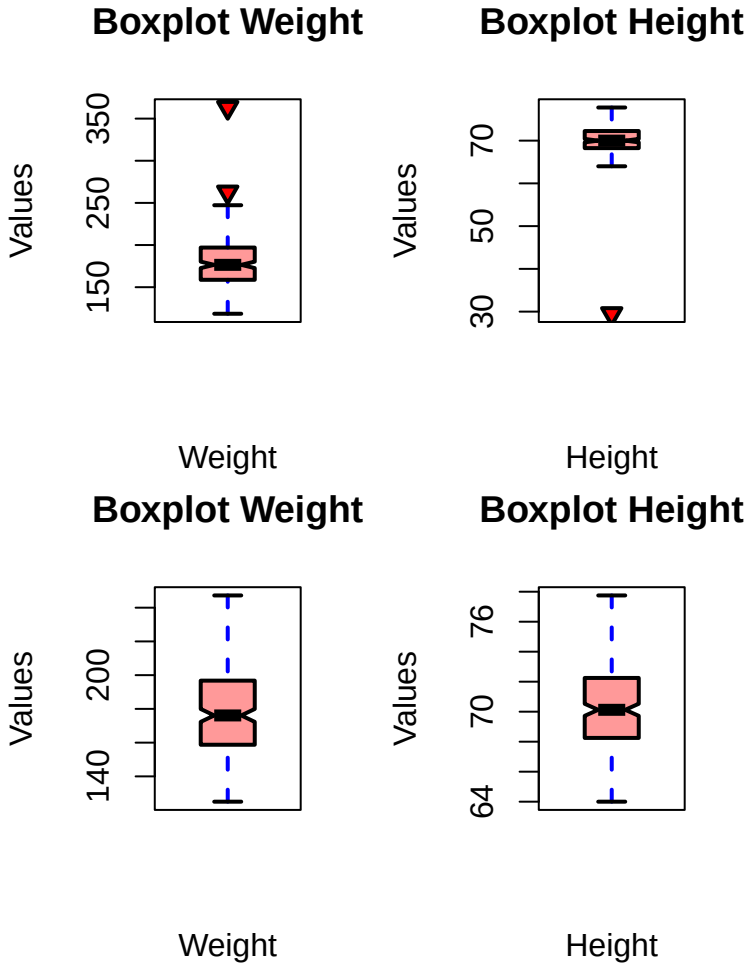
Selección de Variables

Saúl & Leobardo

2024-05-12

Para este ejercicio lo que nos interesa es usar las variables clínicas observadas en pacientes para estudiar cuales son los factores que ayudan a modelar mejor el promedio del porcentaje de grasa corporal en Hombres.



Ya que se identificaron los outliers existentes en las variables de Height y Weight lo que se hizo fue eliminarlos, dicho procedimiento se observa en los boxplot realizados.

Con el pre procesamiento realizado lo que sigue es crear modelos para datos continuos con liga identidad y distribución Gaussiana,ademas de hacer selección de variables, etc.

```
## (Intercept)      height       abdom       wrist
##   8.5938000  -0.4184531   0.7231168  -1.4836737
```

```
## [1] 1412.142
```

En este primer ajuste lo que se hizo fue hacer la selección de las mejores variables, es decir, las que aportan mas información al modelo y fueron height, abdom, wrist, de las cuales se obtuvo un BIC de 1412.142

```
##  (Intercept)       weight        abdom        wrist
## -25.32118026  -0.08768903   0.89001276  -1.22551890
```

```
## [1] 1412.255
```

Con el ajuste del modelo stepwise(forward) se obtuvo un BIC de 1412.255 el cual es muy parecido al obtenido con el del primer ajuste realizado con la seleccion de variables.

```
##  (Intercept)          age        abdom        wrist
## -10.86151335   0.07109929   0.71409295  -2.16060130
```

```
## [1] 1415.872
```

Como resultado del ajuste Backward obtuvimos un BIC de 1415.872 el cual comparado con los anteriores 2 BIC resulta mas alto.

```
## [1] 77
```

```
## (Intercept)         age      height       abdom       wrist
##  4.57171015  0.04702087 -0.32711867  0.71391555 -1.68478028
```

```
## [1] 1413.107
```

```
## [1] 1413.107
```

Con los métodos anteriormente realizados obtuvimos BIC muy similares entre si por lo que escoger uno como mejor modelo seria usar el mas parsimonioso, es decir, que resulte fácil de construirse y de interpretarse.

Ahora ajustaremos modelos parecidos a los anteriormente realizados con la diferencia de que incluiremos interacciones para ver si mejoran los modelos.

```
##   (Intercept)           hip    height:hip    neck:abdom      neck:hip
## -44.092260115   1.141649113  -0.004176813   0.020750934  -0.024928240
```

```
## [1] 1416.311
```

```
##   (Intercept)         abdom  height:wrist     chest:hip
## -24.906984524   0.873134763  -0.018543532  -0.001293605
```

```
## [1] 1405.596
```

```
##                                                                  (Intercept)
##                                                                -2.060208e+01
##        X2_aux[, unlist(coeficientes2[which.min(BIC_lasso_comp), c(-1)])]abdom
##                                                                 7.065840e-01
##    X2_aux[, unlist(coeficientes2[which.min(BIC_lasso_comp), c(-1)])]age:abdom
##                                                                -5.744683e-04
##    X2_aux[, unlist(coeficientes2[which.min(BIC_lasso_comp), c(-1)])]age:thigh
##                                                                 1.844340e-03
## X2_aux[, unlist(coeficientes2[which.min(BIC_lasso_comp), c(-1)])]height:wrist
##                                                                -2.201041e-02
```

```
## [1] 1411.985
```

```
## [1] 1411.985
```

Con los nuevos cambios en el modelo lasso con interacciones obtuvimos un BIC de 1411.985 por lo que notamos una pequeña mejoría respecto a los ajustes de modelos aplicados en la primer parte.

