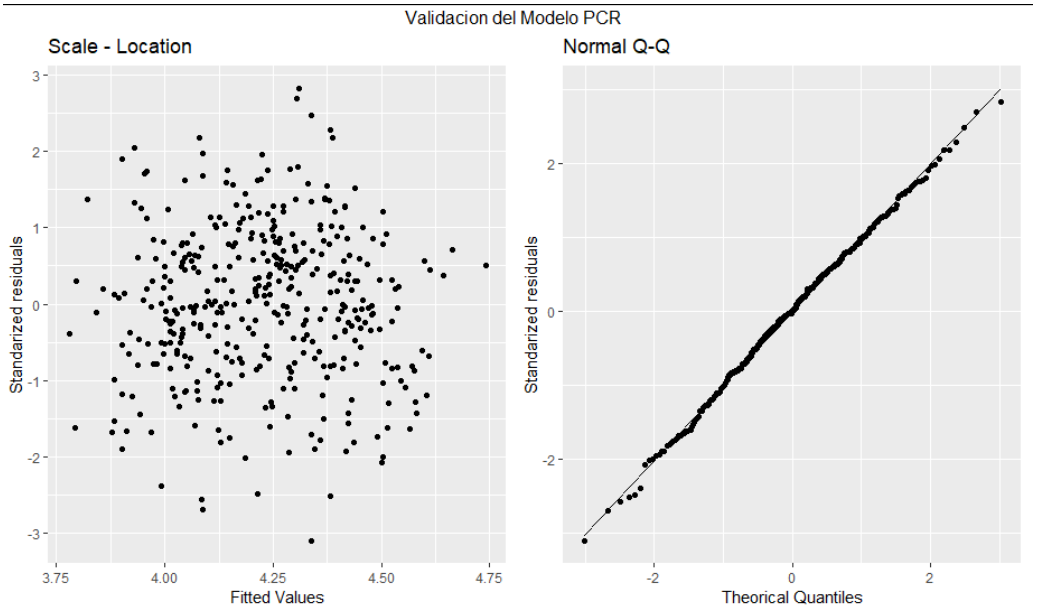

Fig 11

Ya construido el modelo y obtenidos las componentes principales, se debe definir la cantidad a utilizar. En base a la Fig 12, observamos que a partir de la séptima componente principal el porcentaje de varianza explicada es muy bajo. La suma acumulada de la varianza explicada para la séptima componente es 90%, dado esto utilizaremos las primeras siete componentes para nuestro modelo.

