# Regresión, modelos y métodos - PEC 1

*Alejandro Keymer*

*24/11/2019*

## Contents



```
pacman::p_load(rcompanion, faraway, tidyverse, broom, corrplot, GGally)
```

## Problema 1 (30 pt.)

**Gladstone (1905) estudió la relación entre el peso del cere-bro humano y algunas medidas de la cabeza en personas deambos sexos y diferentes edades, fallecidas por diversas cau-sas. Los datos se hallan agrupados por sexos y grupos de edad en el documento original y en el archivo `cbrain.dat`, los dos adjuntos. Estos datos fueron estudiados por Blakeman et al. (1905) desde el punto de vista descriptivo y con diversos modelos de regresión.**

### (a)

**Incorporar a R esta base de datos, teniendo en cuenta la codificación de los valores faltantes *(missingvalues)*. Identificar con su número correspondiente las observaciones con algún dato faltante. ¿Cuantas observaciones son? Eliminar de la base de datos todas las observaciones para las que falta algún dato. Fijaros que conservamos la columna `obs` que permite identificar las observaciones en la base de datos original.**

```
# leer la tabla y modificar tipos de variable.
cbrain_raw <-
    read.table("cbrain.dat", skip = 18, header = T, na.strings = '-1') %>%
    mutate(obs = as.character(obs),
           cause = factor(cause, levels = c(1,2,3), labels = c('A', '-', 'C')),
           sex = factor(sex, levels = c(0,1), labels = c('F', 'M')),
           ageclass = factor(ageclass, levels = c(0,1,2),
                             labels = c('child', 'age 20-45', 'over 45'))
    )
```

Importo la base con `read.table` y utilizo operados de `dplyr` para convertir el *tipo* de algunas de las variables a factores.

```
# identificar filas con algun valor con NA
cbrain_raw %>%
    filter(!complete.cases(.))
```

```
##   obs cause   age height headht length breadth size brnweight circum
## 1   1     -    NA     67    141  200.0     160 4512      1530    574
## 2  25     A    NA     71    142  199.0     168 4747      1635    602
## 3  76     -    NA     63    126  184.5     146 3394      1195    540
## 4  89     A    NA     65    130  186.0     145 3506      1280    538
## 5 135     C 36.00     NA    114  179.0     140 2857      1027    570
## 6 141     -    NA     64    137  175.0     149 3572      1270    532
## 7 188     C 46.00     66    116  185.0     143 3069      1022     NA
## 8 228     A 50.00     NA    142  191.0     155 4204      1380    559
## 9 239     A  0.75     28    114  155.0     117 2067        NA     NA
##   cephalic sex  ageclass
## 1     80.0   M age 20-45
## 2     84.4   M age 20-45
## 3     79.1   M   over 45
## 4     78.0   M   over 45
## 5     78.2   F age 20-45
## 6     85.1   F age 20-45
## 7     77.3   F   over 45
## 8     81.2   F   over 45
## 9     75.5   M     child
```

```
cbrain <-
    cbrain_raw %>%
    filter(complete.cases(.))
```

Para identificar los valores *NA* utilizo un filtro con el inverso de `complete.cases`. De esta forma obtengo el listado de las filas con *al menos* un valor *NA*.

Para responder exactamente a la pregunta;

- El número de observaciones con al menos un valor *NA* es: 9.
- Las observaciones `obs` con valores *NA* son: 1, 25, 76, 89, 135, 141, 188, 228, 239.

---

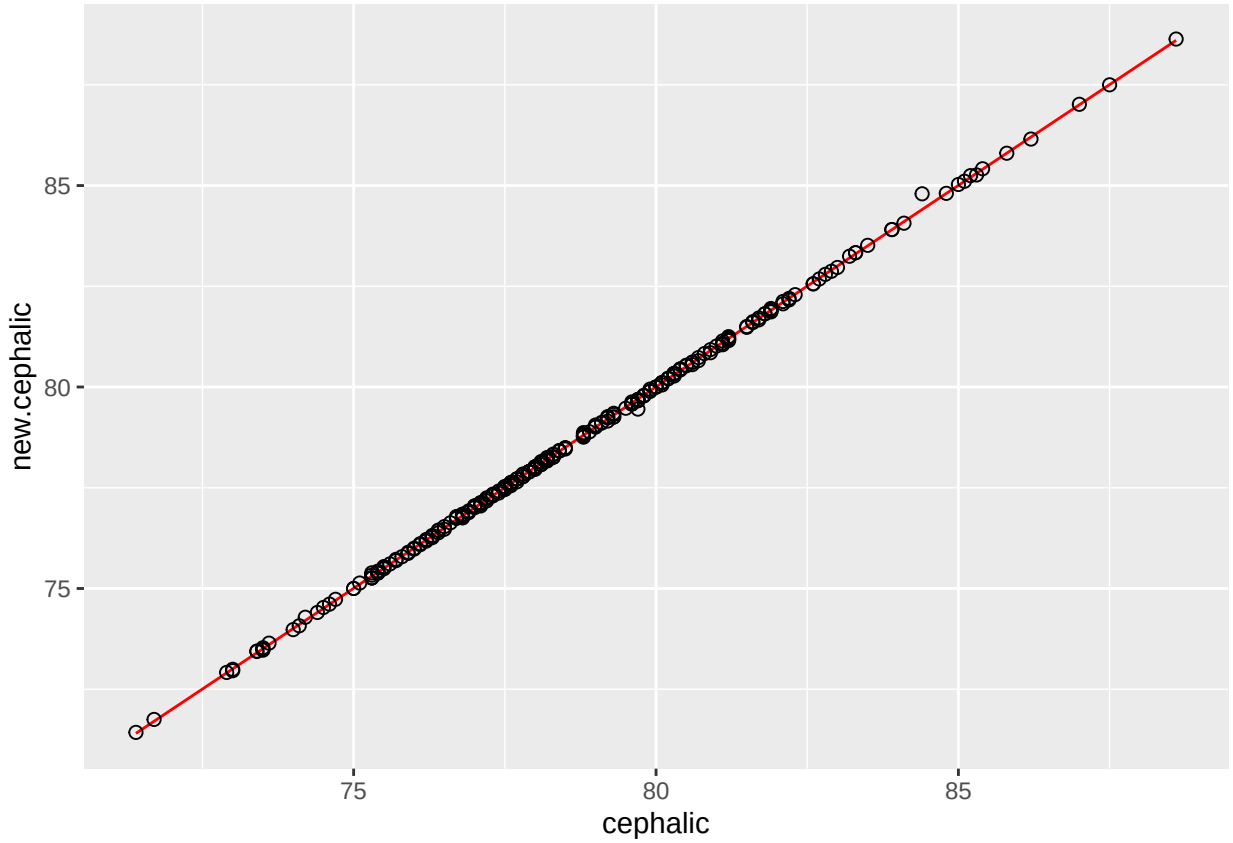
**(b)**

**Calcular la variable `new.cephalic` calculando la relación $(B/L)$»100 y estudiar la discrepancia con `cephalic` con una recta de regresión.**

```
cbrain <-
    cbrain %>%
    mutate(new.cephalic = breadth / length * 100)

# crear modelo lineal
model_new <-
    cbrain %>%
    column_to_rownames("obs") %>%
    lm(cephalic ~ new.cephalic, data = .)
```

```r
# Creamos una gráfica con la curva de regresion
cbrain %>%
    select(cephalic, new.cephalic) %>%
    ggplot(aes(cephalic, new.cephalic)) +
    geom_smooth(method = 'lm', formula = y ~ x, color = "red", size = .5) +
    geom_point(shape = 1, size = 2)
```




En la recta de regresion se puede observar una relacion que sigue de manera muy fidedigna la recta, salvo por dos puntos, uno mas discordante que el otro.

---

**(i)**

**¿Cual es la correlación entre ambas?**

```r
(cef_cor <- cor(cbrain$cephalic, cbrain$new.cephalic))
```

```
## [1] 0.9998977
```

Para el cálculo de al correlacion de ambas variables utilizo la función `cor`. La correlación es 0.9998977, Lo que se acerca mucho a 1.

---

**(ii)**

**¿Podemos aceptar que la pendiente es 1?**

$H0 : \beta_1 = 1$

```r
# construir IC para el 95%
confint(model_new)
```

```
##                   2.5 %    97.5 %
## (Intercept)  -0.1296286 0.1428785
## new.cephalic  0.9981275 1.0015873
```

Se puede construir el intrevalo de confianza para el modelo, y es posible observar que el intervalo incluye el 1. Lo mas correcto sería utilizar los intervalos de confianza, ya que además de el valor de la pendiente (o los límites de los valores) da una idea del tamaño del efecto del modelo.

**(iii)**

**¿Podemos aceptar con un contraste que el coeficiente de intercepción es cero y la pendiente es 1 (las dos cosas a la vez)?**

```r
model_1 <- lm(cephalic ~ 0 + offset(1*new.cephalic), cbrain)

anova(model_1, model_new)
```

```
## Analysis of Variance Table
##
## Model 1: cephalic ~ 0 + offset(1 * new.cephalic)
## Model 2: cephalic ~ new.cephalic
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    267 0.46036
## 2    265 0.45466  2 0.0056951 1.6597 0.1922
```

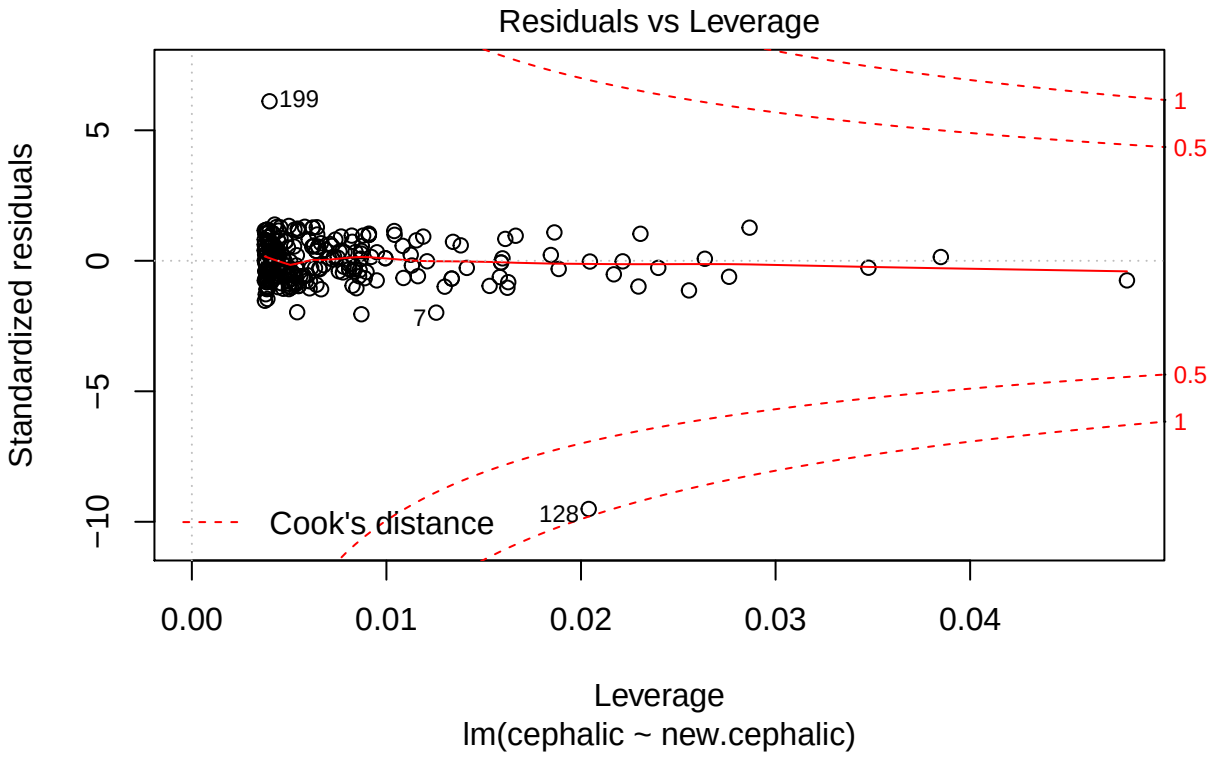
No se puede rechazar la Hipótesis nula , por lo que se podría aceptar lo propuesto en la pregunta, aceptando el modelo mas simple que es una recta de pendiente 1 que pasa por el 0.

