Regresión Lineal

Edmond Géraud

Estimación del peso estándar del hígado

- gender
- weight [kg]
- height [cm]
- liver_weight [g]
- liver_volume [ml]

Figure 1: Intervención de cirugía hepática mayor por laparascopia

Necesitamos cargar las librerías

```
if (!(require(car))) install.packages("car", dep=TRUE)
```

```
Loading required package: car
```

```
Loading required package: carData
```

```
if (!(require(DescTools))) install.packages("DescTools",
                                            dependencies = T)
```

```
Loading required package: DescTools
```

```
Attaching package: 'DescTools'
```

```
The following object is masked from 'package:car':

    Recode
```

```
if (!(require(faraway))) install.packages("DescTools",
                                            dependencies = T)
```

```
Loading required package: faraway
```

```
Attaching package: 'faraway'
```

```
The following objects are masked from 'package:car':

    logit, vif
```

```
ruta <- "./data/chan_data.csv"
datos <- read.csv(ruta)
```

```r
class(datos)
```

```
[1] "data.frame"
```

```r
str(datos)
```

```
'data.frame':   158 obs. of  5 variables:
 $ gender      : chr  "F" "F" "F" "F" ...
 $ weight      : num  50.3 47.4 44.3 44.1 52.1 51.3 51.8 42 52.1 44.1 ...
 $ height      : num  152 151 155 159 175 ...
 $ liver_weight: num  596 635 641 645 669 ...
 $ liver_volume: num  697 759 890 790 818 ...
```

Cálculo de BMI y BSA (body surface area)

```r
logBSA <-
log(0.007184) + 0.425 * log(datos$weight) + 0.725 * log(datos$height)
datos$BSA <- exp(logBSA)
datos$BMI <- datos$weight/(datos$height/100)^2
str(datos)
```

```
'data.frame':   158 obs. of  7 variables:
 $ gender      : chr  "F" "F" "F" "F" ...
 $ weight      : num  50.3 47.4 44.3 44.1 52.1 51.3 51.8 42 52.1 44.1 ...
 $ height      : num  152 151 155 159 175 ...
 $ liver_weight: num  596 635 641 645 669 ...
 $ liver_volume: num  697 759 890 790 818 ...
 $ BSA         : num  1.45 1.41 1.39 1.42 1.63 ...
 $ BMI         : num  21.7 20.7 18.4 17.4 17 ...
```

¿Qué es necesario saber de la OLS y MLS?

1. $$Y = X\beta + \epsilon$$

   - La Y es la variable respuesta dependiente
   - La X es/son las variables independientes

- La ϵ es el error

2. Supuestos

   - $\epsilon \sim N$
   - Linealidad: Al graficar no es una parábola por ejemplo
   - Independencia, las variables deben ser independientes entre sí
   - Homocedasticidad´

Estudiemos la normalidad de las variables

Es una buena práctica realizar dichos análisis por motivos del modelo, aunque no se cumplan la normalidad de las variables, es importante, que una vez hecho el modelo, los residuos, es decir la diferencia entre la variable respuesta y la predecida.

Lo podemos realizar de dos maneras, realizar un bucle for, o un apply

```
p.values <- vector("numeric",length=ncol(datos)-1)
for(i in 2:ncol(datos)){

  print(paste(round(JarqueBeraTest(datos[,i])$p.value,4),colnames(datos)[i]))

}
```

```
[1] "0.0824 weight"
[1] "0.1861 height"
[1] "0.0056 liver_weight"
[1] "0.0505 liver_volume"
[1] "0.0718 BSA"
[1] "4e-04 BMI"
```

