Martinez_Sequera_Amelia_RMM_PEC1

Amelia Martínez Sequera

Abril 2020

Problema 1

```
Male<-c(rep(0,17), rep(1,14))
Female<- c(rep(1,17), rep(0,14))
Alcoholic <- c(1,1,rep(0,15),rep(1,5),rep(0,9))
alcohol<- cbind(alcohol0, Male, Female, Alcoholic)
head(alcohol)
```

```
##   Metabol Gastric      Sex         Alcohol Male Female Alcoholic
## 1     0.6     1.6 "Female"     "Alcoholic"    0      1         1
## 2     1.5     1.5 "Female"     "Alcoholic"    0      1         1
## 3     0.4     2.2 "Female" "Non-alcoholic"    0      1         0
## 4     0.1     1.1 "Female" "Non-alcoholic"    0      1         0
## 5     0.2     1.2 "Female" "Non-alcoholic"    0      1         0
## 6     0.3     0.9 "Female" "Non-alcoholic"    0      1         0
```

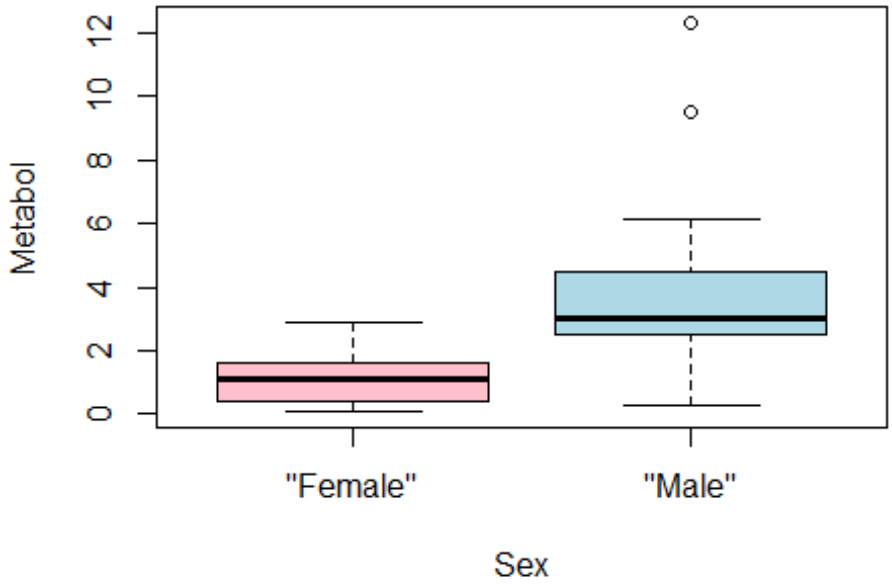
Relación beta0(2)= Metabol - beta1(1)Gastric + beta2(1)Female - error

Diferencias de Metabol entre hombres y mujeres

```
M1<- lm(Metabol~Gastric+Female, data=alcohol )
M2<- lm(Metabol~Gastric+Male, data=alcohol)
M3<- lm(Metabol~Gastric+Male+Female, data=alcohol)
M4<- lm(Metabol~0+Gastric+Male+Female, data=alcohol)
```

Vemos gráficamente con un diagrama de cajas que existen diferencias. Si construimos un modelo con la variable Sex, es el coeficiente de ésta la que nos muestra la diferencia entre ambos sexos:

```
M<- lm(Metabol~Gastric+Sex, data=alcohol)
boxplot(Metabol~Sex, alcohol, col=c("pink", "lightblue"))
```




```
M$coefficients

## (Intercept)     Gastric   Sex"Male"
##   -2.002702    1.979403    1.642197
```

#Mejor modelo según coeficiente de determinación: M4, R-Squared=0.87, explica mayor porcentaje de varianza.

```
sumary(M1);sumary(M2); sumary(M3); sumary(M4)

##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -0.36050    0.71554 -0.5038  0.618330
## Gastric      1.97940    0.27289  7.2535 6.765e-08
## Female      -1.64220    0.52159 -3.1485  0.003878
##
## n = 31, p = 3, Residual SE = 1.35010, R-Squared = 0.76

##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -2.00270    0.54192 -3.6955 0.0009446
## Gastric      1.97940    0.27289  7.2535 6.765e-08
## Male         1.64220    0.52159  3.1485 0.0038780
##
## n = 31, p = 3, Residual SE = 1.35010, R-Squared = 0.76

##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -2.00270    0.54192 -3.6955 0.0009446
```

```
## Gastric      1.97940    0.27289  7.2535 6.765e-08
## Male         1.64220    0.52159  3.1485 0.0038780
##
## n = 31, p = 3, Residual SE = 1.35010, R-Squared = 0.76

##         Estimate Std. Error t value  Pr(>|t|)
## Gastric  1.97940    0.27289  7.2535 6.765e-08
## Male    -0.36050    0.71554 -0.5038 0.6183296
## Female  -2.00270    0.54192 -3.6955 0.0009446
##
## n = 31, p = 3, Residual SE = 1.35010, R-Squared = 0.87
```

Mejor modelo según RMSE: Todos tienen el mismo valor.

