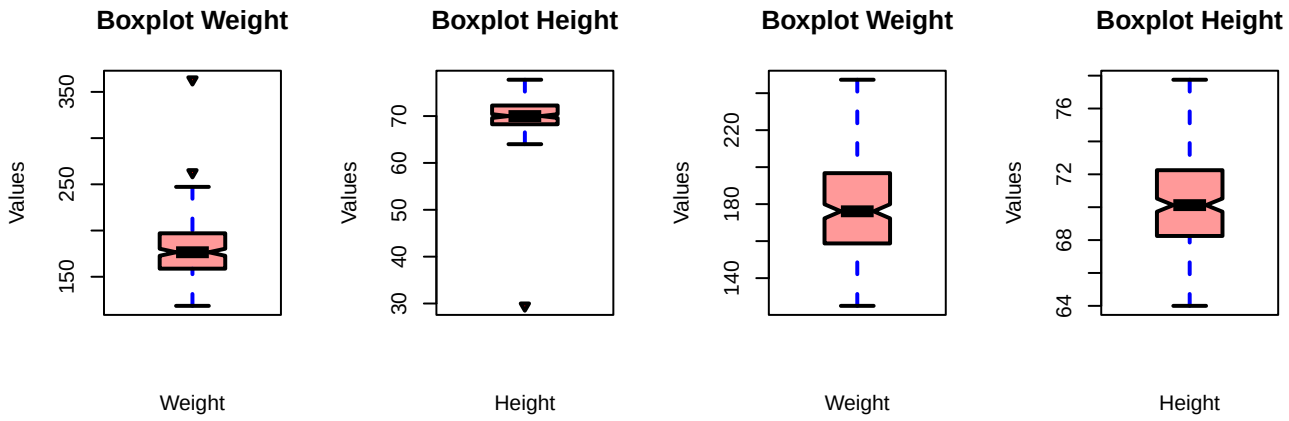
2. Selección de variables.

Nos interesa es usar las variables clínicas observadas en pacientes de la base de datos `fat` del paquete `faraway` para estudiar cuales son los factores que ayudan a modelar mejor el promedio del porcentaje de grasa corporal en Hombres (brozek). Omitiremos las variables siri, density y free, se eliminaron los valores nulos de la variable brozek, y los outliers de weight y height. Esto último se puede apreciar en la siguiente Gráfica.



Con el pre procesamiento realizado lo que sigue es crear subconjuntos de modelos para datos continuos con liga identidad y distribución Gaussiana, además de hacer selección de variables considerando efectos principales usando el mejor subconjunto, un método stepwise y lasso, con el criterio BIC para el mejor modelo. Además se sonsiderarán los subconjuntos con interacciones, términos cuadráticos para las variables, etc.

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

```
## [1] "BIC: 1412.141592"
```

En un primer subconjunto de ajuste (`fitBestSubset`) con la función `regsubsets` se hizo una selección de las mejores combinaciones de variables de las 14, el mejor resultado fue la combinación de tres variables height, abdom, wrist, con las cuales se obtuvo un menor BIC de 1412.141592. (Chunk fitBestSubset, linea 150)

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

```
## [1] "BIC: 1412.254566"
```

En el segundo subconjunto `modeloforward` con el ajuste del modelo stepwise(forward) se obtuvo un BIC de 1412.2545657el cual es muy parecido pero ligeramente mayor al obtenido con el del primer ajuste realizado con la seleccion de variables.(Chunk modeloforward, linea 198)

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

```
## [1] "BIC: 1415.872346"
```

En el tercer subconjunto `modelobackward` con el ajuste Backward obtuvimos un BIC de 1415.8723455 el cual comparado con los dos anteriores BIC resulta mas alto.(Chunk modelobackward, linea 212)

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

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

```
## [1] "BIC: 1413.106619"
```

El cuarto subconjunto de modelos `AjusteModeloLasso`, corresponde al modelo lasso, donde se obtuvo un BIC de 1413.1066192. (Chunk AjusteModeloLasso, linea 235)

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

```
## [1] "BIC: 1405.595944"
```

Para el quinto subconjunto `Ajusteforward2`, el resultado del forward con interacciones muestra un BIC de 1405.5959444.(Chunk Ajusteforward2, linea 292)

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

```
## [1] "BIC: 1416.310922"
```

Para el sexto subconjunto `Ajustebackward2`, el resultado del backward con interacciones muestra un BIC de 1416.3109222. (Chunk Ajustebackward2, linea 318)

```
##   (Intercept)         abdom     abdom:age     age:thigh  height:wrist
## -2.060208e+01  7.065840e-01 -5.744683e-04  1.844340e-03 -2.201041e-02
```

```
## [1] "BIC: 1411.984695"
```

Para el séptimo subconjunto `AjusteLassoInteracciones`, con los nuevos cambios en el modelo lasso con interacciones obtuvimos un BIC de 1411.9846953. (Chunk AjusteLassoInteracciones, linea 345)

Con las interacciones notamos una pequeña mejoría del BIC.