Es decir solamente el BMI y el peso del higado no siguen una normal. Pero no hemos considerado los grupos por separado

```
JarqueBeraTest(datos[datos$gender=="F","weight"])
```

```
    Robust Jarque Bera Test

data:  datos[datos$gender == "F", "weight"]
X-squared = 3.4764, df = 2, p-value = 0.1758
```

```
JarqueBeraTest(datos[datos$gender=="M","weight"])
```

```
    Robust Jarque Bera Test

data:  datos[datos$gender == "M", "weight"]
X-squared = 1.4866, df = 2, p-value = 0.4755
```

```
JarqueBeraTest(datos[datos$gender=="M","height"])
```

```
    Robust Jarque Bera Test

data:  datos[datos$gender == "M", "height"]
X-squared = 1.3449, df = 2, p-value = 0.5104
```

```
JarqueBeraTest(datos[datos$gender=="M","height"])
```

```
    Robust Jarque Bera Test

data:  datos[datos$gender == "M", "height"]
X-squared = 1.3449, df = 2, p-value = 0.5104
```

Los cuatro test resultan no significativos y podemos considerar normales estas dos variables en ambas poblaciones.

Regresión Lineal Simple (OLS)

Procedemos a calcular la regresión del peso del hígado en función del BSA

```r
reg <- lm(liver_weight ~weight,data=datos)
summary(reg)
```

```
Call:
lm(formula = liver_weight ~ weight, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-254.39  -86.29   -9.92   66.44  396.58

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  141.725     71.419   1.984    0.049 *
weight        14.043      1.254  11.195   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 131 on 156 degrees of freedom
Multiple R-squared:  0.4455,    Adjusted R-squared:  0.4419
F-statistic: 125.3 on 1 and 156 DF,  p-value: < 2.2e-16
```

Esto nos quiere decir, que por cada kilogramo de peso en las personas, hay un incremento de 14 gramos en el peso del hígado.

REGRESION MÚLTIPLE

Ahora trabajaremos con el dataset `prostate` el cual se encuentra en el paquete `faraway`. Este dataset donsiste en 97 filas y 9 columnas,los cuales se les realizo una prostatectomía

```r
summary(prostate)
```

```
     lcavol           lweight           age             lbph
 Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863
 1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863
 Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001
 Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004
 3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581
```

```
 Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263
      svi              lcp              gleason          pgg45
 Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00
 1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00
 Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00
 Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38
 3rd Qu.:0.0000   3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00
 Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00
      lpsa
 Min.   :-0.4308
 1st Qu.: 1.7317
 Median : 2.5915
 Mean   : 2.4784
 3rd Qu.: 3.0564
 Max.   : 5.5829
```

```
regr.pros <- lm(lpsa~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45,prostate)
summary(regr.pros)
```

```
Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
    gleason + pgg45, data = prostate)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7331 -0.3713 -0.0170  0.4141  1.6381

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669337   1.296387   0.516  0.60693
lcavol       0.587022   0.087920   6.677 2.11e-09 ***
lweight      0.454467   0.170012   2.673  0.00896 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
svi          0.766157   0.244309   3.136  0.00233 **
lcp         -0.105474   0.091013  -1.159  0.24964
gleason      0.045142   0.157465   0.287  0.77503
pgg45        0.004525   0.004421   1.024  0.30886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.7084 on 88 degrees of freedom
Multiple R-squared:  0.6548,    Adjusted R-squared:  0.6234
F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

```r
head(x0pros <- data.frame(lcavol=1.44692,
                          lweight=3.62301,
                          age=65,
                          lbph=0.30010,
                          svi=0,
                          lcp=-0.79851,
                          gleason=7,
                          pgg45=15))
```

```
   lcavol lweight age   lbph svi      lcp gleason pgg45
1 1.44692 3.62301  65 0.3001   0 -0.79851       7    15
```

```r
predict(regr.pros, x0pros, interval="prediction", level=0.95)
```

```
       fit       lwr      upr
1 2.389053 0.9646584 3.813447
```

El intervalo con el valor de 20 en `age` es más amplio que cuando es 65 debido a que ese valor está fuera del rango de valores para esa variables, y el modelo está extrapolando sobre valores que quedan fuera de aquellos sobre los que se ha contruido el modelo de ajuste. Cuanto más alejados sean los valores predictores de ese rango de valores originales, más amplio será el intervalo, mayor el error y menos ajustada la predicción.

