1. Selección de variables

Considere la base de datos fat del paquete faraway, considere todas las variables, excepto siri, density y free. También eliminé del análisis los casos con valores extraños en weight y height, así como valores cero en **brozek**. Suponga que el objetivo del estudio es usar las variables clinicas observadas en los pacientes para estudiar cuáles de éstas son los factores que ayudan a modelar mejor el promedio del porcentaje de grasa corporal en los hombres (var **brozek**).

DESARROLLO:

Los datos registrados en este estudio son la edad, el peso, la altura y ciertas mediciones de la circunferencia corporal de 250 hombres, donde el porcentaje de grasa corporan fue estimada con presición mediante una técnica de pasaje bajo el agua. En resumén:

- El porcenjate de grasa corporal promedio observado fue de 18.96.
- El estudio comprende personas desde los 22 años hasta los 81 años. La edad media del estudio es de 45 años.
- El peso promedio observado en las personas del estudio es de 179.1 libras.
- La altura promedio observada es de 70.32 pulgadas.

Buscamos saber que variables nos podrían ayudar a modelar mejor el promedio del porcentaje de grasa corporal en los hombres. Para ello se realizarón varios algoritmos con distintas especificaciones como se muestra en la siguiente tabla:

NOTAS:

1. Se utilizarón 4 métodos para la selección de variables, **Mejor Subconjunto**, **Stepwise (Backward)**, **Stepwise (Forward)** y **Método Lasso**.
2. Se consdierarón modelos con distribución **Gaussiana** y liga **identidad**, y modelos **Gamma** con ligas **Identidad** y **Logarítmica**.
3. La columna "**M**"hace referencia a le método que se utilizó para obtener el modelo, donde **1:=** Mejor Subconjunto, **2:=** Stepwise (Backward), **3:=** Stepwise (Forward) y **4:=** Lasso.
4. La columna "**I**"hace referencia a si los modelos consideran interacciones ("SI") o si solo se consideran efectos principales ("NO").
5. La columna "**D**"hace referencia a las modificaciones que se aplicarón a los datos, donde **1:=** No se aplicó ningúna modificación, **2:=** Se adicionó la versión al cuadrado de las variables a los datos originales.

M	I	D	DISTRIBUCIÓN Y LIGA	MODELO	BIC
1	NO	1	Gaussina (Identidad)	**brozek** ∼weight + abdom + forearm + wrist	1432.163
			Gamma (Identidad)	**brozek** ∼height + neck + abdom	1485.957
			Gamma (Logarítmica)	**brozek** ∼weight + abdom	1530.582
		2	Gaussina (Identidad)	**brozek** ∼abdom + $weight^2$ + $wrist^2$	1420.689
2	NO	1	Gaussina (Identidad)	**brozek** ∼weight + abdom + forearm + wrist	1432.163
			Gamma (Identidad)	**brozek** ∼neck + abdom + wrist	1506.148
			Gamma (Logarítmica)	**brozek** ∼age + abdom + hip + thigh + wrist	1535.799
		2	Gaussina (Identidad)	**brozek** ∼abdom + hip + wrist + $adipos^2$ + $chest^2$ + hip^2	1427.45
			Gamma (Identidad)	**brozek** ∼weight + neck + biceps + wrist + age^2 + $weight^2$ + $neck^2$ + $abdom^2$ + $biceps^2$ + $forearm^2$	1500.884
			Gamma (Logarítmica)	**brozek** ∼neck + abdom + biceps + wrist + age^2 + $neck^2$ + $abdom^2$ + $biceps^2$	1508.849
	SI	1	Gaussina (Identidad)	**brozek** ∼age + height + abdom + hip + wrist + age:wrist + abdom:hip	1427.132
			Gamma (Identidad)	**brozek** ∼ age + weight + height + neck + abdom + thigh + knee + ankle + biceps + forearm + wrist + weight:ankle + height:biceps + neck:abdom + neck:thigh + neck:forearm + knee:ankle + ankle:wrist + forearm:wrist	1515.237
			Gamma (Logarítmica)	**brozek** ∼height + adipos + chest + abdom + biceps + wrist + height:biceps + adipos:chest	1525.074
3	SI	2	Gaussina (Identidad)	**brozek** ∼age + weight + height + adipos + neck + chest + abdom + hip + thigh + knee + ankle + biceps + forearm + wrist + age^2 + $weight^2$ + $height^2$ + $adipos^2$ + $neck^2$ + $chest^2$ + $abdom^2$ + hip^2 + $thigh^2$ + $knee^2$ + $ankle^2$ + $biceps^2$ + $forearm^2$ + $wrist^2$ + age:$wrist^2$ + age^2:$abdom^2$	1520.224
			Gamma (Identidad)	**brozek** ∼age + weight + height + adipos + neck + chest + abdom + hip + thigh + knee + ankle + biceps + forearm + wrist + age^2 + $weight^2$ + $height^2$ + $adipos^2$ + $neck^2$ + $chest^2$ + $abdom^2$ + hip^2 + $thigh^2$ + $knee^2$ + $ankle^2$ + $biceps^2$ + $forearm^2$ + $wrist^2$ + biceps:wrist + chest:$abdom^2$ + thigh:$thigh^2$ + abdom:$abdom^2$ + ankle:$ankle^2$ + biceps:$wrist^2$	1562.624
			Gamma (Logarítmica)	**brozek** ∼age + weight + height + adipos + neck + chest + abdom + hip + thigh + knee + ankle + biceps + forearm + wrist + age^2 + $weight^2$ + $height^2$ + $adipos^2$ + $neck^2$ + $chest^2$ + $abdom^2$ + hip^2 + $thigh^2$ + $knee^2$ + $ankle^2$ + $biceps^2$ + $forearm^2$ + $wrist^2$ + thigh:$thigh^2$	1596.693

