Ejercicio 1: Predicción en el caso continuo

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 clínicas observadas en los pacientes para predecir el porcentaje de grasa corporal en los hombres (var brozek).

Solución:

En la figura 1 se muestran los boxplots correspondientes a los valores que se tienen para las variables *height* y *weight*.

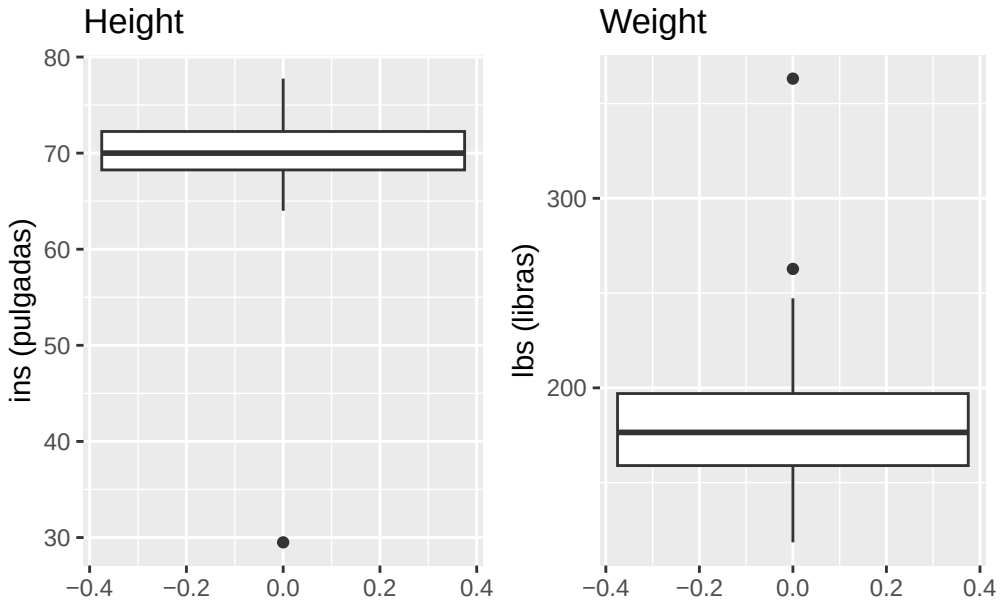

Figura 1: Boxplot de las variables height (izquierda) y weight (derecha)

En estos, se observa la presencia de dos outliers para el caso de *weight*, mientras que para *height* únicamete se presenta un dato "raro", el cual tiene un valor de 29.50 pulgadas. Para evitar sesgos en los modelos de predicción, se optó por eliminar los registros correspondientes a estos datos atípicos.

También se eliminaron observaciones con valores cero en la variable *brozek*.

Una vez realizado el preprocesamiento anterior, se comenzó ajustando un modelo lineal generalizado con liga identidad y distribución Gaussiana. Para esto, se consideraron cuatro modelos: con efectos principales, efectos principales más interacciones de segundo orden, efectos principales más las variables al cuadrado y un últimos más complejo, con efectos principales más interacciones de segundo orden y variables al cuadrado. Para hacer una comparación de los modelos, todos se entrenaron bajo el mismo esquema (K-Cross Validation) con $K = 5$ iteraciones, y se calcularon métricas para analizar los errores de los resultados: MSE, MAE y el coeficiente de correlación al cuadrado R^2.

Después, se utilizaron dos métodos distintos para la selección de variables en los modelos: el método por pasos con dirección "both", y el método de selección lasso. Esto con el objetivo de reducir el número de variables y analizar si con esto se obtienen métricas de error más reducidas. Por tanto para estos nuevos modelos también se calcularon el MSE, MAE y R^2, para poder hacer una comparación directa del poder predictivo.

Finalmente, para abarcar una mayor gama de modelos a comparar, se optó por considerar un modelo lineal generalizado con distribución Gamma, con todas sus ligas posibles, es decir: identidad, inversa y logarítmica. Para estos también se realizó una selección de variables por el método lasso, ya que, fue el método con el que se obtuvieron los resultados con más rapidez.

Finalmente, se presenta a continuación los resultados de los 15 modelos mencionados anteriormente, entrenados bajo el esquema 5-CV.

Cuadro 1: Esquemas de entrenamiento con sus respectivas métricas para comparar el poder predictivo de cada modelo

