1. Predicción en el caso continuo

Considere la base de datos "fat" del paquete faraway, 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).

Mediante el summary se verificó y eliminaron datos que parecían atípicos, veáse en el chunk "datos_raros". Quedandonós así con 247 observaciones y 15 columnas.

Tras revisar la base se procedio a realizar la predicción en sus dos etapas, recopilando la información de estas en la siguiente tabla con las siguientes columnas

REGLA: En todos los casos se usaron MLG para datos continuos, en los 9 primeros se uso una liga identidad y distribución Gaussiana, mintras que en el 10 fue una liga inversa con distribución Gamma, es decir su liga canónica. Los últimos cuatro modelos.

SELECCIÓN: Aquí se especifíca el método o criterio mediante el cuál se hizo la selección de variables para cada una una de las reglas. En los primeros tres no se uso ningún método de selección, en los siguientes tres se utilizó optimización discreta mediante BIC y para los últimos cuatro se seleccionaron mediante lasso, en estos para poder determinar el valor óptimo de lambda, es decir del hiperparámetro, se usó el método de remuestreo K-CV con k=5.

VARIABLES: Las covariables involucradas en cada una de las regresiones.

Para poder determinar el poder predictivo de estas reglas se tomaron como parámetros las siguientes tres métricas, calculadas bajo el método de remuestreo "Repeated Holdout Method" (RHM) con B=50. Cabe mencionar que en los modelos donde se hizo uso de lasso, al tener un hiperparámetro entonces tenemos el uso de remuestreo anidado, una primera vez para el calculo del hiperparámetro óptimo, el cual fue calculado bajo K-CV con k=5 y el segundo, para medir el poder predictivo como el resto de modelos mediante RHM.

MSE: Criterio para el poder predictivo, error cuadrático medio

MAE: Criterio para el poder predictivo, media de la diferencia en valor absoluto

CORR: Criterio para el poder predictivo, coeficiente de correlación al cuadrado entre "y" y "y_gorrito"

Tras revisar los resultados podemos concluir que:

Ls variables que más veces aparecen entre los modelos tanto con selección de variables con BIC y Lasso son los siguientes: "abdom,height,neck,thigh"

De los modelos que se probaron el que tuvo un mejor desempeño en cuanto a poder predictivo es el MLG con liga identidad y distribución Gaussiana, sin selección de variables, minimizando los errores al cuadrado y también la diferencia entre los valores quese predicen con los observados, además de alcanzar una alta correlación, mostramos sus valores a continuación:

brozek ~ . + I(variables)^2	ninguna	efectos principales y variables al cuadrado	MSE 1.182450	MSA 0.7881561	Correlaión 0.9901339