Rango M3

```
X<- model.matrix(~Gastric+Male+Female, data = alcohol)
Y<- alcohol$Metabol
qr(X)$rank

## [1] 3
```

No podemos calcular la inversa de X´X porque es una matriz singular. El rango de x (matriz del modelo M3) es 3, ya que las variables Male y Female son linealmente dependientes.

Utilizamos la inversa de Moore-Penrose:

```
library(matlib)
Xtxi<- Ginv (t(X)%*% X)
coef<- Xtxi%*%t(X)%*%Y
coef; M3$coefficients

##           [,1]
## [1,] -2.002703
## [2,]  1.979402
## [3,]  1.642197
## [4,]  0.000000

## (Intercept)     Gastric        Male      Female
##   -2.002702    1.979403    1.642197          NA
```

Los coeficientes son los mismos. Se desestima la variable Female porque es una combinación lineal de Male.

Residuos:

También coinciden.Los restamos para ver que la diferencia es 0

```
M3sum<- summary(M3)
```

```r
residuals<- Y-X %*% coef
head(round(residuals-M3sum$residuals,4))
```

```
##   [,1]
## 1    0
## 2    0
## 3    0
## 4    0
## 5    0
## 6    0
```

Intervalos de confianza para beta0+beta1 en M2 al 95%

Hipótesis nula: 2alpha-beta0-beta1=0

```r
M2sum<- summary(M2)
X2<- model.matrix(~Gastric+Male, alcohol)
betas<- Ginv(t(X2) %*% X2) %*% t(X2) %*% Y
```

Residuos, sigma:

```r
res<- Y-X2 %*% betas
n<- length(Y)
r<- qr(X2)$rank
sigma2<- sum(res^2)/(n-r)
```

t Student:

```r
a <- c(2,-1,-1)
numerador <- t(a) %*% betas
denominador <- sqrt(sigma2 * t(a) %*% Ginv(t(X2) %*% X2) %*% a)
t.est <- numerador/denominador
p.value <- pt(abs(t.est), df = n-r, lower.tail = F) * 2
c(t.est,p.value)
```

```
## [1] -5.395144e+00  9.425730e-06
```

```r
qt(0.975, n-r)
```

```
## [1] 2.048407
```

```r
betas;M2sum$coefficients
```

```
##           [,1]
## [1,] -2.002703
## [2,]  1.979401
## [3,]  1.642197
```

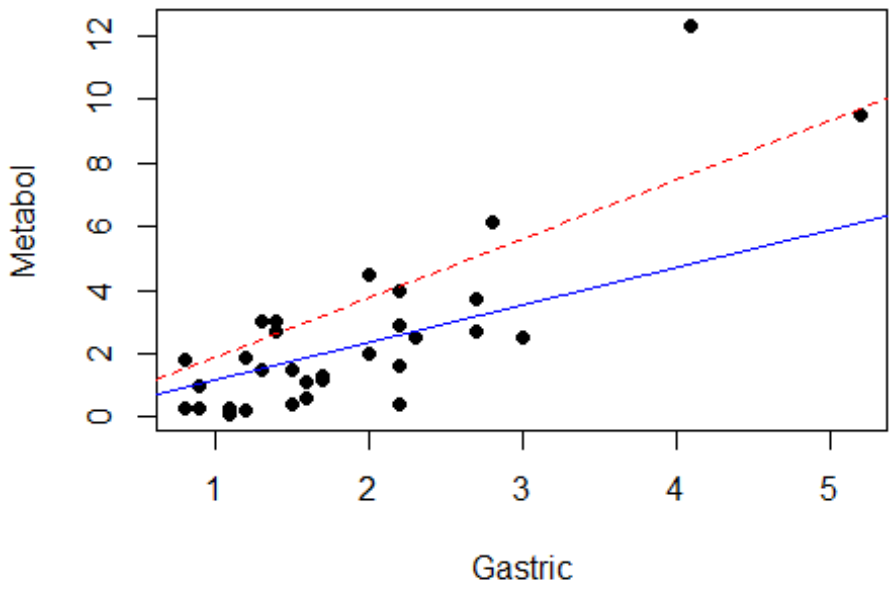
```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -2.002702  0.5419240 -3.695540 9.445837e-04
## Gastric      1.979403  0.2728908  7.253460 6.765435e-08
## Male         1.642197  0.5215852  3.148474 3.877991e-03
```

```
(1.979401-2.002703)+c(-1,1)*2.048407*(0.541924+0.2728908)

## [1] -1.692374  1.645770
```