M	I	D	DISTRIBUCIÓN Y LIGA	MODELO	BIC
4	NO	1	Gaussina (Identidad)	**brozek** ∼ age + weight + height + adipos + neck + chest + hip + thigh ankle + biceps + forearm + wrist	1468.525
			Gamma (Identidad)	**brozek** ∼abdom	1476.451
			Gamma (Logarítmica)	**brozek** ∼abdom	1551.975
		2	Gaussina (Identidad)	**brozek** ∼age + weight + adipos + neck + abdom + thigh + biceps + wrist + age^2 + $weight^2$ + $neck^2$ + $chest^2$ + hip^2 + $ankle^2$ + $forearm^2$	1463.738
			Gamma (Identidad)	**brozek** ∼abdom	1525.976
			Gamma (Logarítmica)	**brozek** ∼abdom	1551.975
	SI	1	Gaussina (Identidad)	**brozek** ∼ age + weight + height + neck + abdom + hip + thigh + knee + biceps + forearm + wrist + age:weight + age:height + age:adipos + age:chest + age:abdom + age:hip + age:thigh + age:ankle + age:biceps + age:forearm + age:wrist + weight:adipos + weight:hip + weight:thigh + weight:knee + weight:biceps + weight:forearm + weight:wrist + height:adipos + height:neck + height:thigh + height:ankle + height:forearm + adipos:neck + adipos:chest + adipos:thigh + adipos:ankle + neck:abdom + neck:thigh + neck:ankle + chest:knee + chest:biceps + abdom:ankle + abdom:forearm + abdom:wrist + hip:thigh + thigh:biceps + knee:biceps + ankle:biceps + ankle:wrist + biceps:wrist + forearm:wrist	1614.336
			Gamma (Identidad)	**brozek** ∼ abdom	1525.976
			Gamma (Logarítmica)	**brozek** ∼ abdom	1551.975
		2	Gaussina (Identidad)	**brozek** ∼ abdom + age:$biceps^2$ + age:$forearm^2$ + abdom:forearm + age^2:$ankle^2$ + $weight^2$:$adipos^2$ + $height^2$:$neck^2$ + $height^2$:$wrist^2$ + $neck^2$:$wrist^2$	1439.59
			Gamma (Identidad)	**brozek** ∼ abdom	1476.451
			Gamma (Logarítmica)	**brozek** ∼ abdom	1476.451

En términos generales, en la mayoria de los modelos obtenidos las variables que más se repiten son **weight** (la variables asociada al peso de los hombres) y **abdom** (la variable asociada a la circunferencia del abdomen), esto nos da una primera intuición a cerca de que algunas de las variables que mejor podrían ayudar a modelar el promedio de porcentaje de grasa corporal en los hombres, pueden ser las variables asociadas al peso y a la circunferencia del individuo, y seguramente con algunas otras variables más.

Ahora analicemos brevemente los resultados con los distintos métodos. En cuanto al método de **Mejor Subconjunto**, se puede observar que los modelos obtenidos mediante este método son bastante sencillos, donde la mayoria de ellos incluyen a las variables *weight* y/o *abdom*, los mejores modelos en cuanto a su BIC se obtuvieron considerando una distribución *Gaussiana* con liga *identidad*, y el modelo en el que se cosnsiderarón la versión al cuadrado de las variables mostro un BIC aún más bajo.