#	Fórmula	Selección	Variables	MSE	MAE	CORR
1	brozek ~ .	NO	Variables principales	17.69641	3.448992	0.2494696

2	brozek ~ .^2	NO	Variables principales, Interacciones de segundo orden	43.73025	5.072932	0.1281249
3	brozek ~ . + I(variables)^2	NO	Variables principales, Variables al cuadrado	19.63711	3.572491	0.2255304
4	brozek ~ .^2 + I(variables)^2	NO	Variables principales, Interacciones de segundo orden, Variables al cuadrado	114.95717	6.316812	0.0999587
5	brozek ~ .	Step	(Intercept), age, adipos, chest, abdom, wrist	17.99459	3.501602	0.2680490
6	brozek ~ .^2	Step	Variables principales, age:weight, age:knee, age:biceps, age:wrist, weight:adipos, weight:knee, height:neck, height:hip, height:thigh, height:biceps, adipos:thigh, adipos:knee, neck:ankle, neck:forearm, hip:thigh, hip:biceps, thigh:biceps, ankle:wrist, forearm:wrist, age:weight:knee, height:hip:biceps, height:thigh:biceps	31.38045	4.373213	0.1755370
7	brozek ~ . + I(variables)^2	Step	(Intercept), weight, wrist, I(weight^2), I(abdom^2)	17.96944	3.504692	0.2593080
8	brozek ~ .^2 + I(variables)^2	Step	Variables principales, I(height^2), I(adipos^2), I(neck^2), I(chest^2), I(abdom^2), I(hip^2), I(thigh^2), I(ankle^2), I(forearm^2), age:weight, age:height, age:adipos, age:neck, age:chest, age:abdom, age:knee, age:ankle, age:biceps, age:forearm, weight:neck, weight:chest, weight:thigh, weight:biceps, height:adipos, height:neck, height:chest, height:thigh, height:knee, height:biceps, height:forearm, adipos:neck, adipos:chest, adipos:abdom, adipos:hip, adipos:thigh, adipos:knee, adipos:ankle, adipos:forearm, neck:chest, neck:biceps, neck:forearm, neck:wrist, chest:abdom, chest:thigh, chest:knee, chest:ankle, chest:forearm, abdom:hip, abdom:forearm, abdom:wrist, hip:thigh, hip:ankle, thigh:forearm, knee:wrist, ankle:biceps, ankle:wrist, forearm:wrist, biceps:I(height^2), age:I(neck^2), neck:I(neck^2), height:I(adipos^2), age:height:knee, height:chest:thigh, height:neck:biceps, neck:chest:forearm	1160.74049	11.779350	0.0519870
9	brozek ~ .	Lasso	(Intercept), age, height, abdom, wrist	15.49616	3.260415	0.7257507
10	brozek ~ .^2	Lasso	(Intercept), height, abdom, age:adipos, age:abdom, height:wrist	15.24163	3.227453	0.7301582
11	brozek ~ . + I(variables)^2	Lasso	(Intercept), age, abdom, wrist, I(height^2)	15.49077	3.259581	0.7258467
12	brozek ~ .^2 + I(variables)^2	Lasso	(Intercept), abdom, I(height^2), age:adipos, age:abdom, height:wrist	15.24163	3.227453	0.7301582
13	brozek ~ .^2 + I(variables)^2	Lasso	(Intercept), abdom, thigh, biceps, wrist, I(adipos^2), I(chest^2), I(abdom^2), I(hip^2), I(forearm^2), age:adipos, age:abdom, age:ankle, height:neck, height:ankle, height:wrist, adipos:abdom, neck:ankle	14.36807	3.113016	0.7444579
14	brozek ~ .^2 + I(variables)^2	Lasso	(Intercept), abdom, age:adipos	16.17066	3.310683	0.7141214
15	brozek ~ .^2 + I(variables)^2	Lasso	(Intercept), abdom, I(height^2), age:abdom, age:thigh, height:wrist	15.63948	3.201793	0.7211586

Como una descripción de estos resultados, cabe mencionar que todos los modelos son ajustes lineales genralizados con distribución Gaussiana y liga identidad, excepto por los últimos tres modelos: 13, 14 y 15; para los cuales se utilizó la distribución Gamma con ligas inversa, identidad y logarítmica; respectivamente. También se debe mencionar que para los casos en los que se hizo una

selección de variables con el método lasso, se realizó cross-validation (con $k = 5$) para tunear el parámetro lambda (escogiendo el *lambda.min* en todos los casos), y en los que se realizó el método step, se hizo en ambas direcciones.

Notamos que, el incluir todas las variables posibles con distintas transformaciones no necesariamente es de ayuda para mejorar el poder predictivo, por ejemplo, el modelo 8 es el que cuenta con la mayor cantidad de variables pero es el que peor le va en la predicción del porcentaje de grasa corporal, y en cambio, modelos más sencillos como el 9, que cuenta con sólo 4 variables, se encuentra entre los 5 más competitivos.

Las variables con mayor poder predictivo resultan ser las que aparecen con mayor frecuencia en los modelos analizados, en este caso resultan ser: *age*, *height*, *weight*, *wrist*, *abdom* y *adipos*.

Finalmente, el modelo a elegir es con el que se obtienen los errores más pequeños para la predicción en general y este es el número 13, el cual se realizó con distribución Gamma y liga inversa (liga canónica de dicha distribución). Dicho modelo contiene sólo 4 variables principales, algunas variables al cuadrado e interacciones, por lo que se aprecia la utilidad de incluir esas posibilidades en el análisis. Además, este fue el único con el que se obtuvo un error MSE menor a 15, y con el coeficiente de correlación al cudrado podemos ver que esta regla explica casi el 75 % de la variabilidad en los datos, por lo que es un modelo mejorable pero con cierta solidez. Cabe mencionar que el lambda tuneado para este modelo tiene un valor de $\lambda = 0.004052$ (chunk: $lambda_tun$).