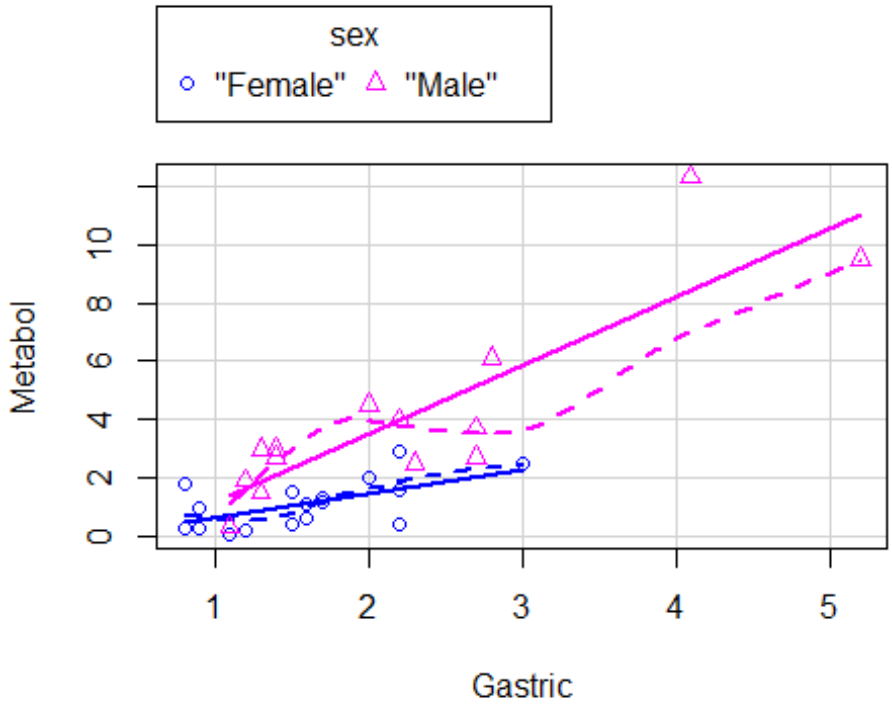
Como elintervalo contiene el 0, la hipótesis nula no sería rechazada para una significación del 5%.

Rectas de regresión:

```
sex<- as.factor(alcohol$Sex)
levels(sex)

## [1] "\"Female\"" "\"Male\""

plot(Metabol~0+Gastric, pch=ifelse(sex=="Male",1,16),data=alcohol)
M12<-lm(Metabol~0+Gastric+Female, data=alcohol)
M22<- lm(Metabol~0+Gastric+Male, data=alcohol)
abline(M22, lty=1, col="blue")
abline(M12, lty=2, col="red")
```




```
scatterplot(Metabol~Gastric|sex, data=alcohol)
```



#Contraste de coincidencia:

```
r1<- lm(Metabol~Gastric*Sex, data = alcohol)
r1

##
## Call:
## lm(formula = Metabol ~ Gastric * Sex, data = alcohol)
##
## Coefficients:
##       (Intercept)            Gastric          Sex"Male"
Gastric:Sex"Male"
##           -0.1887             0.8330            -0.9970
1.5108
```

#Contraste paralelismo:

Suponiendo que las rectas estan a sociadas a un modelo lineal normal, planteamos la hipótesis nula de que los coeficientes de la variable Gastric (pendiente) son iguales en los 2 casos(hombre/mujer):

```
r2<- lm(Metabol~Gastric+Sex, data=alcohol)
anova(r2,r1)

## Analysis of Variance Table
##
## Model 1: Metabol ~ Gastric + Sex
## Model 2: Metabol ~ Gastric * Sex
```

```
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     28 51.038
## 2     27 40.811  1    10.227 6.7659 0.01489 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se rechaza la hipótesis. No son paralelas

#Significación de la variable concomitante(Metabol):

```
anova(r2)
```

```
## Analysis of Variance Table
##
## Response: Metabol
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Gastric    1 146.562 146.562 80.4058 1.009e-09 ***
## Sex        1  18.069  18.069  9.9129  0.003878 **
## Residuals 28  51.038   1.823
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modelo completo

```
MC<-
lm(Metabol~Gastric+Male+Alcoholic+Gastric*Male+Gastric*Alcoholic+Male*Alc
oholic+Gastric*Male*Alcoholic, data=alcohol)
MCsum<- summary(MC)
M2sum<- summary(M2)
MCsum$r.squared; M2sum$r.squared;MCsum$sigma; M2sum$sigma
```

```
## [1] 0.8271248
```

```
## [1] 0.7633509
```