---

**(iv)**

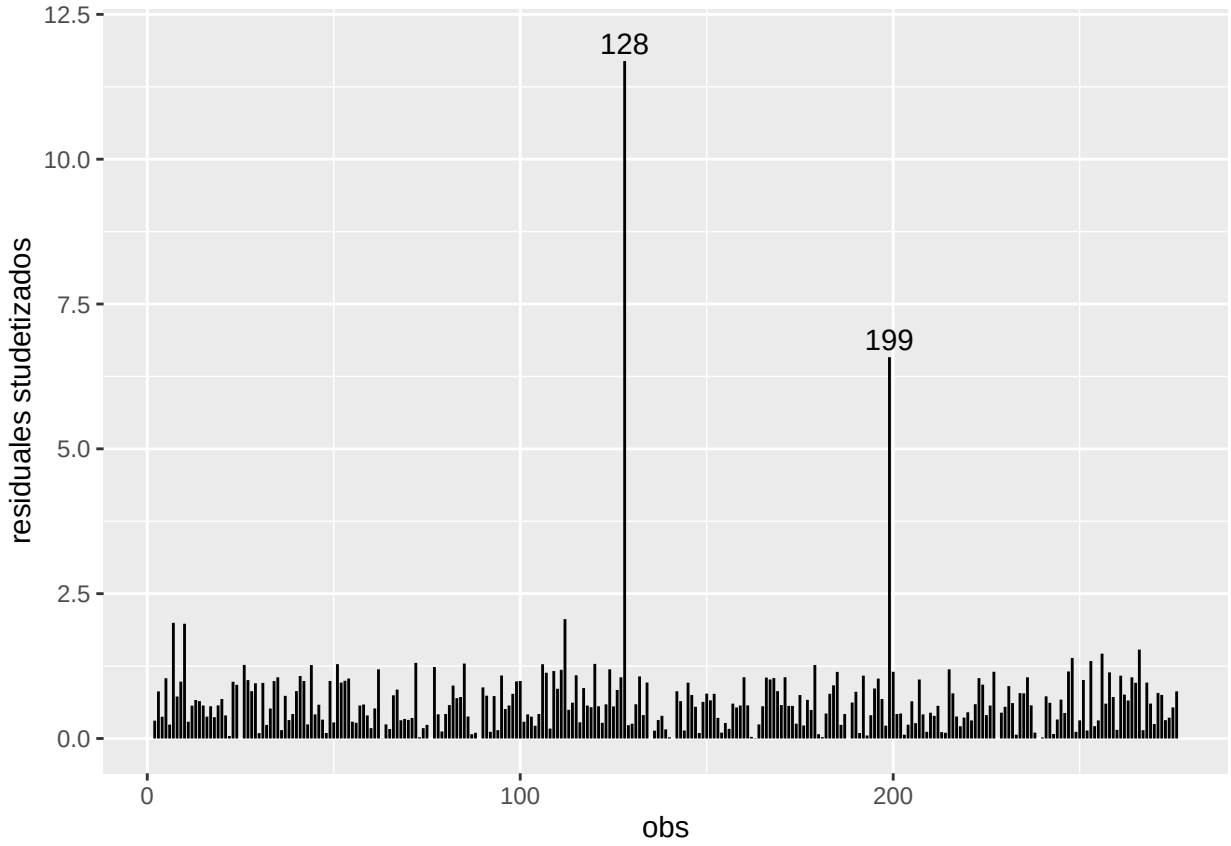
**En esta regresión, identificar posibles residuos atípicos o outliers (sin hacer ningún contraste) y valorar si el dato de cephalic anotado en la base de datos original es una errata.**

```r
# grafica de distancias de hatvalues v/s residuales std
plot(model_new, 5)
```




```
# grafica de residuales studetizados
p_df <-
augment(model_new) %>%
    mutate(t.resid = rstudent(model_new),
           obs = as.integer(cbrain$obs))

ggplot(p_df, aes(obs, abs(t.resid))) +
    geom_segment(aes(xend = obs, yend = 0)) +
    geom_text(data = filter(p_df, abs(t.resid) > abs(qt(
        .05 / length(resid(model_new)) * 2, (length(resid(model_new)) - length(coef(model_new)))
    ))), aes(label = obs), nudge_y = .3) +
    labs(y = "residuales studetizados")
```




Utilizo tres métodos para valorar la prescencia de *outliers*.

- Gráfica de *residuos estandatizados* v/s *valores extremos* + curvas de distancias de Cook
- Gráfica de los *residuos studentizados* v/s observaciones

En las gráficas es claramente visible que hay dos valores que sobrepasan al resto de manera peculiar. En el gráfico de valores extremos, podemos observar que si bien son valores que no tienen mucha palanca, poseen una distancia de Cook alta y alejada del resto de los valores para este modelo.

Finalmente en el cálculo de los residuos studentizados, se ve que los valores sobrepasan el valor de la corrección de Bonferroni, estrategia que se plantea en el libro de Faraway.

En este sentido se puede establcer que el valores `192`, es efectivamente un *outlier* y habría que revisar mejor la obs `128`, ya que tambien es algo anómala.

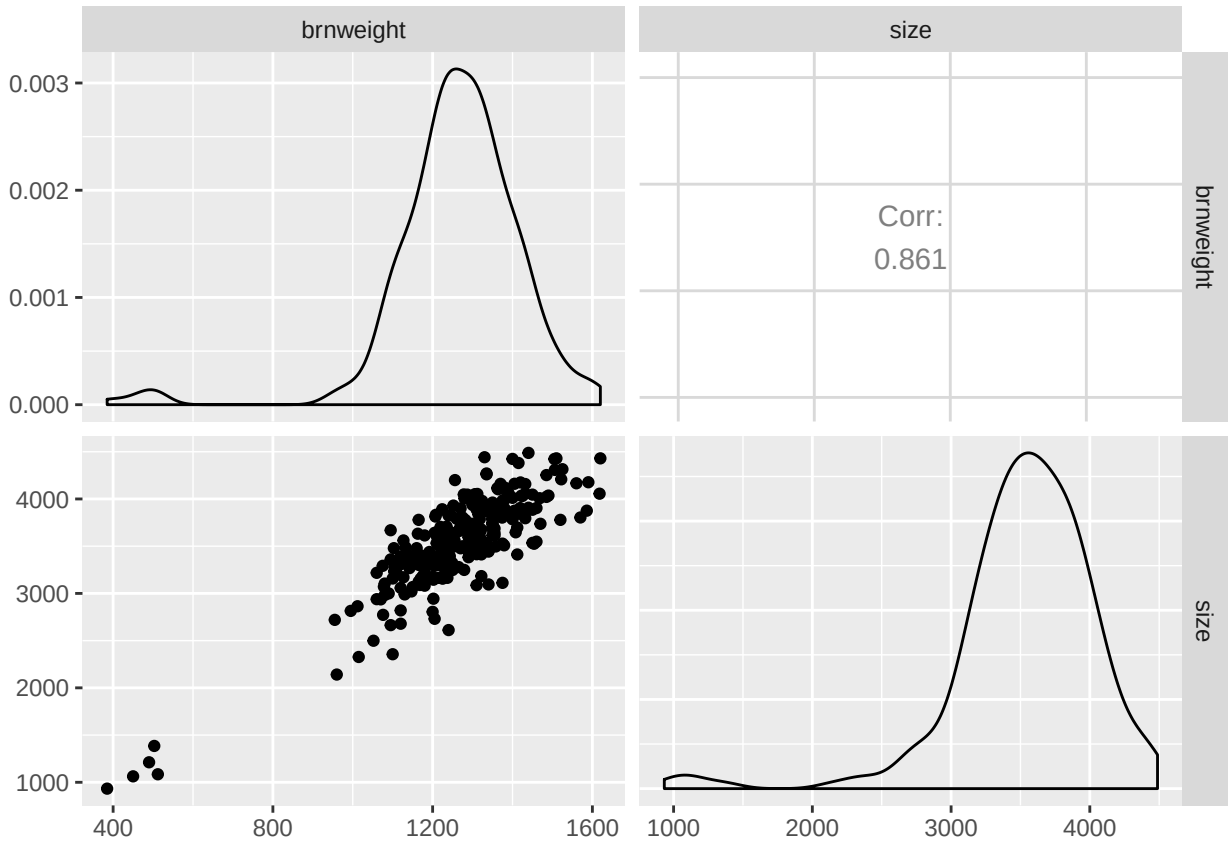
Se puede valorar que este dato pueden constituir un *errata*, en la medida que difieren bastante de la medida dada por la formula. Mirando el archivo de texto escaneado, efectivamente la obs `192` constituye una errata.

---

**(c)**

**Estudiar la regresión del peso del cerebro `brnweight` con la variable producto `size`. ¿Veis alguna dificultad?**

```
# gráfica de dispersion
cbrain %>%
    select(brnweight, size) %>%
    ggpairs()
```

