Regresion lineal

Contents

Introducción

Los datos de este estudio se pueden incorporar a R desde el data.frame bodyfat del paquete TH.data.

Las variables del estudio son:

· age es la edad en años.

· DEXfat es la grasa corporal medida con DXA, variable respuesta.

· waistcirc es la circunferencia de la cintura.

· hipcirc es el perímetro de la cadera.

· elbowbreadth es la anchura del codo.

· kneebreadth es la anchura de la rodilla.

· anthro3a es la suma del logaritmo de tres medidas antropométricas.

· anthro3b es la suma del logaritmo de tres medidas antropométricas.

· anthro3c es la suma del logaritmo de tres medidas antropométricas.

· anthro4 es la suma del logaritmo de tres medidas antropométricas.

Exploración de los datos

Primera observación de las variables, vemos que todas son numéricas.

```
## 'data.frame':    71 obs. of  10 variables:
##  $ age         : num  57 65 59 58 60 61 56 60 58 62 ...
##  $ DEXfat      : num  41.7 43.3 35.4 22.8 36.4 ...
##  $ waistcirc   : num  100 99.5 96 72 89.5 83.5 81 89 80 79 ...
##  $ hipcirc     : num  112 116.5 108.5 96.5 100.5 ...
##  $ elbowbreadth: num  7.1 6.5 6.2 6.1 7.1 6.5 6.9 6.2 6.4 7 ...
##  $ kneebreadth : num  9.4 8.9 8.9 9.2 10 8.8 8.9 8.5 8.8 8.8 ...
##  $ anthro3a    : num  4.42 4.63 4.12 4.03 4.24 3.55 4.14 4.04 3.91 3.66 ...
##  $ anthro3b    : num  4.95 5.01 4.74 4.48 4.68 4.06 4.52 4.7 4.32 4.21 ...
##  $ anthro3c    : num  4.5 4.48 4.6 3.91 4.15 3.64 4.31 4.47 3.47 3.6 ...
##  $ anthro4     : num  6.13 6.37 5.82 5.66 5.91 5.14 5.69 5.7 5.49 5.25 ...
```

A continuación vemos un resumen de los resultados, para confirmar que no vemos nada raro en los datos.

```
##       age            DEXfat        waistcirc         hipcirc     
##  Min.   :19.00   Min.   :11.21   Min.   : 65.00   Min.   : 88.00  
##  1st Qu.:42.00   1st Qu.:22.32   1st Qu.: 78.50   1st Qu.: 96.75  
##  Median :56.00   Median :29.63   Median : 85.00   Median :103.00  
##  Mean   :50.86   Mean   :30.78   Mean   : 87.38   Mean   :105.28  
##  3rd Qu.:62.00   3rd Qu.:39.33   3rd Qu.: 99.75   3rd Qu.:111.15  
##  Max.   :67.00   Max.   :62.02   Max.   :117.00   Max.   :132.00  
##   elbowbreadth    kneebreadth        anthro3a        anthro3b    
##  Min.   :5.200   Min.   : 7.200   Min.   :2.400   Min.   :2.580  
##  1st Qu.:6.200   1st Qu.: 8.600   1st Qu.:3.540   1st Qu.:4.060  
##  Median :6.500   Median : 9.200   Median :3.970   Median :4.390  
##  Mean   :6.508   Mean   : 9.301   Mean   :3.869   Mean   :4.291  
##  3rd Qu.:6.900   3rd Qu.: 9.800   3rd Qu.:4.155   3rd Qu.:4.660  
##  Max.   :7.400   Max.   :11.800   Max.   :4.680   Max.   :5.010  
##     anthro3c        anthro4     
##  Min.   :2.050   Min.   :3.180  
##  1st Qu.:3.480   1st Qu.:5.040  
##  Median :3.990   Median :5.530  
##  Mean   :3.886   Mean   :5.398  
##  3rd Qu.:4.345   3rd Qu.:5.840  
##  Max.   :4.620   Max.   :6.370
```

Gráficas y calculo de la matriz de correlaciones entre variables 2 a 2

En la figura 1 podemos ver la correlación entre las varianbles. La correlación mide la relación lineal entre dos variables. Si nos dirigimos a la segunda fila, vemos la relación de DEXfat con el resto de variables, observamos una correlación positiva. Las variables con menor correlación son age con un 0,271 y elbowbreadth con un 0,354.