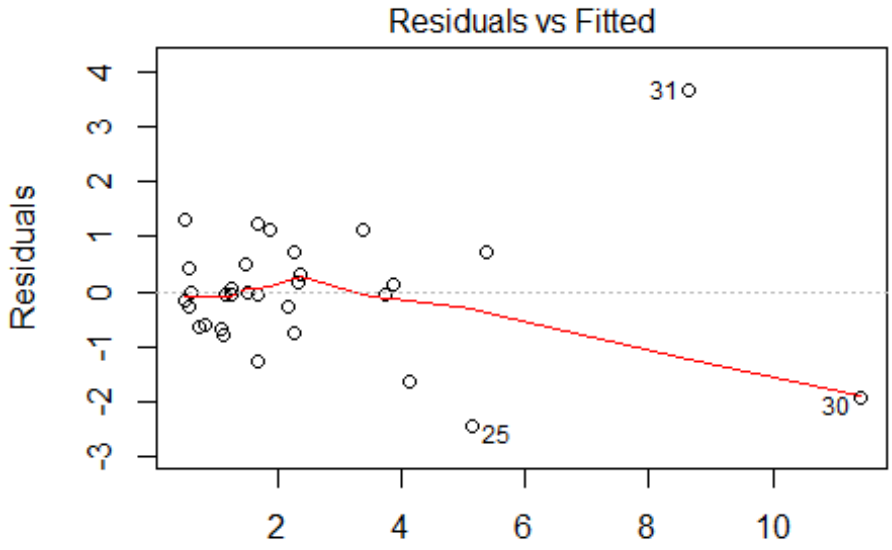
```
## [1] 1.273197
```

```
## [1] 1.350102
```

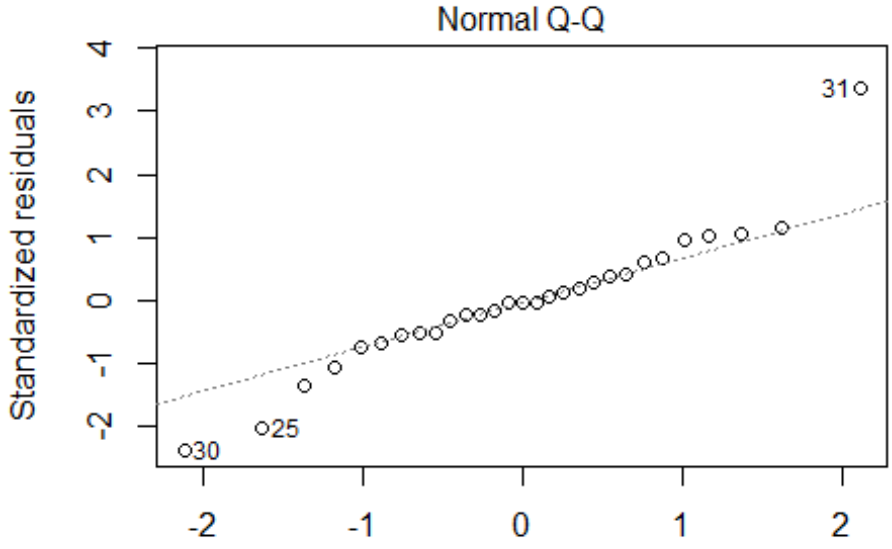
El modelo completo mejora los resultados de R-Squared, explica mayor porcentaje de la varianza.

```
plot(MC)
```

```
## Warning: not plotting observations with leverage one:
##   1, 2
```