```r
summary(regr.pros)$coef[,4]<0.05
```

```
(Intercept)      lcavol     lweight         age        lbph         svi
      FALSE        TRUE        TRUE       FALSE       FALSE        TRUE
        lcp     gleason       pgg45
      FALSE       FALSE       FALSE
```

```r
confint(regr.pros)
```

```
                  2.5 %      97.5 %
(Intercept) -1.906960983 3.245634379
lcavol       0.412298699 0.761744954
lweight      0.116603435 0.792331414
age         -0.041840618 0.002566267
lbph        -0.009101499 0.223209561
svi          0.280644232 1.251670420
lcp         -0.286344443 0.075395916
gleason     -0.267786053 0.358069248
pgg45       -0.004260932 0.013311395
```

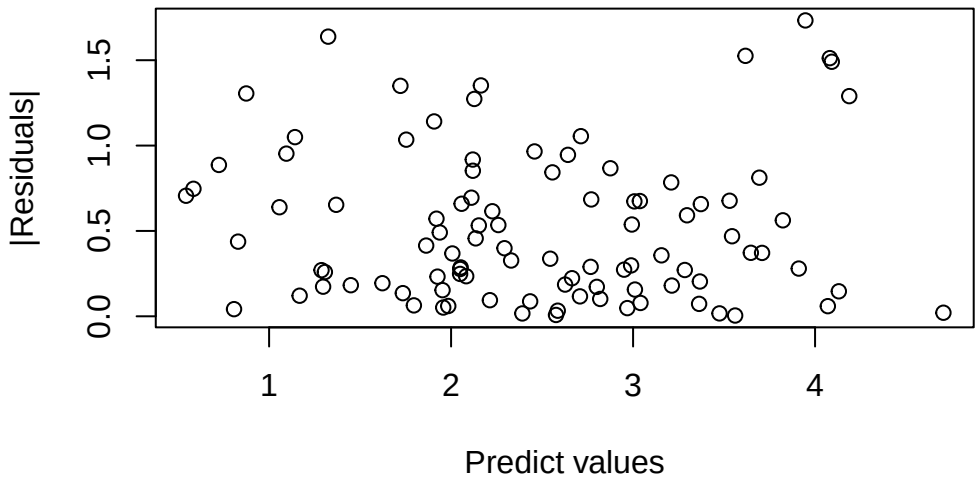
```r
regr.pros2 <- lm(lpsa~lcavol+lweight+svi,prostate)
summary(regr.pros2)
```

```
Call:
lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.72964 -0.45764  0.02812  0.46403  1.57013

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.26809    0.54350  -0.493  0.62298
lcavol       0.55164    0.07467   7.388  6.3e-11 ***
lweight      0.50854    0.15017   3.386  0.00104 **
svi          0.66616    0.20978   3.176  0.00203 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7168 on 93 degrees of freedom
Multiple R-squared:  0.6264,    Adjusted R-squared:  0.6144
F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

Suposiciones

Varianza constate

```
model <- lm(lpsa~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45,prostate)
plot(fitted(model),abs(residuals(model)),xlab="Predict values",ylab="|Residuals|")
```