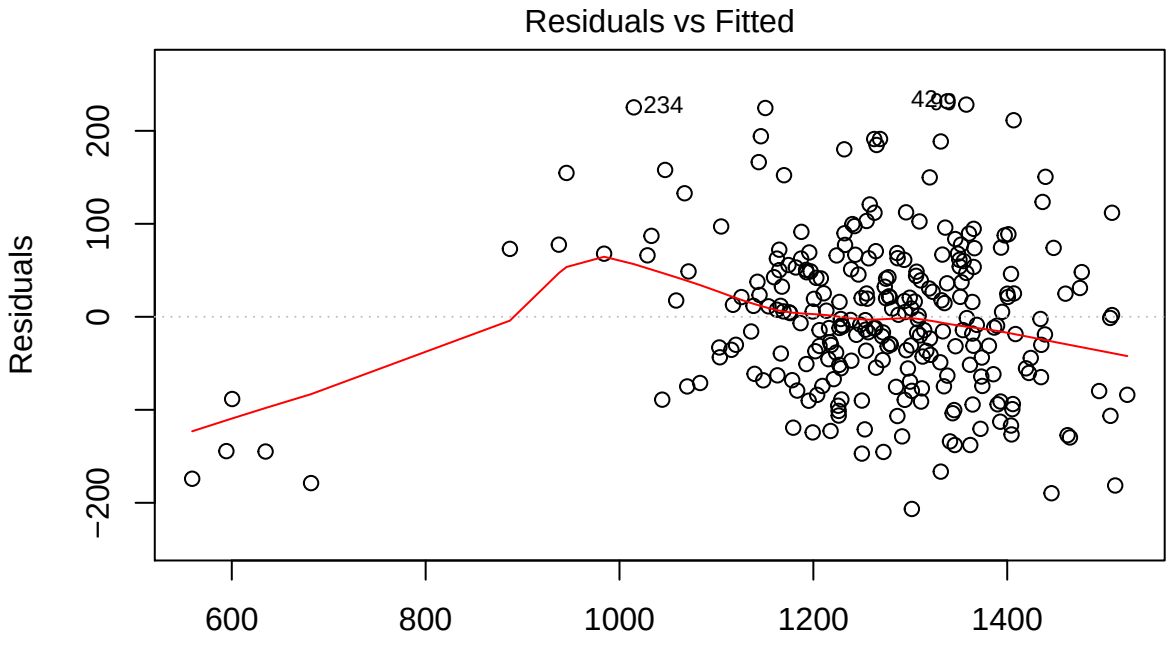


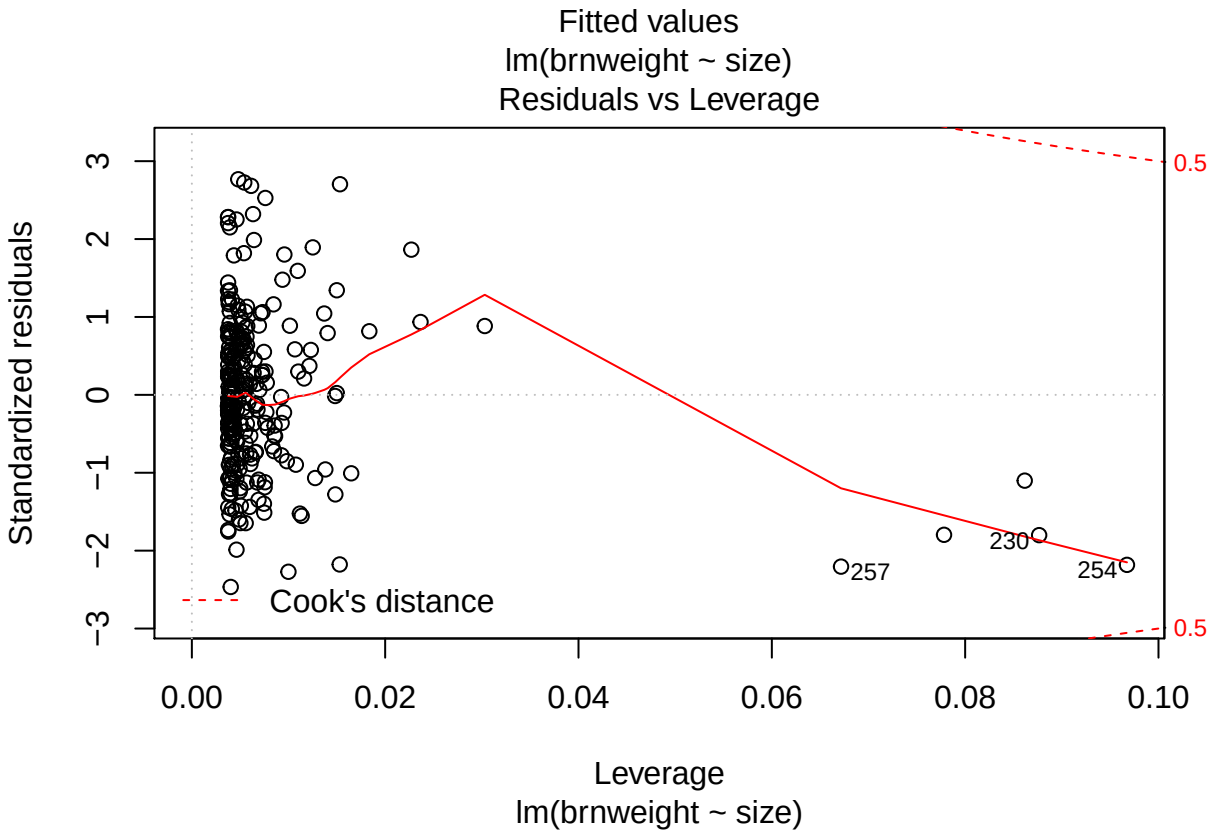

```
# modelo
model_w <- lm(brnweight ~ size, cbrain)
```

La dificultad es que si bien hay una correlación directa entre las dos variables, las variables tienen una distribución con un componente bimodal. Es probable que exitan varios puntos con *valores extremos* que puedan ejercer un efecto de *palanca (leverage)*

Mirando los datos estos corresponden a cerebros de la categoría `child` por lo que se podría asumir que en la muestra hay un número de cerebros muy pequeños ( y que pesan poco ) pero que distorcionan la distribución.

```
plot(model_w, c(1,5))
```



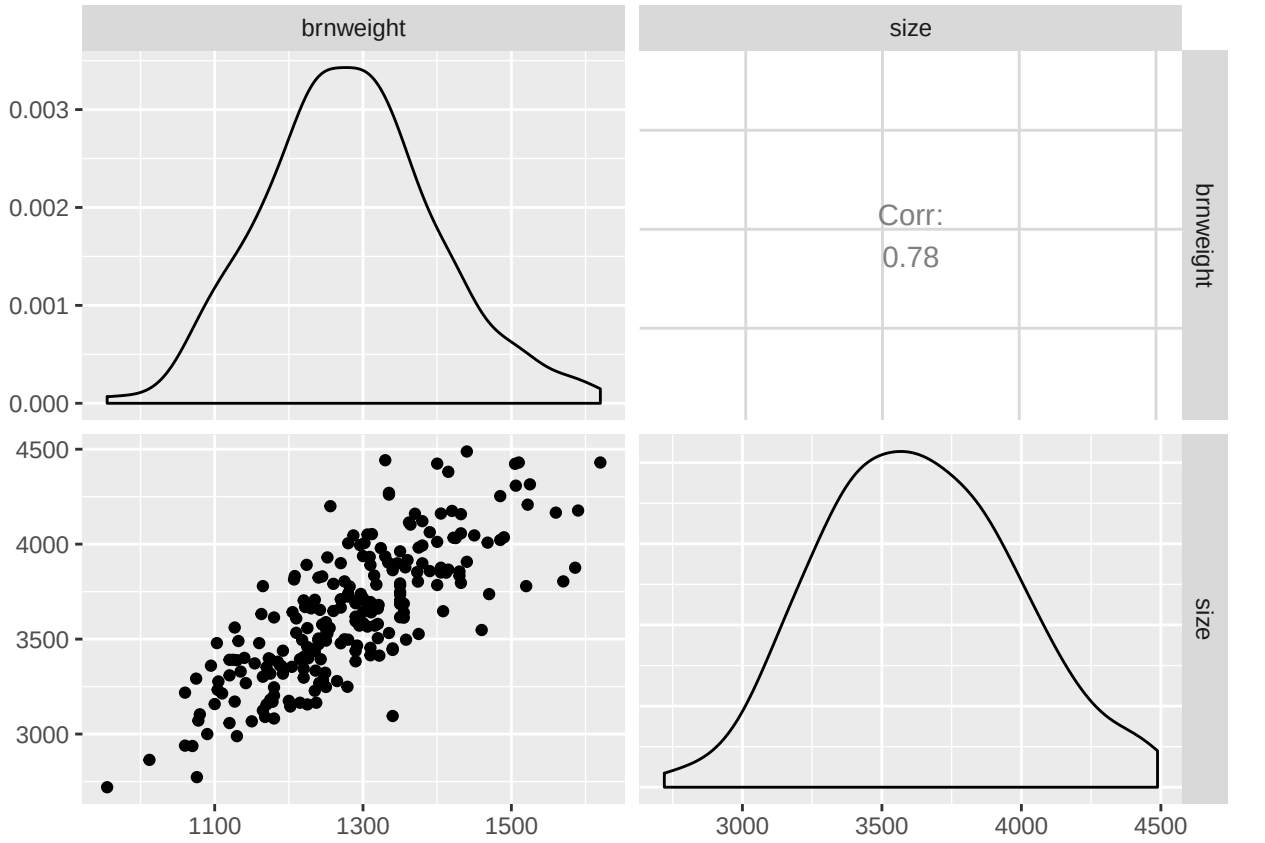

---

**(i)**

**Eliminar del estudio los individuos de menos de 20 años y repetir la regresión. ¿Mejora?**

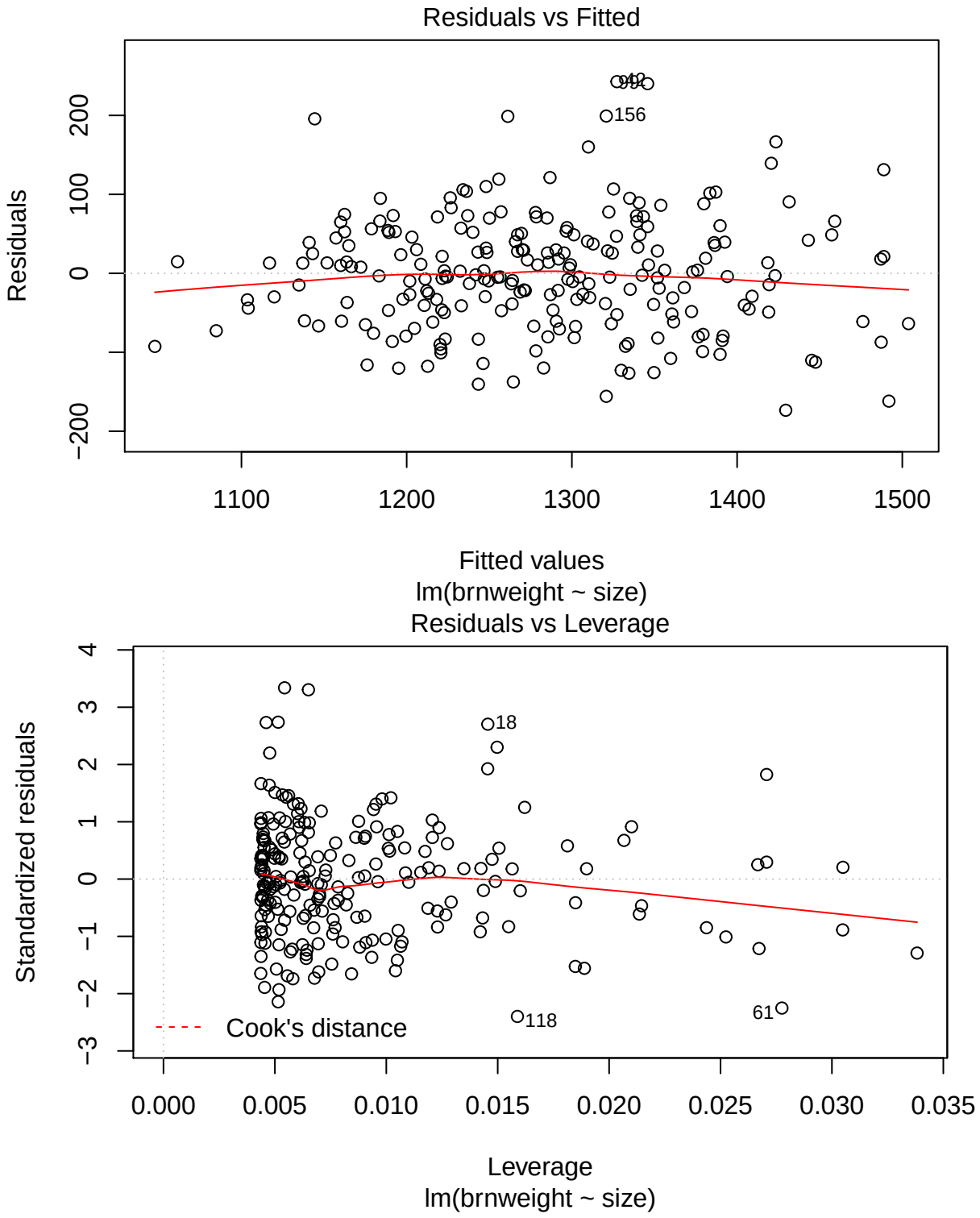
```r
ad_cbrain <-
    cbrain %>%
    filter(ageclass != 'child')

ad_cbrain %>%
    select(brnweight, size) %>%
    ggpairs()
```




```r
model_ad <- lm(brnweight ~ size, data = ad_cbrain)
```

```r
plot(model_ad, c(1,5))
```




El modelo con los mayores de 20 años mejora en la medidad de que gráfico de *fitted* v/s *residuals* se observa un patrón mucho mas homogeneocomo detereminado con el caso del modelo completo.