Agregaremos ajustes nuevos probando con distintas funciones ligas con el fin de ver si con esto logramos mejorar el puntaje de BIC obtenidos hasta este momento.

```
## [1] 1493.091
## [1] 1512.765
## [1] 1494.935
## [1] 1509.475
## [1] 1493.091
## [1] 1512.765
## [1] 1490.06
## [1] 1518.263
## [1] 1491.893
## [1] 1504.872
## [1] 1496.997
## [1] 1511.293
## [1] 7
## [1] 1490.06
##   (Intercept)           hip    hip:height    neck:abdom      hip:neck
## -0.5645385544  0.0598604549 -0.0001725943  0.0011757588 -0.0014368866
## [1] 1423.089
## [1] 1430.734
## [1] 1432.196
## [1] 1497.667
## [1] 1502.136
## [1] 1500.844
## [1] 1500.353
## [1] 1510.742
## [1] 1508.554
## [1] 2
## [1] 1423.089
##   (Intercept)        height   I(height^2)         abdom    I(abdom^2)
## -34.506777195   1.051899470  -0.010616136   1.429932998  -0.003731687
##         wrist    I(wrist^2)
##  -5.892918565   0.118790824
```

De todos los modelos generados con las modificaciones y tomando el cuadrado de las variables tenemos que el mejor modelo escogido a partir del criterio BIC fue el que toma el mejor subconjunto de variables tomando su cuadrado con un BIC de 1423.089.

No. de Modelo	Método de selección de variables	BIC
1	fitBestSubset	1412.142
2	modeloforward	1412.255
3	modelobackward	1415.872
4	AjusteModeloLasso	1413.107
5	Ajusteforward2	1405.596
6	Ajustebackward2	1416.311
7	AjusteLassoInteracciones	1411.985
8	Backward, distribución Gamma, liga identidad	1490.06
9	ajusteCuadraticosubset	1423.089

Table 1: Resultados de los métodos de selección

De acuerdo a los modelos que se observan en la tabla, la variable que mas veces aparece es *abdom*, la cual aparece un total de 6 veces, sin contar aquellos modelos en los que se toman las interacciones, seguida de la variable *wrist*, la cual aparece 4 veces, tambien sin contar interacciones. Además los coeficientes asociados a la variable abdom son positivos en los primeros cuatro modelos por lo que podemos afirmar que la variable más significativa asociada a un incremento del valor promedio de la grasa corporal en los hombres es la medida del abdomen [2]. En cambio, en los primeros cuatro modelos de la tabla que incluyen a la variable wrist los coeficientes asociados a la variable wrist son negativos en los modelos en que esta variable aparece y por lo tanto es posible inferir que a medida que la medición de la muñeca aumenta entonces habra una disminución del valor promedio de la grasa corporal en hombres [3]. En conclusión las variables más significativas para modelar el promedio de la grasa corporal en los hombres son la medida del abdomen y de la muñeca. Las variables que no aparecen entre las seleccionadas son: adipos, knee, ankle, biceps, forearm. Por lo tanto ninguna de las variables anteriormente mencionadas tiene algún efecto sobre el valor promedio del porcentaje de grasa en hombres.

Además, según el criterio BIC, se observa que al considerar modelos que toman en cuenta las interacciones entre las variables, el BIC de los modelos con interacciones disminuye. Esto indica que, al incorporar interacciones, el modelo del promedio de porcentaje de grasa corporal en hombres se ajusta mejor a los datos.

Por otra parte, de los modelos observados en el **Cuadro 1** se puede ver que aquel que tiene un menor valor BIC es el modelo cuya selección de variables se realizo con el método stepwise conocido como *Forward* incluyendo interacciones. Los coeficientes de esto modelo pueden ser interpretados de la siguiente manera: Un aumento del 100% de la variable *hip* se asocia a un aumento del 114 % en el promedio de la grasa corporal de los hombres dejando al resto de variables fijas, a su vez las variables asociadas a la altura y a la medida del cuello (*height* y *neck*) interactuan con la variable *hip*, que representa la medida de la cadera, el valor del coeficiente asociado es negativo, en particular, un aumento de una unidad para la variable asociada a la interacción *hip:height* se asocia a una disminución del 0.4 % del promedio de la grasa corporal en hombres y un aumento en una unidad a la variable asociada a la interacción *hip:neck* está asociada a una disminución del 2 % para el promedio de la grasa corporal en hombres, por otro lado la interacción *neck:abdom* está asociada a un aumento del 2% en el promedio de la grasa corporal para hombres. [5].