Por otro lado, al aplicar el método de **Stepwise (Bakcward)**, si interacciones se obtienen modelo de cireta manera sencillos, en cuanto al número de variables, en este caso se mantiene el patrón de que la mayoria de los modelos contienen a la variable *weight* y/o a la variable *abdom*, además se puede observar que en efecto al adicionar la versión al cuadrado de las variables, se obtienen mejores modelos en cuanto a su BIC. Al aplicar este mismo método con interacciones, se obtienen modelos con un BIC más elevado con respecto a los obtenidos cuando no se consideraron interacciones, salvo para el caso de una distribución Gaussiana con liga Identidad, en donde se observa que el BIC disminuyó al considerar las interacciones entre las covariables. Ahora bien, al consdierar un método **Stepwise (Forward)** para los caso de interacciones con los datos que contienen adicionalmente las versión al cuadrado de las variables, se obtienen modelos muy robustos, en donde debido a la gran cantidad de variables que contienen no resultan optimos para trabajar con ellos, además de que por esta misma razón su BIC es mayor al que se obtuvó considerando interacciones sin considerar la versión al cuadrado de las variables. Por lo tanto en este caso, adicionar las versión al cuadrado de las variables no resulta de mucha utlidad.

Por último, al aplicar el método **Lasso** para el caso de la distribución *Gamma* con ligas *Identidad* y *Logarítmica*, bajo todos los posibles casos siempre se obtuvó el mismo modelo, **brozen** $\sim$ abdom, sin embargo, esto no ocurrió en el caso de la distribución *Gaussiana* con liga *Identidad*, donde el mejor modelo considerando el BIC fue el que se obtuvo considerando interacciones y los datos que contienen las variables originales y la versión al cuadrado de las mismas. Por lo que en este caso, el incluir la versión al cuadrado de las variables parece ser una buena opción.

De manera general, se puede observar que los mejores modelos se obtuvieron con una distribución **Gaussiana** con liga **Identidad**.

Ya que tenemos una idea general de nuestros posibles modelos, comencemos a hacer una selección, para ello vamos a inicialmente basarnos el su respectivo BIC. Veamos cuales son los 5 mejores modelos en cuanto a su BIC:

1. **brozek** $\sim$ abdom + weight^2 + wrist^2

   BIC = 1420.689. Distribución **Gaussiana** con liga **Identidad**.

2. **brozek** $\sim$ age + height + abdom + hip + wrist + age:wrist + abdom:hip

   BIC = 1427.132. Distribución **Gaussiana** con liga **Identidad**.

3. **brozek** $\sim$ abdom + hip + wrist + adipos^2 + chest^2 + hip^2

   BIC = 1427.45. Distribución **Gaussiana** con liga **Identidad**.

4. **brozek** $\sim$ weight + abdom + forearm + wrist

   BIC = 1432.163. Distribución **Gaussiana** con liga **Identidad**.

5. **brozek** $\sim$ abdom + age:biceps^2 + age:forearm^2 + abdom:forearm + age^2:ankle^2 + weight^2:adipos^2 + height^2:neck^2 + height^2:wrist^2 + neck^2:wrist^2

   BIC = 1439.59. Distribución **Gaussiana** con liga **Identidad**.

Como podemos observar los **BIC** asociados a estos modelos no son muy distantes, por lo que podríamos decir que cualquiera de estos modelos podría modelar de manera optima el promedio del porcentaje de grasa corporal en los hombres. Sin embargo, dado que estos modelos están diseñados para la inferencia e interpretación, se deberán deberán cumplir los supuestos de linealidad, homocedasticidad, normalidad y aleatoriedad. Al verificar los supuestos anteriormente mencionadios, se obtuvó que todos los modelos salvo el modelo número 4, **brozek** $\sim$ weight + abdom + forearm + wrist, cumplen todos los supuestos. Por lo tanto, para la elección del mejor modelo vamos a basarnos unicamente en el BIC, de esta manera, de acuerdo a todo el análisis realizado el mejor modelo que ayuda a modelar el promedio del porcentaje de grasa corporal en los hombres es:

$$\textbf{brozek} \sim \text{abdom} + \text{weight}^2 + \text{wrist}^2$$

Esto es:

$$\mathbb{E}\left[\textbf{brozek};\ \text{abdom}, \text{weight}^2, \text{wrist}^2\right] = \beta_0 + \beta_1 \cdot \text{abdom} + \beta_2 \cdot \text{weight}^2 + \beta_3 \cdot \text{wrist}^2$$

Donde,

$$\beta_{\mathbf{1}} = 0.9147 \qquad \beta_{\mathbf{2}} = -0.0002893 \qquad \beta_{\mathbf{3}} = -0.03115$$

De esta manera, tomando un valor fijo de las variables wrist^2 y weight^2, por cada centimetro que aumente la circunferencia del abdomen, en promedio el porcentaje de grasa corporal de los hombres aumentará en 0.9147 unidades. Por otro lado, tomando un valor fijo de las variables abdom y wrist^2, por cada libra adicional al cuadrado del peso de un hombre, en promedio el porcentaje de grasa corporal disminuirá 0.0002893 unidades. Por último, tomando un valor fijo de las variables abdom y weight^2, por cada centimetro que aumente el cuadrado de la circunferencia de la muñeca, en promedio el porcentaje de grasa corporal en hombres disminuira en 0.03115 unidades.