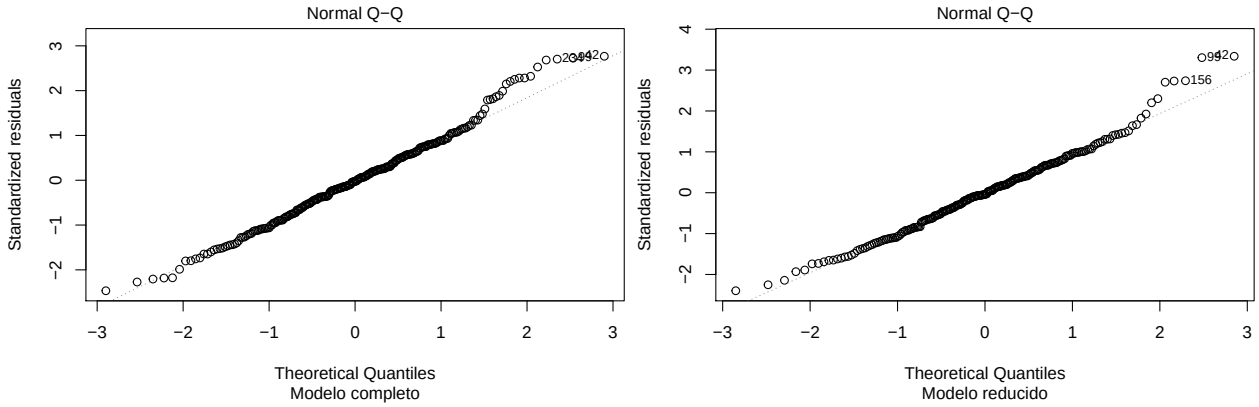
---

**(ii)**

**Estudiar la normalidad del error.**

Estudiamos la normalidad del error de amobos modelos; el modelo completo y el modelo reducido, sin los menores de 20 años.

```
plot(model_w, 2, sub = "Modelo completo")
plot(model_ad, 2, sub = "Modelo reducido")
```




Calculamos un test de Shapiro Wilk para los residuos de ambos modelos.

```
# modelo completo
shapiro.test(resid(model_w))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_w)
## W = 0.98885, p-value = 0.03756
```

```
# modelo reducido
shapiro.test(resid(model_ad))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_ad)
## W = 0.98642, p-value = 0.02814
```

POdemos estudiar la normalidad con un patrón grafico utilizando una grafica de Q-Q que grafica la distribucion de los residuales v/s los quantiles de una distribucion normal. La gráfica debiera acercarse de manera lo mas fidedigna a la recta diagonal para asumir normalidad.

Otra estrategia es una prueba estadistico, como el caso del test de Shapiro-Wilk.

En el caso de la opciño gráfica se puede ver como ambos modelos se aleja de la reacta, sobretodo en las *colas* de la distribución. La prueba de S-W por otra prte permite rechazar la H0 de normalidad en ambos casos. Con estos elementos podemos establcer que los residuos no siguen una distribucion normal en este modelo.

---

**(iii)**

**Dado que la variable `size` es un producto de tres variables, proponer una transformación potencia $h(x) = x$ que mejore su simetría o "normalidad".**

```r
# utilizar transformTukey para obtener lambda
(lambda <- transformTukey(ad_cbrain$size, quiet = T, plotit = F, returnLambda = T))
```

```
## lambda
##  0.225
```

```r
# if (lambda >  0){TRANS = x ^ lambda}
# if (lambda == 0){TRANS = log(x)}
# if (lambda <  0){TRANS = -1 * x ^ lambda}

# tranformar con lambda Finalmente, después de la debate en el foro, utilizo la variable reducida caluc
# (i). Creo que de esta forma el ejercicio queda mas coherente. La Transfromacion es mas efectiva y se
# documento al ejercicio 2
cbrain_tk <-
    ad_cbrain %>%
    mutate(h_size = case_when(
        lambda > 0 ~ size ^ lambda,
        lambda == 0 ~ log(size),
        lambda < 0 ~ -1 * size ^ lambda
    ))
```

**(iv)**

**Repetir la regresión con la variable size transformada y valorar el modelo.**

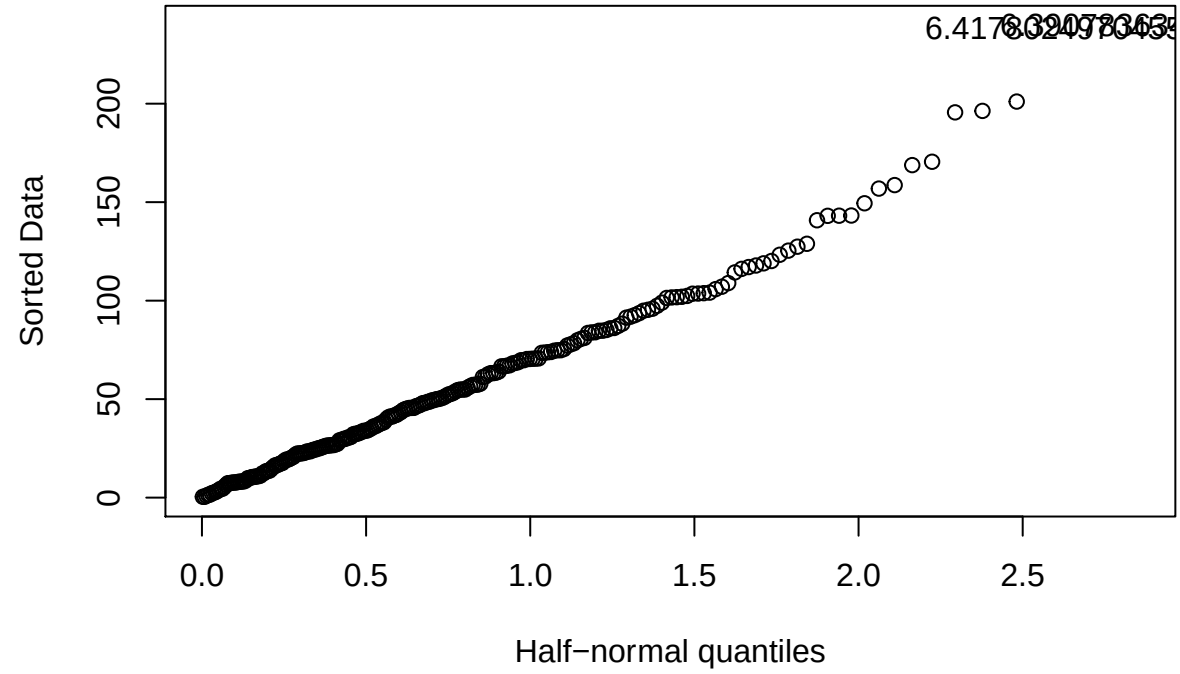
```r
model_tk <- lm(brnweight ~ h_size, data = cbrain_tk)

# resumen del modelo
summary(model_tk)
```

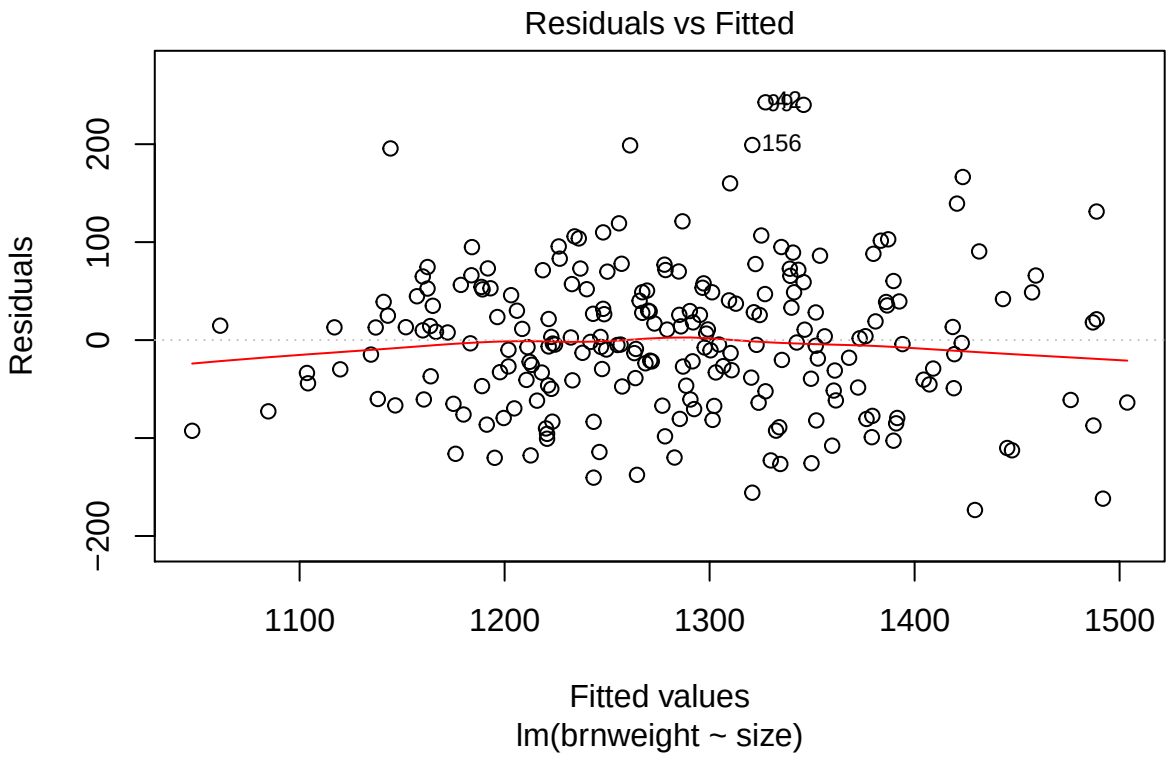
```
##
## Call:
## lm(formula = brnweight ~ h_size, data = cbrain_tk)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -168.806  -49.552   -2.793   45.562  240.090
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2881.67     220.40  -13.07   <2e-16 ***
## h_size        659.01      34.87   18.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72.62 on 227 degrees of freedom
## Multiple R-squared:  0.6114, Adjusted R-squared:  0.6097
## F-statistic: 357.2 on 1 and 227 DF,  p-value: < 2.2e-16
```