Estudio de VIF

Es muy frecuente que varias variables explicativas estén correlacionadas. Para confirmar que hay multicolinealidad, esta relación debe ser fuerte (>0.8) pero no perfecta (1). Esto lo hemos podido comprobar en la gráfica anterior.

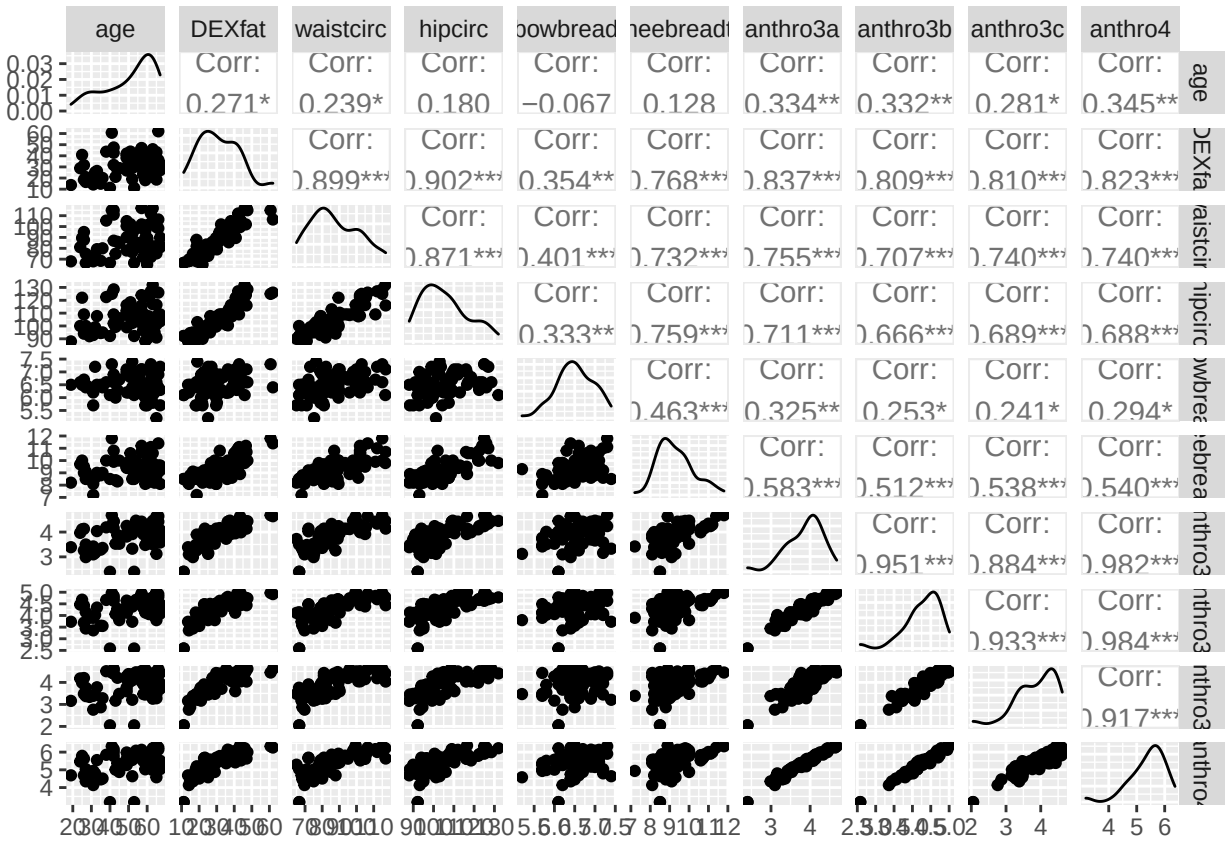


Figure 1: Gráfico de dispersión, densidad y correlación. *** significa estadísticamente significativa

Hay varias formas de detectar la multicolinealidad. Una de ellas es viendo si R2 es alta, pero los coeficientes t´s no son significativos. Para ello podemos ver el resumen de la regresión lineal. En este caso podemos sospechar, porque nos encontramos con un R2 de 0.9117 y con solo 3 de las 9 variables con resultado significativo.

```
##
## Call:
## lm(formula = DEXfat ~ ., data = bodyfat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.954 -1.949 -0.219  1.169 10.812
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -69.02828    7.51686  -9.183 4.18e-13 ***
## age            0.01996    0.03221   0.620  0.53777
## waistcirc      0.21049    0.06714   3.135  0.00264 **
## hipcirc        0.34351    0.08037   4.274 6.85e-05 ***
## elbowbreadth  -0.41237    1.02291  -0.403  0.68826
## kneebreadth    1.75798    0.72495   2.425  0.01829 *
## anthro3a       5.74230    5.20752   1.103  0.27449
## anthro3b       9.86643    5.65786   1.744  0.08622 .
## anthro3c       0.38743    2.08746   0.186  0.85338
## anthro4       -6.57439    6.48918  -1.013  0.31500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.281 on 61 degrees of freedom
## Multiple R-squared:  0.9231, Adjusted R-squared:  0.9117
## F-statistic: 81.35 on 9 and 61 DF,  p-value: < 2.2e-16
```