Ahora, probaremos con distintas funciones ligas (identidad, log) en combinación con el modelo Gama con el fin de ver si con esto logramos mejorar el puntaje de BIC obtenidos hasta este momento.

```
## [1] 7
```

```
##   (Intercept)           hip    hip:height    neck:abdom      hip:neck
## -0.5645385544  0.0598604549 -0.0001725943  0.0011757588 -0.0014368866
```

```
## [1] "BIC: 1490.059783"
```

Para el octavo subconjunto de modelos `GamaLigasBackForLasso`, el mejor modelo considerando el modelo Gama con distintas ligas (identidad, log) y distintos métodos tales como backward, forward y lasso, es el que tiene un BIC de 1490.0597827, el cual es una Gama con liga log. (Chunk GamaLigasBackForLasso, linea 396)

```
##   (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
```

```
## [1] "BIC: 1423.088935"
```

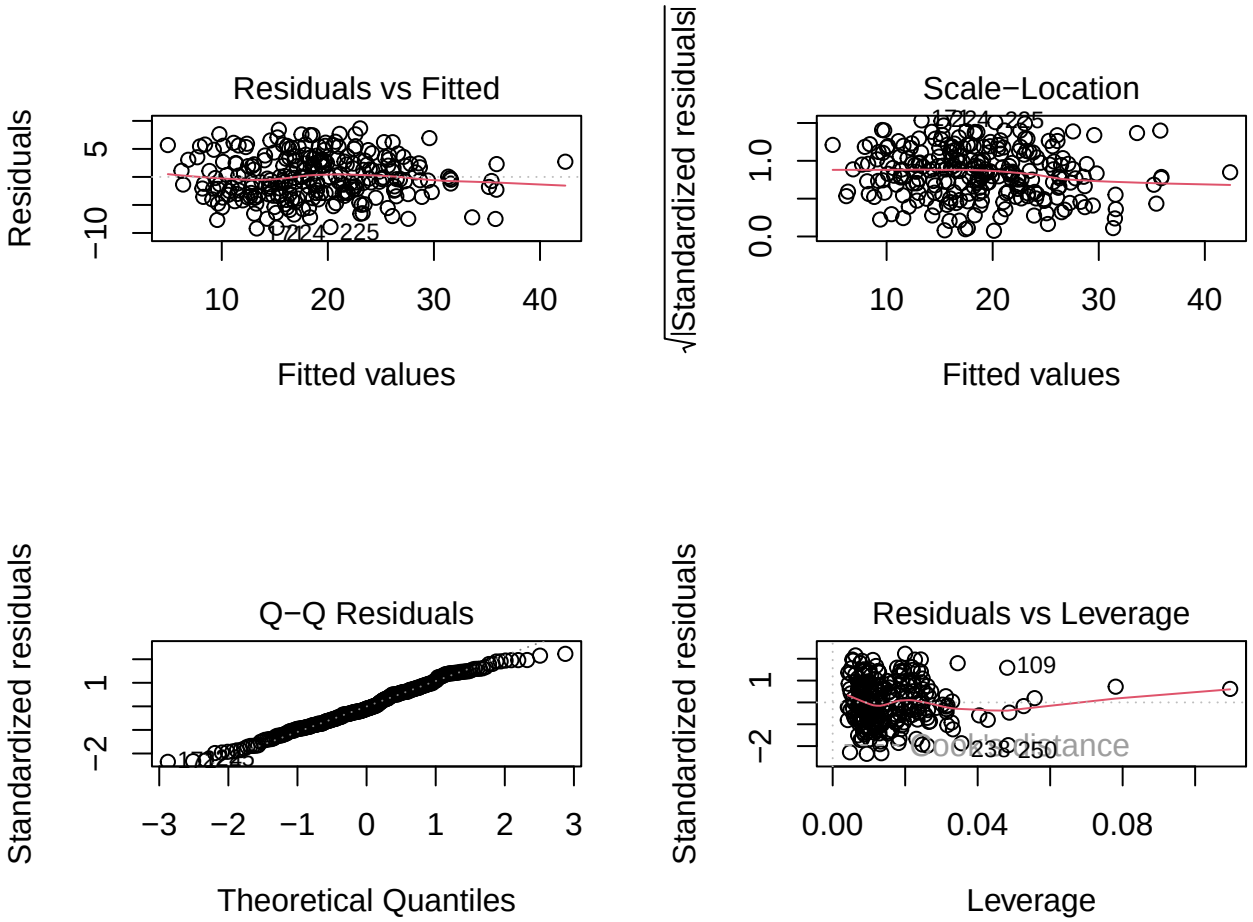
Por último, en el subconjunto noveno de modelos `ajusteCuadraticosubset`, usando una versión extendida que integra el cuadrado de las variables, se tiene un BIC de 1423.0889354 como el mejor. (Chunk ajusteCuadraticosubset, linea 463)