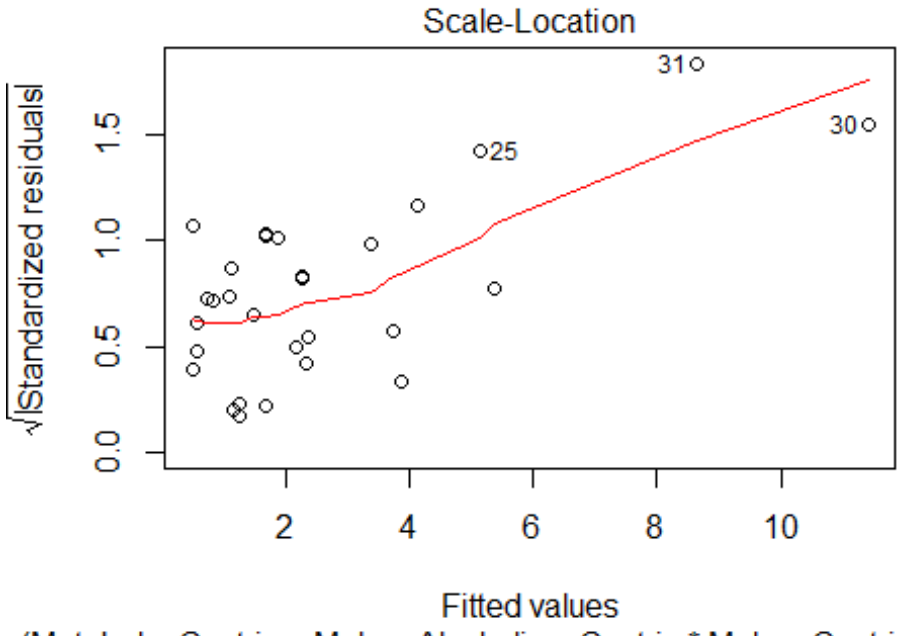


n(Metabol ~ Gastric + Male + Alcoholic + Gastric * Male + Gastric * Alc



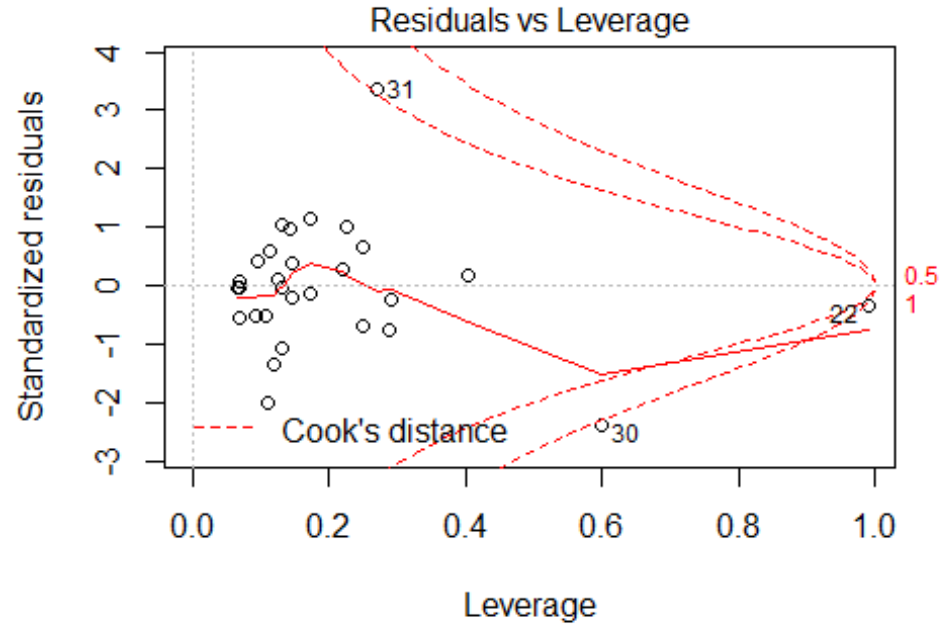
n(Metabol ~ Gastric + Male + Alcoholic + Gastric * Male + Gastric * Alc

```
## Warning: not plotting observations with leverage one:
##   1, 2
```



n(Metabol ~ Gastric + Male + Alcoholic + Gastric * Male + Gastric * Alc

```
## Warning in sqrt(crit * p * (1 - hh)/hh): Se han producido NaNs
## Warning in sqrt(crit * p * (1 - hh)/hh): Se han producido NaNs
```



n(Metabol ~ Gastric + Male + Alcoholic + Gastric * Male + Gastric * Alc

```
anova(M2,MC)
```

```
## Analysis of Variance Table
##
## Model 1: Metabol ~ Gastric + Male
## Model 2: Metabol ~ Gastric + Male + Alcoholic + Gastric * Male +
Gastric *
##     Alcoholic + Male * Alcoholic + Gastric * Male * Alcoholic
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 51.038
## 2     23 37.284  5    13.754 1.6969 0.1754
```

De todas maneras, si comparamos ambos modelos vemos que por el momento no podemor rechazar la hipótesis nula. M2 sigue siendo aceptable.