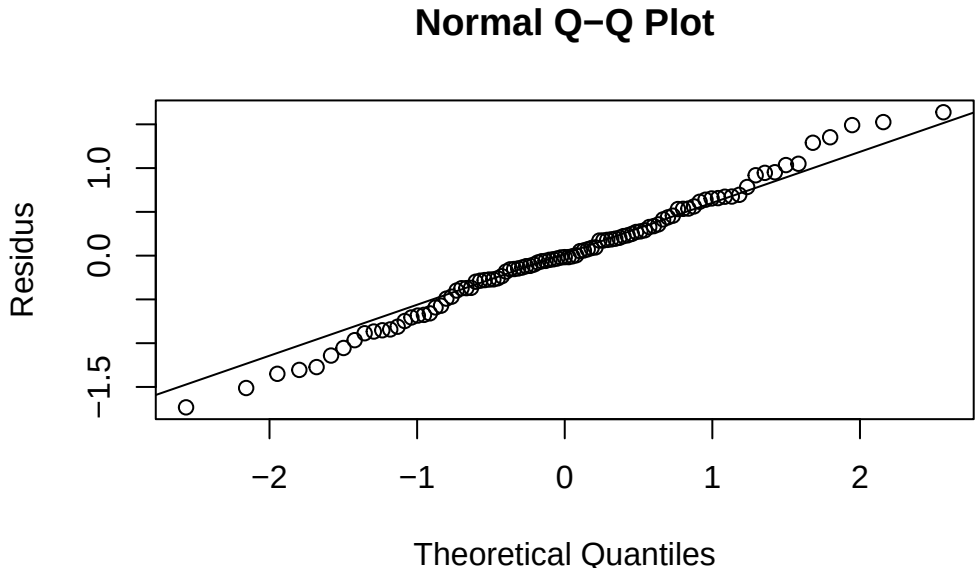

```
sumary(lm(sqrt(abs(residuals(model)))~fitted(model)))
```

```
               Estimate Std. Error t value  Pr(>|t|)
(Intercept)    0.703381   0.090607  7.7630 9.475e-12
fitted(model) -0.021990   0.034232 -0.6424    0.5222

n = 97, p = 2, Residual SE = 0.31328, R-Squared = 0
```

No existen indicios de heterocedasticidad

Normalidad

```
qqnorm(residuals(model),ylab="Residus")
qqline(residuals(model))
```




```
shapiro.test(residuals(model))
```

```
    Shapiro-Wilk normality test

data:  residuals(model)
W = 0.99113, p-value = 0.7721
```

Hay una cierta desviación respecto a la normal, con las colas algo alargadas.

Leverage

O influencia de los puntos

```
hatv <- hatvalues(model)
head(sort(hatv,decreasing=T))
```

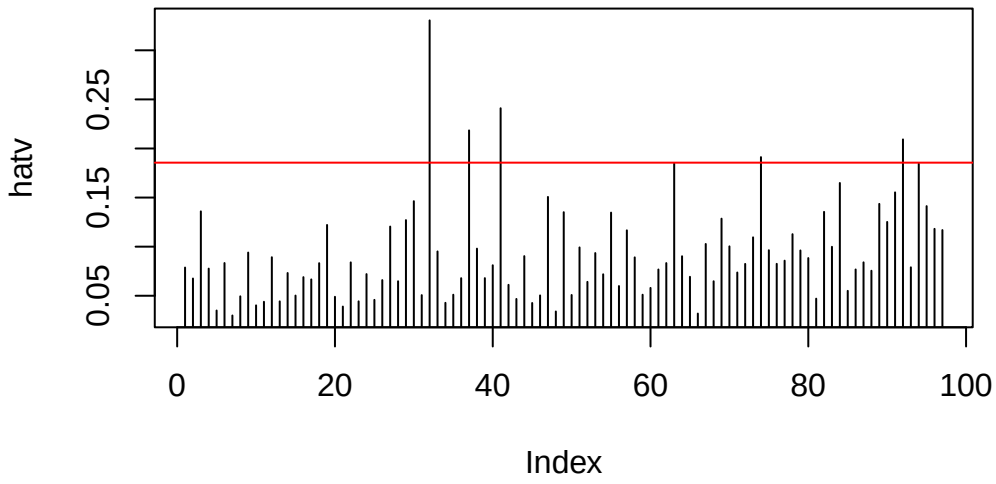
```
       32        41        37        92        74        63
0.3304757 0.2410079 0.2184392 0.2092421 0.1912109 0.1846807
```

```
p <- length(model$coefficients) # k+1
n <- length(model$fitted.values)
```

```r
which(hatv > 2*p/n)
```

```
32 37 41 74 92
32 37 41 74 92
```

```r
plot(hatv, type="h")
 abline(h=2*p/n, col="red")
```



Valores atípicos u outliers

```r
stud <- rstudent(model)
which(abs(stud) > abs(qt(0.05/(2*n),df=n-p-1)))
```

```
named integer(0)
```

TAREA

Leer y realizar resumen de pas pags 225-282, incluido el lab. En un documento de Quarto.