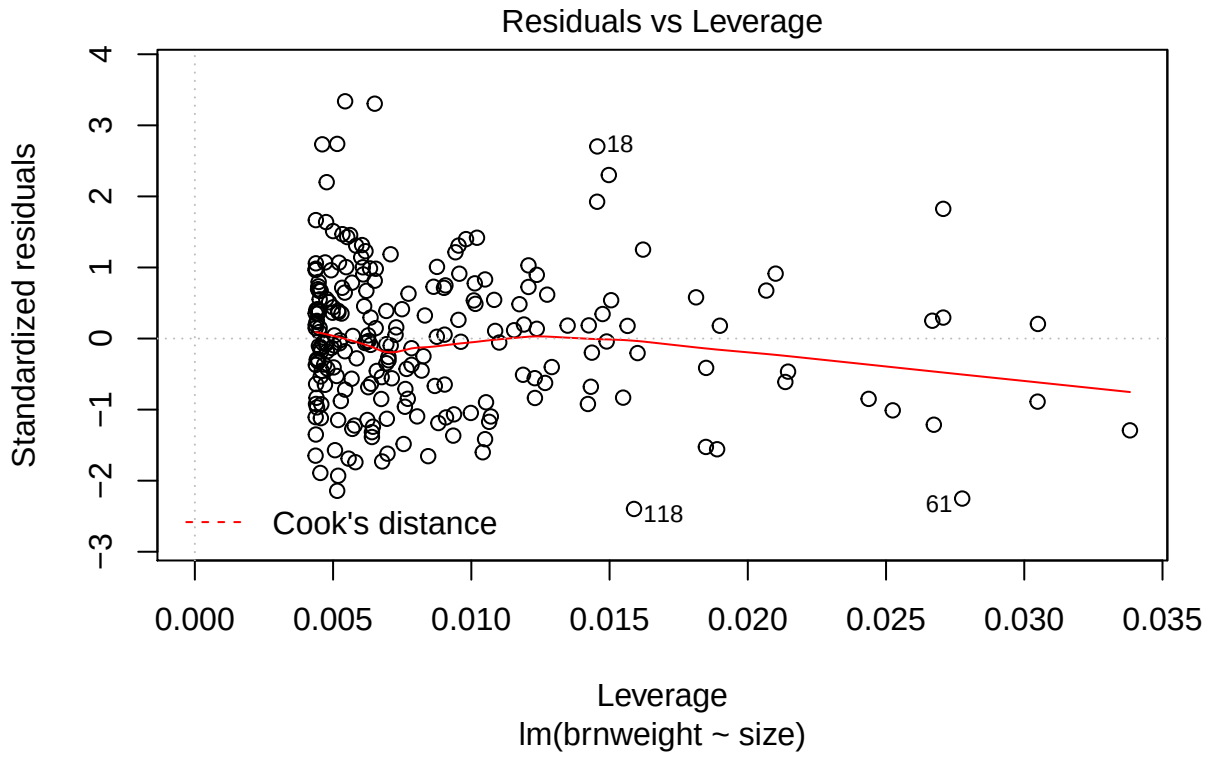
valoramos los valores extremos y los influyente del modelo y la distribucion de la varianza del error

```
halfnorm(resid(model_tk), labs = cbrain_tk$h_size)
```



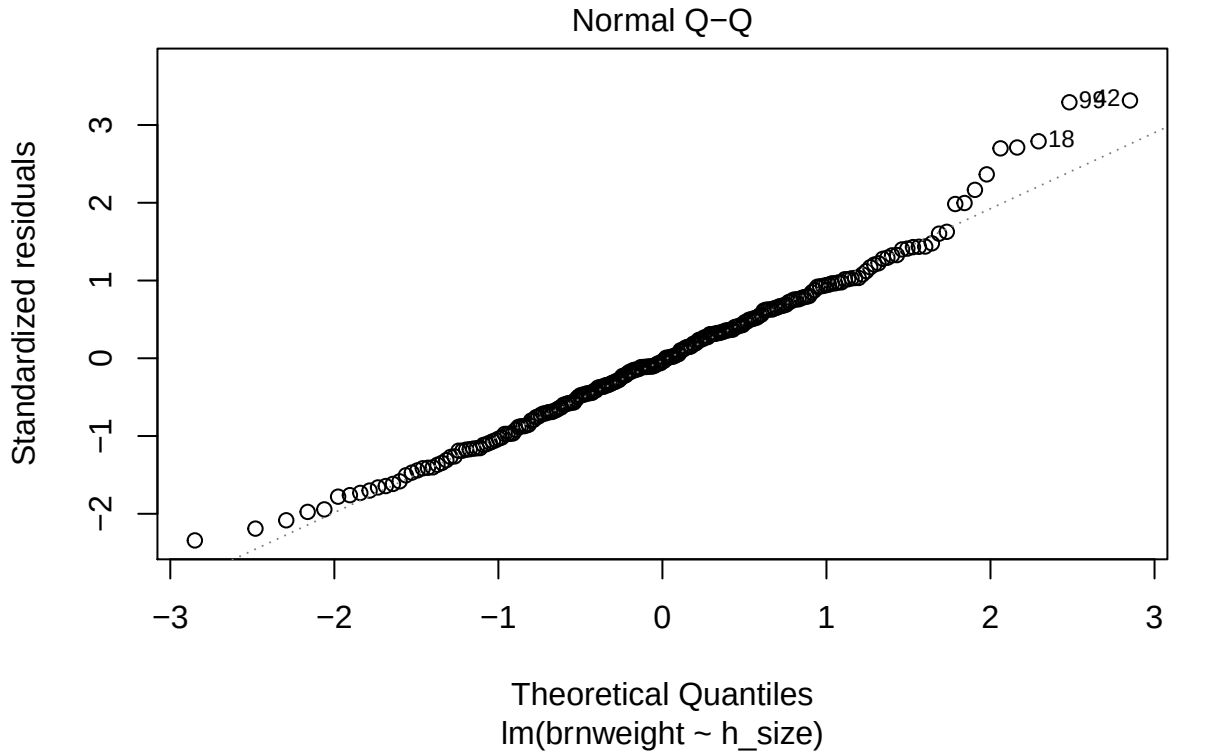

```
plot(model_ad, c(1,5))
```






Valoramos al distribucion del error del modelo

```
plot(model_tk, 2)
```




```
shapiro.test(resid(model_tk))
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  resid(model_tk)
## W = 0.98593, p-value = 0.02317
```

(v)

**Hallar el intervalo de confianza para la pendiente al 97 %.**

```r
# calcular IC al 97 %
confint(model_tk, level = 0.97) %>%
    as_tibble(rownames = "coeficiente")
```

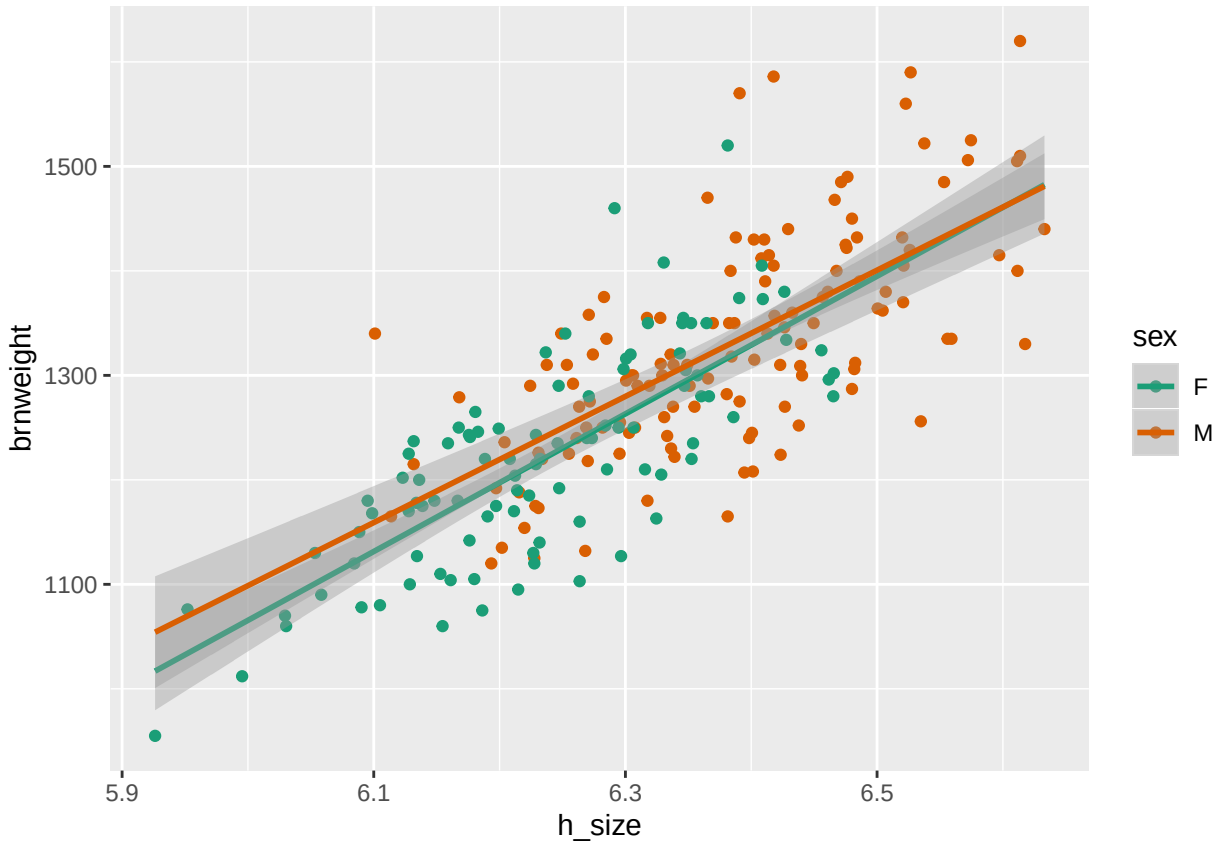
```
## # A tibble: 2 x 3
##   coeficiente `1.5 %` `98.5 %`
##   <chr>         <dbl>    <dbl>
## 1 (Intercept)  -3363.   -2400.
## 2 h_size         583.     735.
```

---

(d)

**Comparar las rectas de regresión que relacionan el peso del cerebro `brnweight` con la variable transformada h(`size`) para hombres y para mujeres sin los menores de 20 años. ¿Son paralelas? ¿Son iguales?**

```r
cbrain_tk %>%
    ggplot(aes(h_size, brnweight, color = sex)) +
    geom_point() +
    stat_smooth(method = "lm", fullrange = T) +
    scale_color_brewer(type = 'qual', palette = 2)
```




```
# modelo "extendido" para valorar pendientes
y <- cbrain_tk$brnweight
x.m <- c(cbrain_tk$h_size[cbrain_tk$sex == 'M'], rep(0,99))
x.f <- c(rep(0,130), cbrain_tk$h_size[cbrain_tk$sex == 'F'])

# contrastes
MM <- c(rep(1,130), rep(0,99))
FF <- c(rep(0,130), rep(1,99))

# modelo completo extendido ( sin interecepto porque agregamos contrastes )
mod_c <- lm(y ~ 0 + MM + FF + x.m + x.f)

# modelo con = pendiente
mod_ll <- lm(y ~ 0 + MM + FF + cbrain_tk$h_size)

# son las rectas paralelas?
anova(mod_ll, mod_c)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 0 + MM + FF + cbrain_tk$h_size
## Model 2: y ~ 0 + MM + FF + x.m + x.f
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    226 1185528
## 2    225 1183224  1    2303.8 0.4381 0.5087
```

Para valorar si las rectas son paralelas, en primer lugar volvemos a hacer el modelo completo pero en froma "extendida", es decir, sin interepceto por los términos para hacer los contrastes.