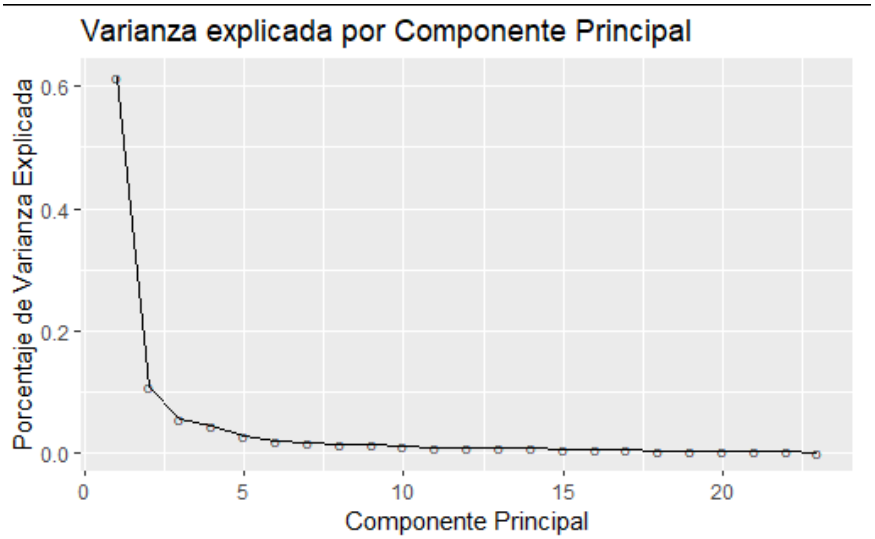

Fig 12. Screeplot

Una vez construido los tres modelos, el primero utilizando todas las variables como explicativas, el segundo utilizando el método de mixed selection y el último de componentes principales, es necesario comparar la capacidad predictiva de los modelos utilizando información nunca antes vista por los mismos. En este momento se utiliza el dataset de testeo que contenía el 20% restante de los datos. La métrica utilizada para comparar los modelos es RMSE y además se agrega esta métrica utilizando el dataset de entrenamiento. Los valores obtenidos se observan en la Tabla 1

Comparacion de RMSE

	Total_Var	Mixed_Sel	PCR
Train	0.0278	0.0283	0.0351
Test	0.029	0.0292	0.0362

Tabla 1

Si bien medir el RMSE ayuda con la elección del modelo a utilizar, también es necesario tener en cuenta que cada modelo puede llevar a una interpretación distinta en relación a las variables explicativas. Por ejemplo, el modelo de componentes principales está construido con una combinación lineal de todas las variables explicativas, de los β poco podría extraerse de cada variable explicativa en particular pero para datasets de gran dimensionalidad podría ser la elección correcta, aun teniendo una capacidad de predicción un poco menor.

Por último, en la Fig 13 observamos las predicciones contra los valores reales del dataset de testeo para los tres modelos. Los resultados mostrados en la Tabla 1 son consistentes con esta figura, siendo el modelo de regresión lineal utilizando componentes principales el que tiene mayor error.

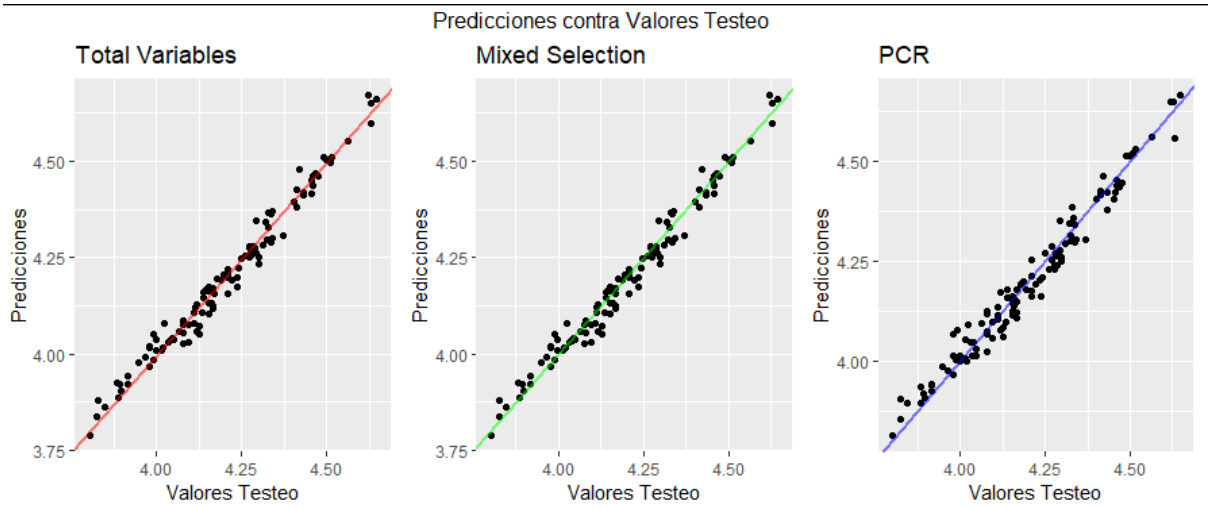
Predicciones contra Valores Testeo
Total Variables
Mixed Selection
PCR
Predicciones
Valores Testeo
4.50
4.25
4.00
3.75

Fig 13