Presentamos a contunuación un Cuadro con los mejores modelos obtenidos en cada subconjunto con su respectivo BIC. Es posible observar que el modelo con el menor BIC de 1405.596 es uno con interacciones, en donde se consideran las covariables `abdom`, y las interacciones de `height:wrist` y `chest:hip`. La variable presente en todos los modelos es `abdom`, seguido de `wrist` en 7 modelos.

Para inferencia e interpretación de los coeficientes del modelo elegido, es necesario el cumplimiento de los supuestos. En la prueba gráfica de los supuestos, tales como la linealidad (Residuals vs Fitted), homocedasticidad (Scale-Location), normalidad (Q-Q Residuals) y presencia de outliers influyentes (Residuals vs Leverage), se observa que no hay problemas graves con los supuestos. (Chunk plotsmodelo, linea 566)

No.	Método de selección	Covariables y coeficientes estimados	BIC
1	fitBestSubset	(Intercept), height, abdom, wrist 8.5938, -0.4184531, 0.7231168, -1.4836737	1412.142
2	modeloforward	(Intercept), weight, abdom, wrist -25.3211803, -0.087689, 0.8900128, -1.2255189	1412.255
3	modelobackward	(Intercept), age, abdom, wrist -10.8615133, 0.0710993, 0.714093, -2.1606013	1415.872
4	AjusteModeloLasso	(Intercept), age, height, abdom, wrist 4.5717101, 0.0470209, -0.3271187, 0.7139155, -1.6847803	1413.107
5	Ajusteforward2	(Intercept), abdom, height:wrist, chest:hip -24.9069845, 0.8731348, -0.0185435, -0.0012936	1405.596
6	Ajustebackward2	(Intercept), hip, height:hip, neck:abdom, neck:hip -44.0922601, 1.1416491, -0.0041768, 0.0207509, -0.0249282	1416.311
7	AjusteLassoInteracciones	(Intercept), abdom, abdom:age, age:thigh, height:wrist -20.6020769, 0.706584, $-5.7446828 \times 10^{-4}$, 0.0018443, -0.0220104	1411.985
8	GamaLigasBackForLasso	(Intercept), hip, hip:height, neck:abdom, hip:neck -0.5645386, 0.0598605, -1.725943×10^{-4}, 0.0011758, -0.0014369	1490.06
9	ajusteCuadraticosubset	(Intercept), height, I(height^2), abdom, I(abdom^2), wrist, I(wrist^2) -34.5067, 1.0519, -0.0106, 1.4299, -0.0037, -5.8929, 0.1188	1423.089

Cuadro 1: Resultados de los métodos de selección



La linealidad se comprueba con la siguiente prueba. (Chunk linealidad, linea 586)

```
##        abdom height_wrist    chest_hip   Tukey test
##   0.26485820   0.71940474   0.08328658   0.50212569
```

De acuerdo con la prueba `studentized Breusch-Pagan` se tiene un p-value de 0.3265423 por lo que no se rechaza

la hipótesis nula de homocesaticidad, por otra parte las pruebas de normalidad Jarque-Bera y Kolmogorov-Smirnov no rechazan la hipótesis nula de normalidad, con p-value de 0.1360153 y 0.1877934, respectivamente. (Chunk pruebasmodelo, linea 575)

Con esto, podemos argumentar que por cada unidad de incremento en la circunferencia del abdomen `abdom` en cm, el porcentaje de grasa corporal (brozek) incrementa en 0.8731348. Por otra parte, con el incremento en una unidad de la interacción estatura - curcunferencia de la muñeca (`height:wrist`) disminuye el porcentaje de grasa corporal en -0.0185435, y el incremento en una unidad de la interacción circunferencia del pecho - circunferencia de cadera disminuye el porcentaje de grasa corporal en -0.0012936.