Al hacer el constraste de modelos, no se puede rechazar la H0, por lo que se acepta que las rectas son paralelas.

```
# modelo null contraste
mod_0 <- lm(y ~ cbrain_tk$h_size)

# son las rectas iguales?
anova(mod_0, mod_ll)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ cbrain_tk$h_size
## Model 2: y ~ 0 + MM + FF + cbrain_tk$h_size
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    227 1197072
## 2    226 1185528  1     11544 2.2007 0.1393
```

Podemos hacer un segundo contraste del modelo de rectas paralelas con el modelo *original* (reescrito para que la anova se lea mejor), para valorar si es que hay diferencias o no. La H0 es que los modelos no son diferentes y por las rectas paralelas del modelo `mod_ll` se pueden considerar que son coincidentes.

En este caso efectivamente no se puede rechazar la H0, por lo que se consideran las rectas como conicidentes.

```
# interacciones conh_size
mod_int <- aov(brnweight ~ h_size * sex, cbrain_tk )

# no intereacciones
mod_no <- aov(brnweight ~ h_size + sex, cbrain_tk)

# miramos si la interaccion es significativa
anova(mod_int, mod_no)
```

```
## Analysis of Variance Table
##
## Model 1: brnweight ~ h_size * sex
## Model 2: brnweight ~ h_size + sex
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    225 1183224
## 2    226 1185528 -1   -2303.8 0.4381 0.5087
```

Otra alternativa es hacer una ANCOVA para evaluar si la interaccion entre la variable `sex` es significativa para $Y$. MAtemáticamente es lo mismo que hemos hecho arriba para valorar si las pendientes son las mismas.

## Problema 2 (45 pt.)

**Con la base de datos `cbrain` del problema anterior sin los menores de 20 años, calcular el modelo de regresión que tiene como respuesta el peso del cerebro `brnweight` y como predictoras las variables `age`, `sex`, `height`, `headht`, `lenght`, `breadth`, `h(size)` y `circum`. La transformación $h()$ es la que hemos decidido en el apartado (c) del problema anterior.**

```
# modelo completo
model_full <-
    lm(brnweight ~ age + sex + height + headht + length + breadth + h_size + circum, data = cbrain_tk)

model_null <-
    lm(brnweight ~ 1, data = cbrain_tk)
```

**(a)**

**Escribir el modelo de regresión estimado e interpretar el coeficiente de la variable $h(`size`)$.**

```
# resumen del modelo
sumary(model_full)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  219.79756 5907.41537  0.0372 0.970354
## age           -1.19993    0.36218 -3.3130 0.001079
## sexM          11.87159   12.57849  0.9438 0.346307
## height         2.21332    1.67643  1.3203 0.188119
## headht        15.36173   30.98295  0.4958 0.620522
## length        10.24069   21.83820  0.4689 0.639581
## breadth       10.53461   27.71008  0.3802 0.704184
## h_size      -781.26607 2867.79328 -0.2724 0.785549
## circum         0.76562    0.52968  1.4454 0.149758
##
## n = 229, p = 9, Residual SE = 69.91662, R-Squared = 0.65
```

```
cfs <- round(coef(model_full),2)

# normalizamos la
abs(coef(model_full)["h_size"])^(1/lambda)
```

```
##       h_size
## 7.192573e+12
```

```
-781.26 ^ lambda
```

```
##    lambda
## -4.475886
```

$$y_i = 219.8 + -1.2 * age + 11.87 * sexM + 2.21 * height + 15.36 * headht + 10.24 * length + 10.53 * breadth + -781.27 * (h) size + 0.77 * circu$$

El coeficiente de h_size. por cada unidad de h_size , baja -781.26 el brnweight, o -4.47 rl brnweight ^ 1/lambda la unidad de h_size es size ^ 1/lambda.

no se pueden establcer muchas conclusioens ya que los coefiecientes son en contexto de los otras varaibles.

**(b)**

**Utilizar un test F para determinar la significación de la regresión del modelo. ¿Qué significa esto último?**

```
anova(model_null, model_full)
```

```
## Analysis of Variance Table
##
## Model 1: brnweight ~ 1
## Model 2: brnweight ~ age + sex + height + headht + length + breadth +
##     h_size + circum
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    228 3080547
## 2    220 1075433  8   2005114 51.273 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La funcion `summary` entrega los resultados de el test de F de la regresion del modelo. El test de hace una prueba de hipotesis entre el modelo completo y la $H0 : \beta_i = 0$. En este caso la prueba permite reachazar $H0$ en la medida que el estadistico $F$ que refleja la diferencia de las RSS, es significativamente diferente del de la H0

**(i)**

**¿Qué predictoras son significativas al 5 %?**

```
tidy(model_full) %>%
    filter(p.value < 0.05)
```

```
## # A tibble: 1 x 5
##   term  estimate std.error statistic p.value
##   <chr>    <dbl>     <dbl>     <dbl>   <dbl>
## 1 age      -1.20     0.362     -3.31 0.00108
```

En este caso el único predictor significativo al 5% es `age`.

**(ii)**

**¿Quiere esto decir que podemos eliminar del model las no significativas al 5 %?**

En el modelo completo, el único valor que tiene un valor significativo es `age`. Esto *no* quiere decir que se pueda precindir de los otros predictores, ya que el modelo completo que **si** resultaba explicativo, toma en cuenta *todos* los predictores y no sólo `age`. Para valorar *que* predictores tienen mayor o menor peso en el modelo de regresión se debe hacer uno a uno, comparando con el modelo completo, o utilizando una aproximación *stepwise*.

**(iii)**

**Calcular la matriz de correlaciones entre las variables continuas del modelo de regresión.**

```
cor_full <-
    augment(model_full) %>%
    select(brnweight:circum, -sex) %>%
    cor()

cor_full %>%
    as_tibble(rownames = "variables")
```

```
## # A tibble: 8 x 9
##   variables brnweight     age  height headht  length breadth h_size  circum
##   <chr>         <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl>  <dbl>   <dbl>
## 1 brnweight     1     -0.214   0.505   0.673  0.675   0.613   0.782  0.714
## 2 age          -0.214  1      -0.0793 -0.206 -0.0221 -0.0409 -0.105 -0.0229
## 3 height        0.505 -0.0793  1       0.461  0.446   0.435   0.535  0.568
## 4 headht        0.673 -0.206   0.461   1      0.546   0.493   0.805  0.621
## 5 length        0.675 -0.0221  0.446   0.546  1       0.584   0.843  0.804
## 6 breadth       0.613 -0.0409  0.435   0.493  0.584   1       0.849  0.767
## 7 h_size        0.782 -0.105   0.535   0.805  0.843   0.849   1      0.878
## 8 circum        0.714 -0.0229  0.568   0.621  0.804   0.767   0.878  1
```

```
# grafica de correlacion con corrplot
corrplot.mixed(cor_full)
```


La matriz de correlaciones (y su version gráfica), permiten ver que varias de las variables tienen grados de correlación relativamente altos.

**(iii)**

**¿Cuales son las variables más correlacionadas? ¿Y las que menos? ¿Concuerdan con los resultados del modelo de regresión?**

Con el ejercicio anterior es facil de ver que las variables que mas correlacionan son, `h_size` con `circum`, `breadth`, `length` y `headht`, y `length` y `circum`.

Por otra parte todas las variables de medidas correlacions de manera positiva con la variable independiente del modeo, `brnweight` lo que es coherente con los resultados y hace ver que es un modelo con mucha colinearidad lo que puede ser probelmático.

**(c)**

**Contrastar si podemos aceptar un modelo más simple sin las variables `headht`, `lenght`, `breadth`.**

```
model_red <-
    lm(brnweight ~ age + sex + height + h_size + circum, data = cbrain_tk)

anova(model_red, model_full)
```

```
## Analysis of Variance Table
##
## Model 1: brnweight ~ age + sex + height + h_size + circum
## Model 2: brnweight ~ age + sex + height + headht + length + breadth +
##     h_size + circum
```

```
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    223 1097487
## 2    220 1075433  3     22054 1.5038 0.2144
```

En este caso la H0 es que no hay diferencias entre los dos modelos. En este caso la prueba resulta no significativa por lo que no se puede rechazar la H0 y por criterio de parsimonia se puede elegir el modelo mas simple

### (d)

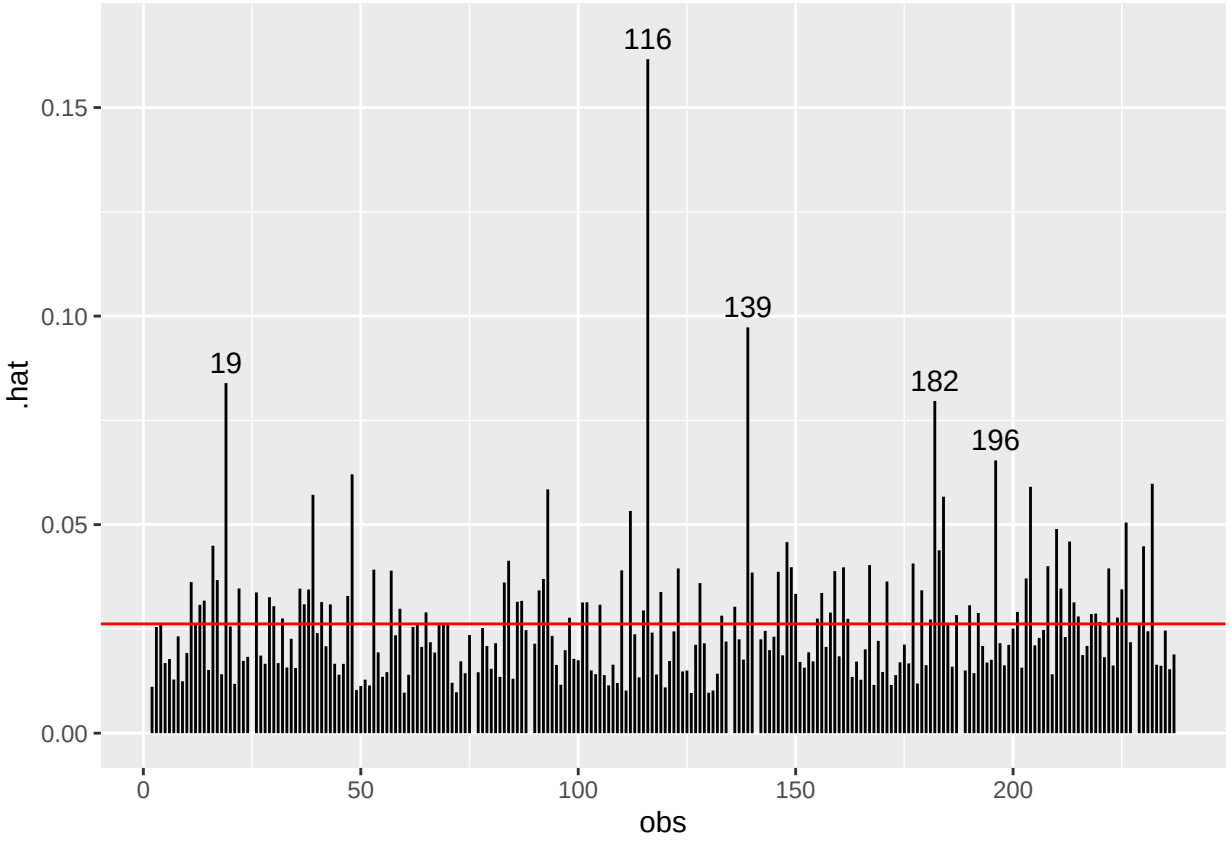
**En el modelo más simple del apartado anterior, ¿hay alguna observación con un alto leverage? ¿Y con una gran influencia? Dibujar los gráficos oportunos para explicar el resultado.**

Primero miramos leverage

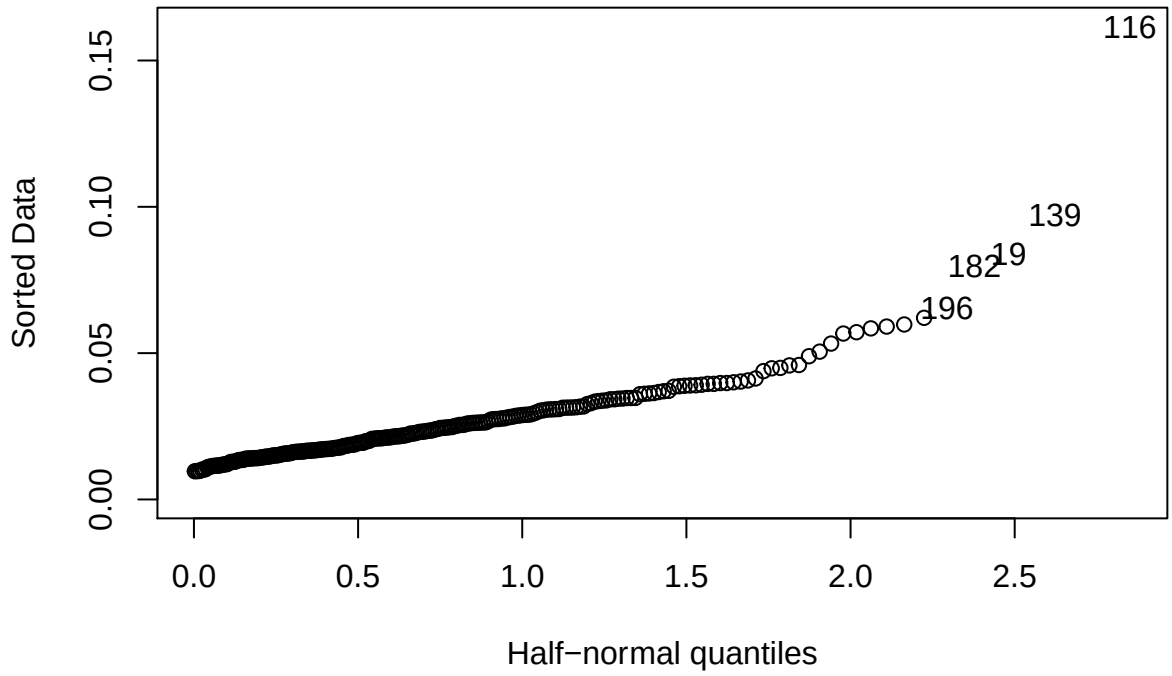
```
n <- dim(model.matrix(model_red))[1]
p <- dim(model.matrix(model_red))[2]
lev.mean <- p/n