Problema 2

```
senic <-
read_table2("C:\\Users\\Meli\\Documents\\UOC\\Regres.mod.met\\PEC1_RMM\\s
enic.txt")
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   stay = col_double(),
##   age = col_double(),
##   infrisk = col_double(),
##   culratio = col_double(),
##   xratio = col_double(),
##   nbeds = col_double(),
##   medschl = col_double(),
##   region = col_double(),
##   census = col_double(),
##   nurses = col_double(),
##   service = col_double()
## )
```

Matriz de correlaciones

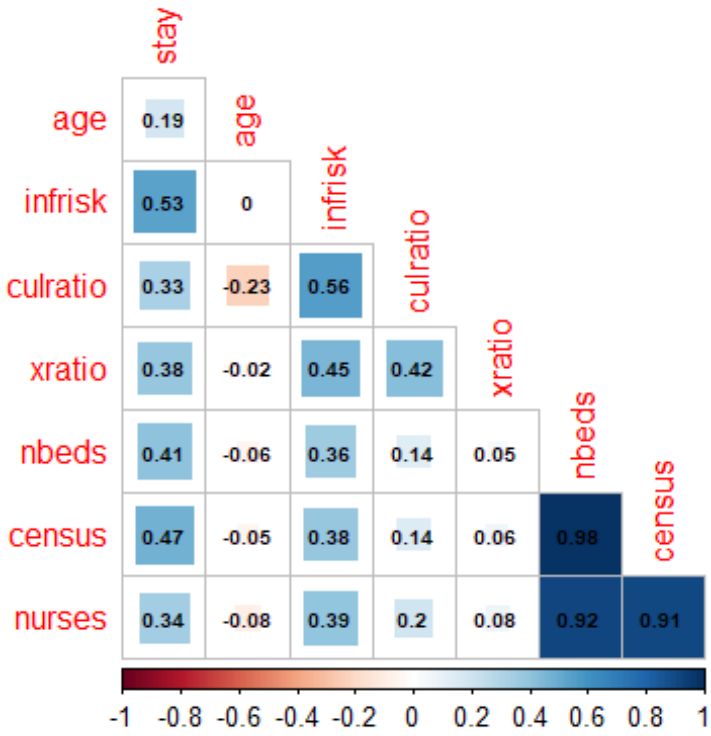
```
head(senic[,-c(1,8,9,12)])
```

```
## # A tibble: 6 x 8
##    stay   age infrisk culratio xratio nbeds census nurses
##   <dbl> <dbl>   <dbl>    <dbl>  <dbl> <dbl>  <dbl>  <dbl>
## 1  7.13  55.7     4.1      9     39.6   279    207    241
## 2  8.82  58.2     1.6      3.8   51.7    80     51     52
## 3  8.34  56.9     2.7      8.1   74     107     82     54
## 4  8.95  53.7     5.6     18.9  123.    147     53    148
## 5 11.2   56.5     5.7     34.5   88.9   180    134    151
## 6  9.76  50.9     5.1     21.9   97     150    147    106
```

```r
senic_cor<- cor(senic[,-c(1,8,9,12)])
senic_cor
```

```
##               stay          age     infrisk   culratio      xratio
nbeds
## stay     1.0000000  0.188913972 0.533443831  0.3266838  0.38248193
0.40926525
## age      0.1889140  1.000000000 0.001093166 -0.2258468 -0.01885490 -
0.05882316
## infrisk  0.5334438  0.001093166 1.000000000  0.5591589  0.45339156
0.35977000
## culratio 0.3266838 -0.225846789 0.559158869  1.0000000  0.42496204
0.13972495
## xratio   0.3824819 -0.018854897 0.453391557  0.4249620  1.00000000
0.04581997
## nbeds    0.4092652 -0.058823160 0.359770000  0.1397249  0.04581997
1.00000000
## census   0.4738855 -0.054774667 0.381411081  0.1429482  0.06291352
0.98099774
## nurses   0.3403671 -0.082944616 0.393981340  0.1988998  0.07738133
0.91550415
##               census      nurses
## stay      0.47388550  0.34036706
## age      -0.05477467 -0.08294462
## infrisk   0.38141108  0.39398134
## culratio  0.14294821  0.19889983
## xratio    0.06291352  0.07738133
## nbeds     0.98099774  0.91550415
## census    1.00000000  0.90789698
## nurses    0.90789698  1.00000000
```