Otra forma es mediante el Factor de Influencia de la Varianza (VIF). Es una medida del grado en que la varianza del estimador de mínimos cuadrados se incrementa por esta colinealidad. Podemos situar la correlación mayor a 0.8 a partir de 5vif y 0.9 con 10vif. Por lo tanto, todos lo que sean mayor presentan multicolinealidad. En este caso podemos decir que son anthro3a, anthro3b, anthro3b las variables que presentan multicolinealidad.

```
##          age    waistcirc      hipcirc elbowbreadth  kneebreadth     anthro3a
##     1.207378     5.770604     5.121272     1.416972     2.853228    39.471164
##     anthro3b     anthro3c      anthro4
##    49.100539     9.037833   111.261438
```

Contraste Test F

Sabemos que muchas veces debemos confirmar si muchos de los predictores que utilizamos son necesarios. Por motivos de simplicidad, siempre se elige el modelo más pequeño, si la diferencia entre modelos no es muy grande.

Vamos a eliminar dos variables para confirmar si el modelo sigue funcionando. Es decir, si H0: anthro3b = anthro4 = 0.

```
## Analysis of Variance Table
```

```
##
## Model 1: DEXfat ~ age + waistcirc + hipcirc + elbowbreadth + kneebreadth +
##     anthro3a + anthro3c
## Model 2: DEXfat ~ age + waistcirc + hipcirc + elbowbreadth + kneebreadth +
##     anthro3a + anthro3b + anthro3c + anthro4
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     63 692.20
## 2     61 656.52  2    35.684 1.6578  0.199
```

El valor p de 0.199 indica que la hipótesis nula no puede rechazarse aquí. Por lo tanto podemos trabajar con el modelo simple.

Propuesta de mejora para el modelo

Las transformaciones en la respuesta y/o predictores pueden hacer ajustes y mejoras en los modelos. En este caso veremos la transformación en la respuesta mediante Box-Cox. Este método selecciona la transformación a utilizar sobre la variable para resolver la supuesta no normalizadad.

Para trabajar necesitamos extraer el parámetro lambda. A continuación observamos que el valor se encuentra en 0.2626263. Si nos vamos a la tabla de transformaciones de lambda, podríamos proponer que se haga una raiz cuadrada.