df_model_red <-
    augment(model_red) %>%
    mutate(obs = as.integer(ad_cbrain$obs))

# grafica de hat values
ggplot(df_model_red, aes(obs, .hat)) +
    geom_segment(aes(yend = 0, xend = obs)) +
    geom_hline(yintercept = lev.mean, color = "red") +
    geom_text(data = top_n(df_model_red, 5, .hat), aes(label = obs), nudge_y = .005)
```



```r
# halfnormal plot de Faraway
halfnorm(hatvalues(model_red), nlab = 5, labs = cbrain$obs)
```

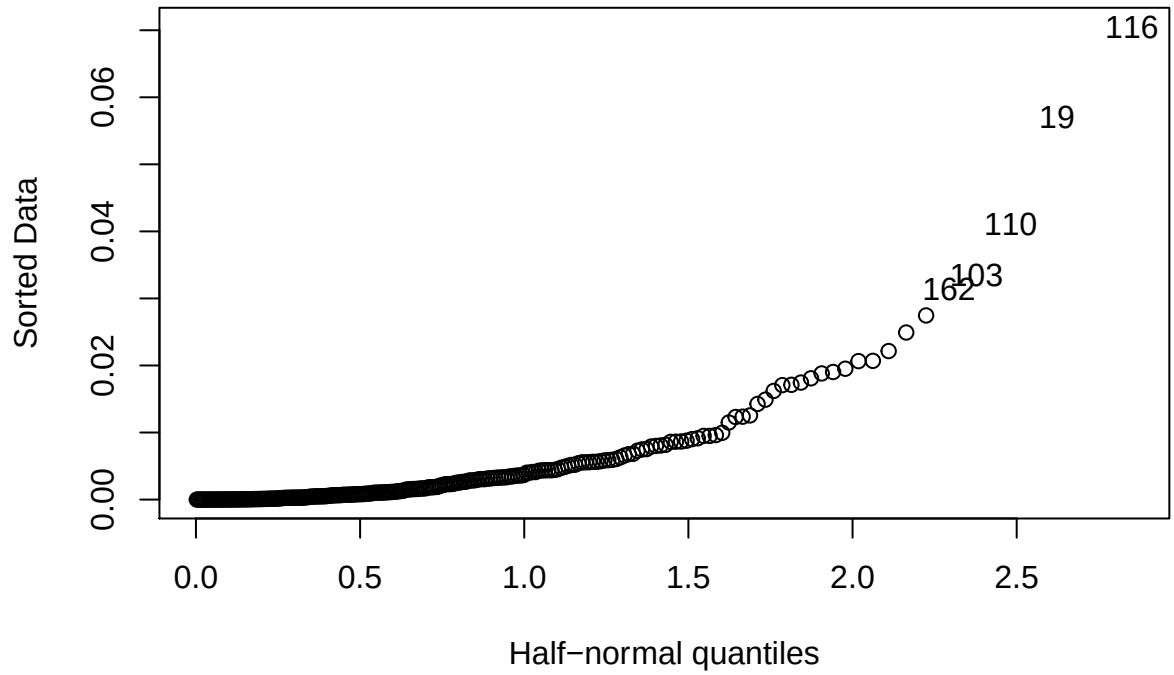



```r
# las top 5
df_model_red %>%
    top_n(5, .hat) %>%
    select(obs, .hat, brnweight:circum) %>%
    arrange(-.hat)
```

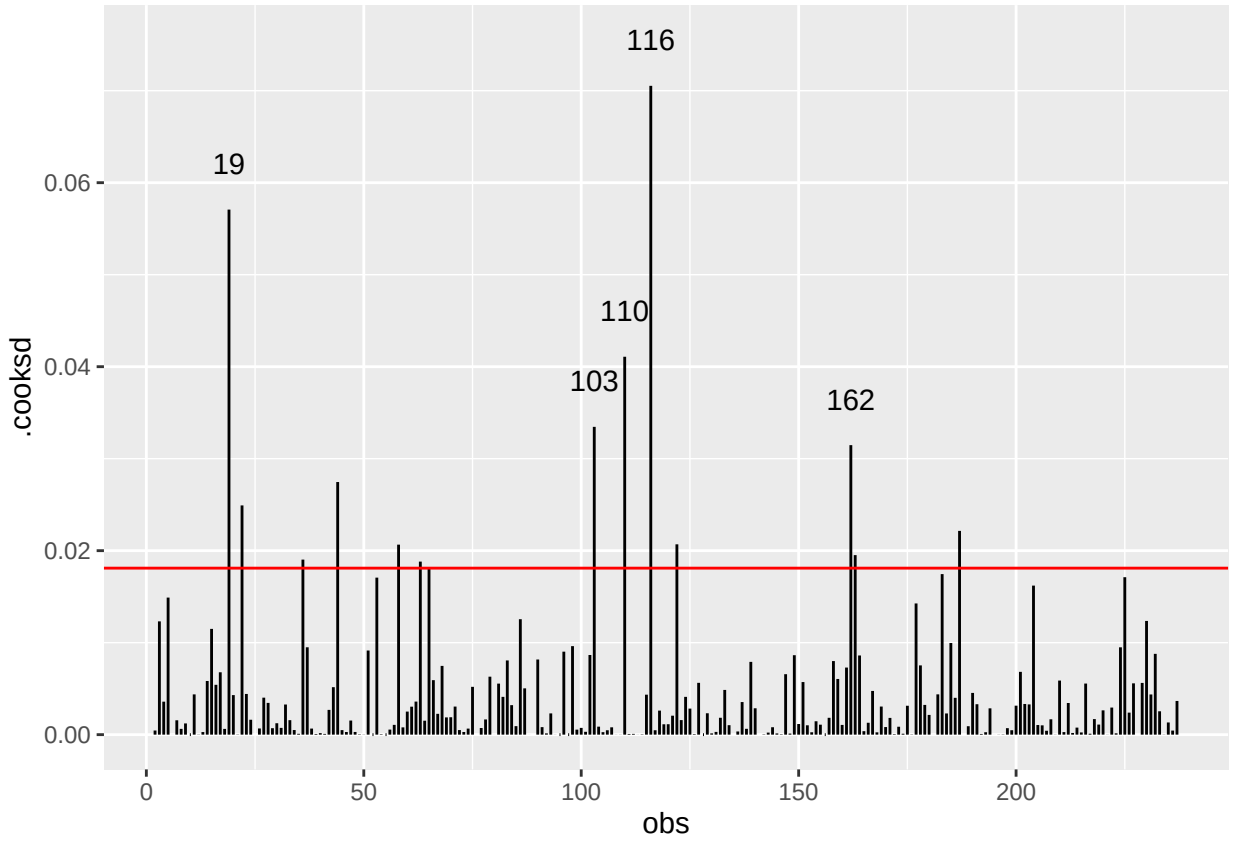
```
## # A tibble: 5 x 8
##     obs   .hat brnweight   age sex   height h_size circum
##   <int>  <dbl>     <int> <dbl> <fct>  <dbl>  <dbl>  <dbl>
## 1   116 0.162       1275    84 M       50.5   6.27   545
## 2   139 0.0973      1080    30 F       51     6.10   498.
## 3    19 0.0839      1340    20 M       71.5   6.10   517
## 4   182 0.0797      1100    33 F       55     6.13   490
## 5   196 0.0654      1076    47 F       66.5   5.95   516
```

Inlfuencia