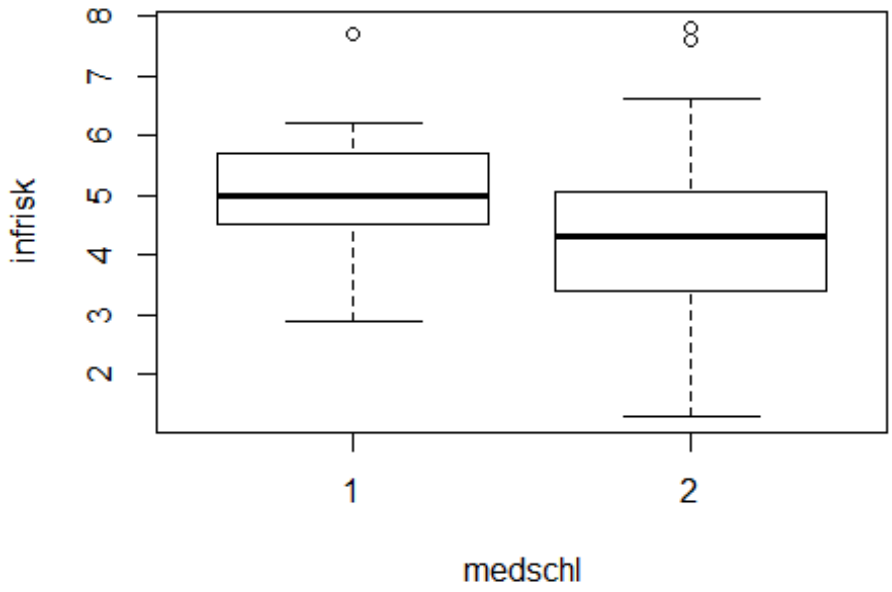
```r
corrplot(senic_cor, method= "square",type="lower", diag=F, addCoef.col =
"black",number.cex = 0.6)
```



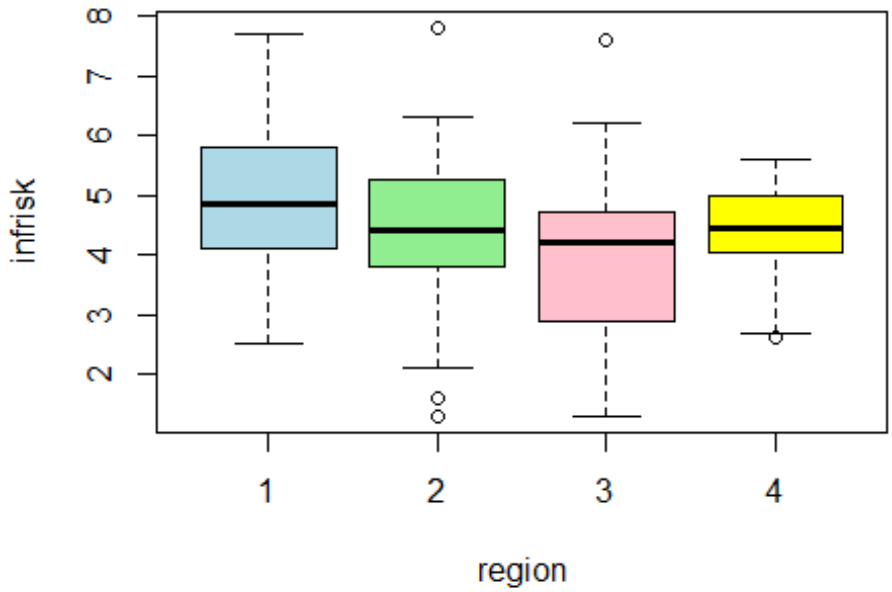
Las variables census, nurses y nbeds están muy correlacionadas entre sí. No se han incluido las variables categóricas: medschl y region. Son variables categóricas pero codificadas numéricamente, tanto medschl(1=si, 2=no), como region(1=NE, 2=NC, 3=S, 4=W). La variable medschl explica, por ejemplo, cuánto valdrá más la variable explicada para los no afiliados(el doble).

#Boxplot

```
boxplot(infrisk~medschl, senic)
```




```
boxplot(infrisk~region, senic,
col=c("lightblue","lightgreen","pink","yellow"))
```



1=NE, 2=NC, 3=S, 4=W

ANOVA

consideramos como hipótesis nula que el coeficiente de la variable medschl es igual a 0 (no influye en el riesgo de infección)

```
lmod<-
lm(infrisk~stay+age+culratio+xratio+nbeds+medschl+region+census+nurses+se
rvice, data=senic)
lmod2<-
lm(infrisk~stay+age+culratio+xratio+nbeds+region+census+nurses+service,
data = senic)
anova(lmod2,lmod)
```

```
## Analysis of Variance Table
##
## Model 1: infrisk ~ stay + age + culratio + xratio + nbeds + region +
census +
##     nurses + service
## Model 2: infrisk ~ stay + age + culratio + xratio + nbeds + medschl +
##     region + census + nurses + service
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1    103 89.273
## 2    102 86.647  1    2.6257 3.091 0.08173 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No podemos rechazar la hipótesis nula (p-value= 0.08) a un 0.05 de significación, pero sí a un 0.01.

Ahora, consideramos como hipótesis nula que el coeficiente de region es 0 (la región no influye en el riesgo de infección)

```
lmod3<-
lm(infrisk~stay+age+culratio+xratio+nbeds+medschl+census+nurses+service,
data = senic)
anova(lmod3, lmod)
```

```
## Analysis of Variance Table
##
## Model 1: infrisk ~ stay + age + culratio + xratio + nbeds + medschl +
##     census + nurses + service
## Model 2: infrisk ~ stay + age + culratio + xratio + nbeds + medschl +
##     region + census + nurses + service
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    103 93.536
## 2    102 86.647  1    6.8893 8.1101 0.005324 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este caso si que rechazamos la hipótesis nula (p= 0.005324). La región influye en el valor de infrisk.