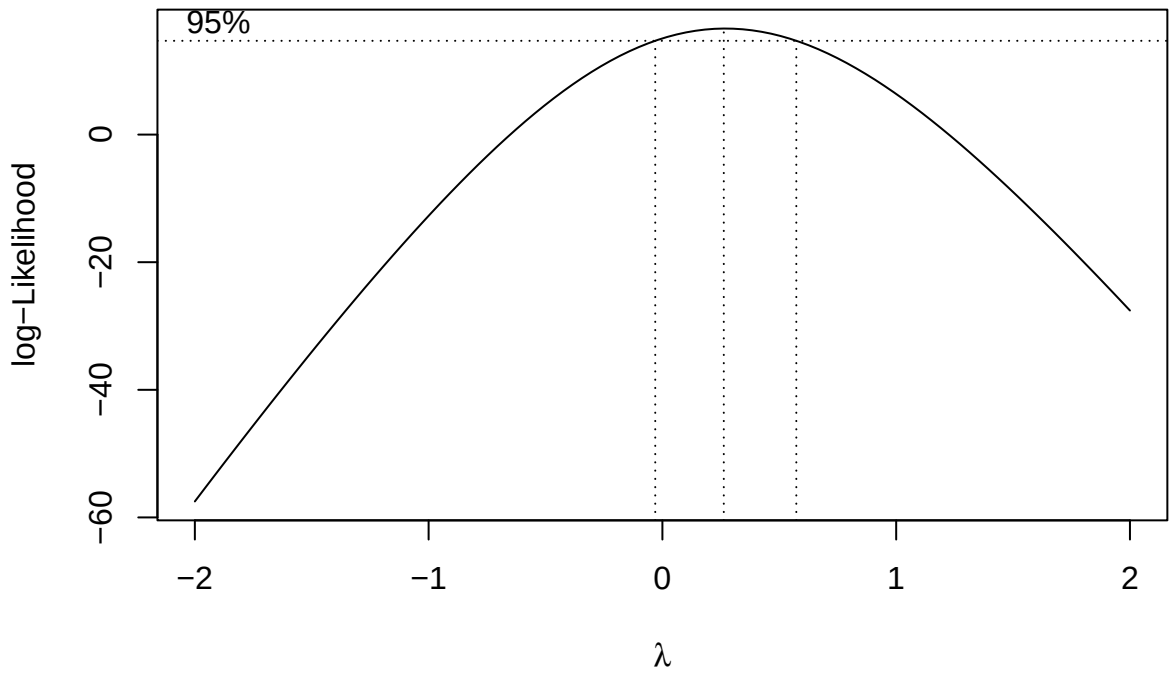


Figure 2: Gráfico de lambda

```
## [1] 0.2626263
```

Nuevo modelo:

```
nuevo_modelo <- lm(sqrt(DEXfat) ~ age + waistcirc +
hipcirc + elbowbreadth + kneebreadth + anthro3a + anthro3c, bodyfat)
```

Análisis de residuos modelo lmod2

Vamos a estudiar si hay homocedasticidad y una distribución normal. Podemos observar en la primera gráfica (Residuos vs ajustados) que la línea roja hace referencia a la relación no lineal entre las variables predictivas y la variable respuesta. La segunda gráfica (Q-Q) obsersevamos que los residuos no siguen una distribución normal en alguno de los puntos. La tercera muestra que los residuos no se distribuyen por igual a lo largo de los rangos de los predictores. La cuarta muesta los valores atípicos.

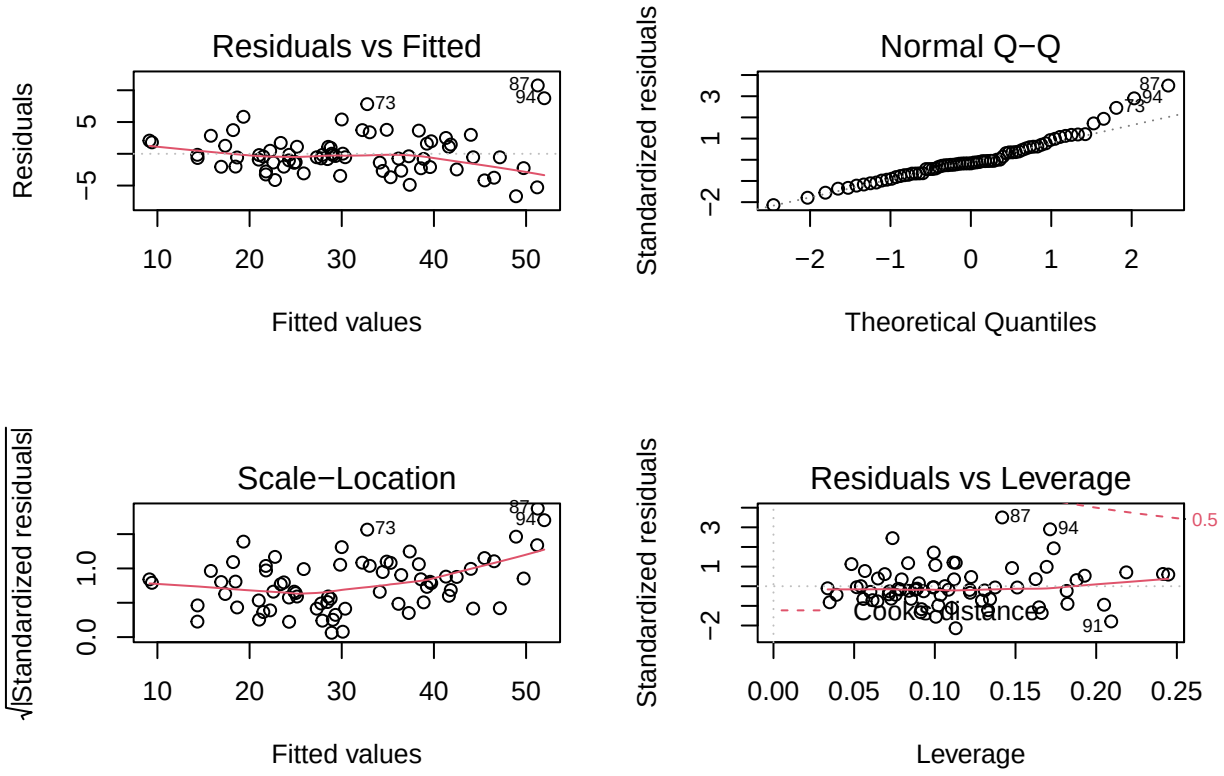


Figure 3: Residuos vs Ajustados, Normal Q-Q, Escala vs Ubicacion y Residuos vs Leverage