```r
halfnorm(cooks.distance(model_red), nlab = 5 ,labs = ad_cbrain$obs)
```




```
# grafica de hat values
ggplot(df_model_red, aes(obs, .cooksd)) +
    geom_segment(aes(yend = 0, xend = obs)) +
    geom_hline(yintercept = 4/(n-p-2), color = "red") +
    geom_text(data = top_n(df_model_red, 5, .cooksd), aes(label = obs), nudge_y = .005)
```



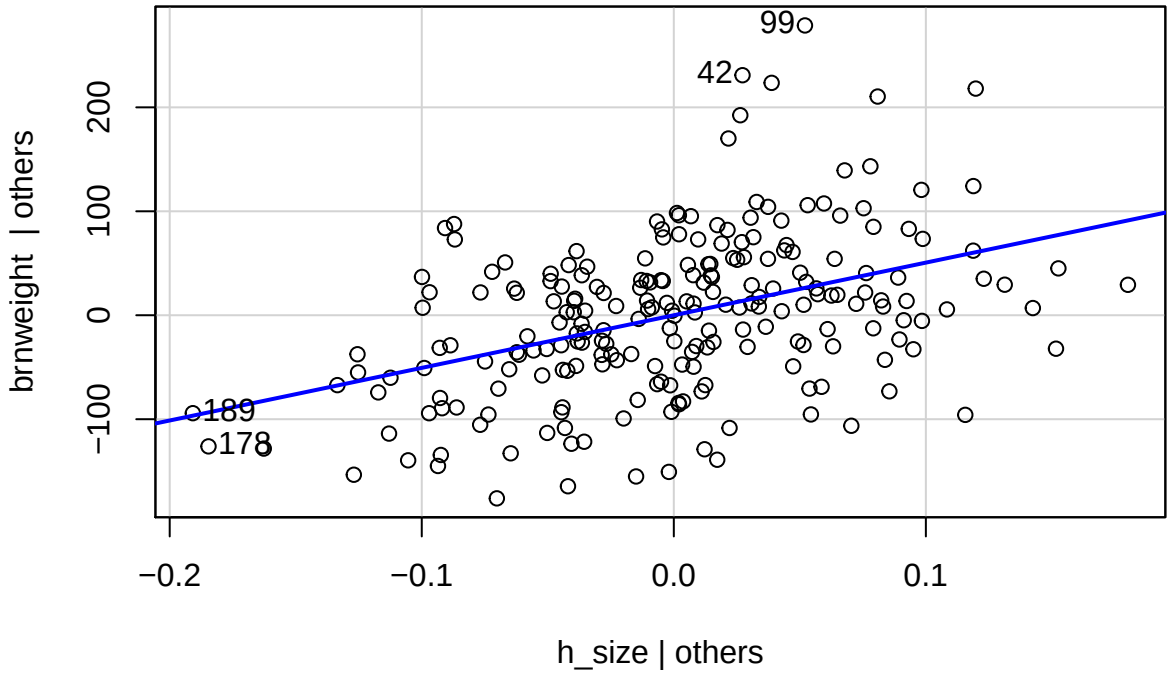
```
# los top 5
df_model_red %>%
    select(obs, .cooksd, brnweight:circum) %>%
    top_n(5, .cooksd) %>%
    arrange(-.cooksd)
```

```
## # A tibble: 5 x 8
##     obs .cooksd brnweight   age sex   height h_size circum
##   <int>   <dbl>     <int> <dbl> <fct>  <dbl>  <dbl>  <dbl>
## 1   116  0.0705      1275    84 M       50.5   6.27    545
## 2    19  0.0571      1340    20 M       71.5   6.10    517
## 3   110  0.0411      1620    65 M       70.5   6.61    585
## 4   103  0.0335      1586    60 M       68     6.42    555
## 5   162  0.0315      1520    25 F       67     6.38    550
```

(e)

**Con el modelo simple del apartado (c), dibujar el gráfico de regresión parcial con la variable predictora `h(size)`.**

```
car::avPlots(model_red, terms = ~ h_size)
```




(f)

**Hallar el intervalo de confianza para la predicción del peso del cerebro, por el modelo simple, cuando un individuo toma los valores de la observación 5 de la base de datos.**

```
ind_5 <- cbrain_tk %>%
    filter(obs == 5) %>%
    select(age, sex, height, h_size, circum)
```

```r
predict(model_red, newdata = ind_5, interval = "conf")
```

```
##       fit      lwr      upr
## 1 1430.93 1413.005 1448.855
```

### (g)

**Estimar la varianza del error y calcular su intervalo de confianza al 99 % en el modelo de regresión más simple del apartado (c).**

Asumimos que la variaza del error se define por el estimador:

$\sigma^2 = \sqrt{MSE}$

$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}$

es posible estimar el intervalo de confianza utilizando la distirbución de chi cuadrado con un nivel de confianza del 99%

```r
n <- length(model_red$residuals)
p <- length(model_red$coefficients)

# calculamos manualmente la varianza del error
MSE <- (sum((ad_cbrain$brnweight - fitted(model_red))^2)) / (n-p)
sqrt(MSE)
```

```
## [1] 70.15318
```

```r
# que coincide con el cálculo dado por el modelo lineal
(S <- glance(model_red)$sigma)
```

```
## [1] 70.15318
```

```r
# calculamos los límites inferior y superior para un IC del 99%
c(inf = (S * (n-p)) / qchisq(0.995, (n-p)), S, sup = (S * (n-p)) / qchisq(0.005, (n-p)))
```

```
##      inf                sup
## 55.64417 70.15318 90.76467
```

## Problema 3 (25 pt.)

**En los ejemplos 5.3.2 y 5.6.3 del libro de Carmona y con los datos de la tabla 5.2 se ha estudiado el diseño crossover simplificado. En este problema vamos a considerar un diseño un poco más sofisticado y con más parámetros. En la página https://newonlinecourses.science.psu.edu/stat509/node/123/ se explican los diseños crossover con todo detalle. Nosotros nos vamos a centrar en el caso 2»2 donde a dos grupos de pacientes se subministran dos fármacos o se realizan dos tratamientos de forma consecutiva en dos períodos de tiempo. En esta situación, los parámetros a considerar son: media general, efecto del fármaco $A$, efecto del fármaco $B$ , $_1$ efecto del grupo 1 o secuencia $AB$, $_2$ efecto del grupo 2 o secuencia $BA$, $_1$ efecto del primer período, $_2$ efecto del segundo período, $_A$ o efecto de arrastre de $A$(first-ordercarryover effect) y $_B$ o efecto de arrastre de $B$. En total 9 parámetros. Como trabajar con 9 parámetros es complicado, en la tabla Design 11 dehttps://newonlinecourses.science.psu.edu/stat509/node/127/nos proponen un conjunto más reducido de la siguiente forma:**

### (a)

**Escribir la matriz de diseño reducida del modelo propuesto en la tabla anterior y calcular su rango. La matriz reducida tendrá 4 filas, una por cada situación experimental, y 6 columnas,**

**una por cada parámetro. El rango representa el número efectivo de parámetros.**

En primer lugar se describen las 4 situaciones experimentales con los parámetros. (El orden es el que luego utilizo para en el apartado d)

$$y_{A1} = \mu_A + \nu + \rho y_{B2} = \mu_B + \nu - \rho + \lambda_A y_{B1} = \mu_B - \nu + \rho y_{A2} = \mu_A - \nu - \rho + \lambda_B$$

Con estas situaciones se puede crear la matriz de diseño:

```
mat <-
     matrix(
         c(1, 0, 1, 1, 0, 0,
           0, 1, 1,-1, 1, 0,
           0, 1,-1, 1, 0, 0,
           1, 0,-1,-1, 0, 1),
         nrow = 4,
         byrow = T
     )

colnames(mat) <- c('mu_A', 'mu_B', 'nu', 'rho', 'lambda_A', 'lambda_B' )
rownames(mat) <- c('A1', 'B2', 'B1' , 'A2')
mat
```

```
##    mu_A mu_B nu rho lambda_A lambda_B
## A1    1    0  1   1        0        0
## B2    0    1  1  -1        1        0
## B1    0    1 -1   1        0        0
## A2    1    0 -1  -1        0        1
```

**(b)**

**Como el objetivo es contrastar si el efecto de los fármacos $A$ y $B$ se puede considerar similar, la hipótesis paramétrica a contrastar es $H0: =$ que es equivalente $H0: _A = _B$ con los nuevos parámetros. Probar que esta hipótesis NO es contrastable con la matriz de diseño del apartado anterior.**

La hipótesis nula en forma vectorial es:

$$H0: \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu A \\ \mu B \\ \nu \\ \rho \\ \lambda_A \\ \lambda_B \end{pmatrix} = 0$$

Para valorar si la Hipótesis es contrastable utilizamos el método de añadir la fila de la hipótesis a la del diseño experimental y ver si cambia o no el rango de la matriz. Si el rango es diferente, la Hipotesis **no** es una combinación lineal de la matriz experiemtnal y por tanto la Hipótesis **no** es constrastable.

```
# vector de Hipotesis nula
A <- c(1,-1,0,0,0,0)

(rank_m <- qr(mat)$rank)
```

```
## [1] 4
```

```r
(rank_h0 <- qr(rbind(mat, A))$rank)
```

```
## [1] 5
```

```r
# El rango es igual? Las hipótesis son contrastables?
rank_m == rank_h0
```

```
## [1] FALSE
```

**(c)**

**Reducir el número de parámetros de forma que $A = B =$ , es decir, el efecto de arrastre de A es igual al efecto de arrastre de B. Probar que la hipótesis principal de igualdad de efectos de los fármacos para este nuevo diseño SI es contrastable.**

Las situaciones experiemtnales quedan así:

$$y_{A1} = \mu_A + \nu + \rho y_{B2} = \mu_B + \nu - \rho + \lambda y_{B1} = \mu_B - \nu + \rho y_{A2} = \mu_A - \nu - \rho + \lambda$$

Con este modelo reducido el contraste de hipotesis es:

$$H0 : \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu A \\ \mu B \\ \nu \\ \rho \\ \lambda \end{pmatrix} = 0$$

Se construye la matriz del diesño reducido:

```r
mat_2 <-
    matrix(
        c(1, 0, 1, 1, 0,
          0, 1, 1,-1, 1,
          0, 1,-1, 1, 0,
          1, 0,-1,-1, 1),
        nrow = 4,
        byrow = T
    )
colnames(mat_2) <- c('mu_A', 'mu_B', 'nu', 'rho', 'lambda' )
rownames(mat_2) <- c('A1', 'B2', 'B1' , 'A2')
mat_2
```

```
##    mu_A mu_B nu rho lambda
## A1    1    0  1   1      0
## B2    0    1  1  -1      1
## B1    0    1 -1   1      0
## A2    1    0 -1  -1      1
```

```r
# vector de Hipotesis nula
A <- c(1, -1, 0, 0, 0)

(rank_mr <- qr(mat_2)$rank)
```

```
## [1] 4
```

```r
(rank_h0 <- qr(rbind(mat_2, A))$rank)
```

```
## [1] 4
```

```
# El rango es igual? Las hipótesis son contrastables?
rank_m == rank_h0
```

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

### (d)

**Contrastar la hipótesis $H0 : A = B$ con el modelo del apartado anterior y los datos de la tabla 5.2 del libro de Carmona. En un cierto orden, los datos de la variable respuesta en R son**

```
Y <- c(17, 34, 26, 10, 19, 17, 8, 16, 13,11,
       17, 41, 26, 3, -6, -4, 11, 16, 16,4,
       21, 20, 11, 26, 42, 28, 3, 3, 16, -10,
       10, 24, 32, 26, 52, 28, 27, 28, 21, 42)

# matriz experimental de X
X <-
    tibble(
        mu_a = c(sapply(mat_2[, 1], function(x)
            rep(x, 10))),
        mu_b = c(sapply(mat_2[, 2], function(x)
            rep(x, 10))),
        nu = c(sapply(mat_2[, 3], function(x)
            rep(x, 10))),
        rho = c(sapply(mat_2[, 4], function(x)
            rep(x, 10))),
        lambda = c(sapply(mat_2[, 4], function(x)
            rep(x, 10)))
    )

# modelo completo
full <- lm( Y ~ 0 + mu_a + mu_b + nu + rho + lambda, X)

# modelo sin los contrastes para valorar mu_a = mu_b
null <- lm( Y ~ 0 + nu + rho + lambda, X)

# contraste de modelos
anova(null, full)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ 0 + nu + rho + lambda
## Model 2: Y ~ 0 + mu_a + mu_b + nu + rho + lambda
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     38 20206.2
## 2     36  5547.3  2     14659 47.565 7.848e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para contrastar la $H0 : \mu_A = \mu_B$, primero creamos los vectores de los contrastes con el modelo reducido. POsterioremtne se construyen dos modelos; el modelo con todos los parámetros y el modelo que exluye los parámetros $\mu_A$ y $\mu_B$. Finalmente se realiza un contraste entre los dos modelos.

El contraste resulta en que se puede rechazar la H0, por lo que se puede asumir que las variables de tratamiento A y B no son iguales.