Para confirmar se realizan dos test: La prueba Breusch-Pagan. hipótesis nula: Homocedaticidad hipótesis alternativa: Heterocedaticidad

```
##
##  studentized Breusch-Pagan test
##
## data:  lmod2
## BP = 17.145, df = 7, p-value = 0.01649
```

Podemos observar que el valor de p es 0.01649, menor que 0.05, por lo tanto se rechaza la hipótesis nula y confirmamos que existe heterocedasticidad.

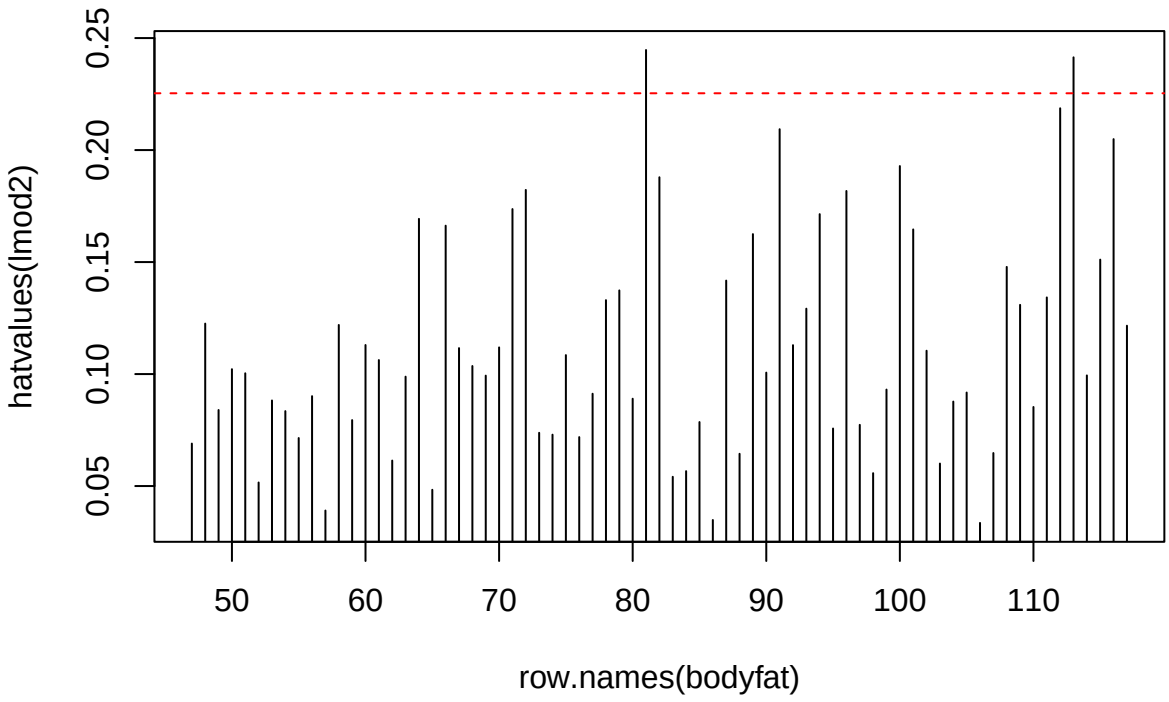
Prueba Shapiro-Wilk: Hipótesis nula: Normalidad Hipótesis alternativa: No normalidad

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(lmod2)
## W = 0.94551, p-value = 0.003922
```

Podemos observar que el valor de p es 0.003922, menor que 0.05, por lo tanto se rechaza la hipótesis nula y confirmamos que no hay normalidad.

Análisis de valores atípicos

En este punto vamos a comprobar si alguna observación alta de **leverage**. Esto se realiza para saber si hay algún valor atípico que pueda influir en las variables.




```
##        81       113       112        91       116       100
## 0.2446719 0.2414016 0.2186666 0.2093134 0.2048922 0.1928627
```

Tenemos dos puntos con alto leverage, el 81 y el 113.

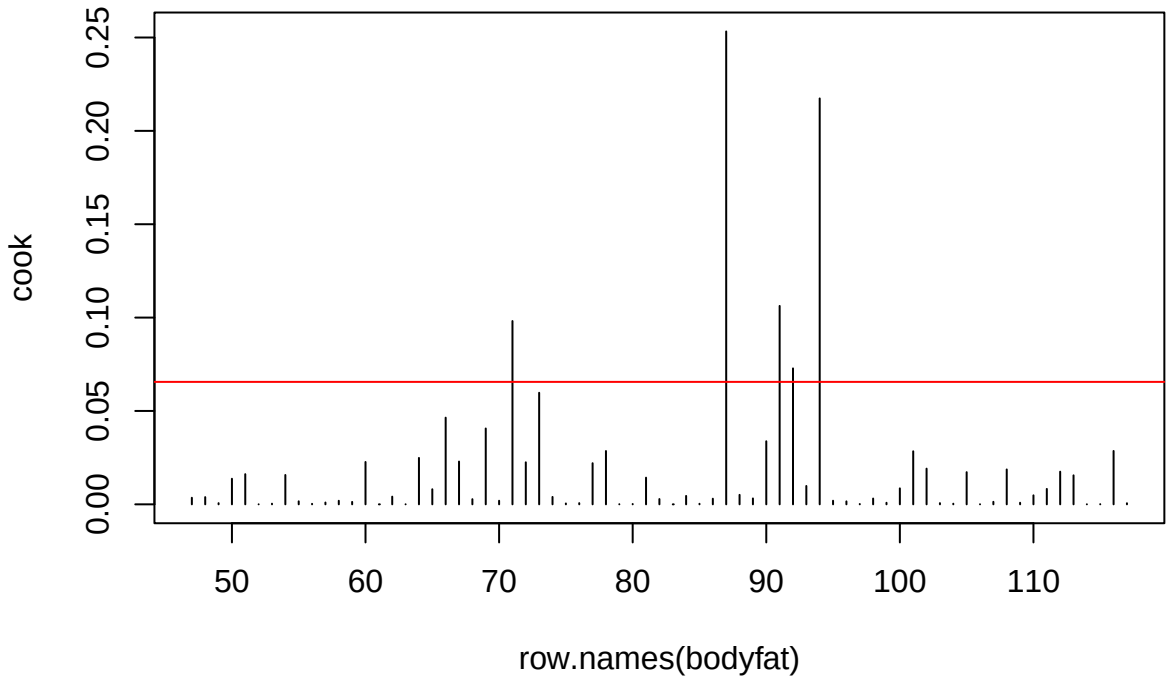
· Outliers:

```
##    rstudent unadjusted p-value Bonferroni p
## 87 3.872077          0.0002623     0.018624
```

Tenemos un punto que es un outlier, el 87.

· Puntos influentes

```
##         87         94         91         71         92         73
## 0.25324431 0.21737723 0.10626480 0.09816008 0.07281841 0.05973962
```




```
##         87         94         91         71         92
## 0.25324431 0.21737723 0.10626480 0.09816008 0.07281841
```

Tenemos 5 puntos influyentes: 87, 94, 91, 71, 92.

Predicción en forma de intervalos de confianza al 90%

Tenemos una serie de valores observados: age = 62, waistcirc = 100, hipcirc = 105, elbowbreadth = 6.8, kneebreadth = 9.5, anthro3a = 4.2, anthro3c = 4.3

Para estos valores queremos hacer una predicción de DEXfat, marcando un intervalo de predicción del 90%.

```
##        fit      lwr     upr
## 1 36.08391 30.35542 41.8124
```

Concluimos que el 90% de las personas con los anteriores parámetros tienen una grasa corporal entre el 30.36 y 41.81 DEXfat.

Ahora vamos a comprobar si los valores observados no suponen una extrapolación. La extrapolación ocurre cuando tratamos de predecir la respuesta para valores del predictor que se encuentran fuera del rango de los datos originales. Vemos a continuación que los valores se encuentran dentro del elipsoide y que la predicción no constituye una extrapolación.

```
h <- max(hatvalues(lmod2))
x <- c(1,62,100,105,6.8,9.5,4.2,4.3)
XtXinv <- summary(lmod2)$cov.unscaled
t(x)%*%XtXinv%*%x < h
```

```
##       [,1]
## [1,] TRUE
```