Modelo

```
lmod<-
lm(infrisk~stay+age+culratio+xratio+nbeds+medschl+region+census+nurses+se
rvice, data=senic)
summary(lmod)
```

```
##
## Call:
## lm(formula = infrisk ~ stay + age + culratio + xratio + nbeds +
##     medschl + region + census + nurses + service, data = senic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86974 -0.56269 -0.02893  0.51925  2.32390
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.239992   1.413934  -2.291 0.023992 *
## stay         0.242240   0.070668   3.428 0.000879 ***
## age          0.010077   0.021500   0.469 0.640280
## culratio     0.053444   0.010545   5.068  1.8e-06 ***
## xratio       0.012632   0.005280   2.392 0.018568 *
## nbeds       -0.003173   0.002660  -1.193 0.235605
## medschl      0.562186   0.319765   1.758 0.081727 .
## region       0.297558   0.104486   2.848 0.005324 **
## census       0.002829   0.003430   0.825 0.411415
## nurses       0.002064   0.001688   1.223 0.224125
## service      0.023272   0.010028   2.321 0.022291 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9217 on 102 degrees of freedom
## Multiple R-squared:  0.5697, Adjusted R-squared:  0.5276
## F-statistic: 13.51 on 10 and 102 DF,  p-value: 7.974e-15
```

Stay y culratio son la variables predictoras más significativas. La variable region también tiene un nivel de significación alto pero, como se ha mencionado anteriormente, es una variable categórica, igual que medschl, que toman valores numéricos (1,2,3,4).

El modelo también nos da el valor del F-test, y un valor de p-value que es significativo. La hipótesis nula es que la media es la misma para los diferentes grupos de análisis y la hipótesis alternativa es que, al menos, dos medias de los diferentes grupos, difieren.

Predictoras significativas al 5%:

```
summary(lmod)$coef[,4]<0.05
```

```
## (Intercept)        stay         age    culratio      xratio
nbeds
```

```
##        TRUE        TRUE       FALSE        TRUE        TRUE
FALSE
##     medschl      region      census      nurses     service
##       FALSE        TRUE       FALSE       FALSE        TRUE
```

Modelo - variables con significación <5% (lmodb)

```
summary(lmod)$coef[,4]<0.05
```

```
## (Intercept)        stay         age    culratio      xratio
nbeds
##        TRUE        TRUE       FALSE        TRUE        TRUE
FALSE
##     medschl      region      census      nurses     service
##       FALSE        TRUE       FALSE       FALSE        TRUE
```

Contraste frente al completo: hipótesis nula: los coeficientes de las variables con significación <5% son igual a 0.

```
lmodb<- lm(infrisk~stay+culratio+xratio+region+service, data=senic)
anova (lmodb, lmod)
```

```
## Analysis of Variance Table
##
## Model 1: infrisk ~ stay + culratio + xratio + region + service
## Model 2: infrisk ~ stay + age + culratio + xratio + nbeds + medschl +
##     region + census + nurses + service
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    107 91.231
## 2    102 86.647  5    4.5837 1.0792 0.3764
```

No rechazamos de momento la hipótesis nula. El modelo reducido es una buena opción.

También podemos calcularlos intervalos de confianza al 95% de los parámetros de cada variable, y descartar aquellas en que su intervalo contiene el 0.

```
confint(lmod)
```

```
##                    2.5 %       97.5 %
## (Intercept) -6.044524003 -0.435459349
## stay         0.102070855  0.382408890
## age         -0.032567512  0.052721582
## culratio     0.032528828  0.074359038
## xratio       0.002159184  0.023104975
## nbeds       -0.008448624  0.002102233
## medschl     -0.072066895  1.196439263
## region       0.090309979  0.504806933
## census      -0.003974149  0.009631925
## nurses      -0.001283282  0.005411258
## service      0.003381908  0.043162558
```

Obtendríamos el mismo modelo lmodb.

Modelo con stay, culratio y region (lmodc). Normalidad y heterocedasticidad. Gráficos.

```
lmodc<- lm(infrisk~stay+culratio+region, data=senic)
lmodcsum<-summary(lmodc)
mean(lmodcsum$residuals);lmodcsum$sigma;cor(fitted(lmodc),lmodcsum$residu
als)
```

```
## [1] -1.101281e-17
```

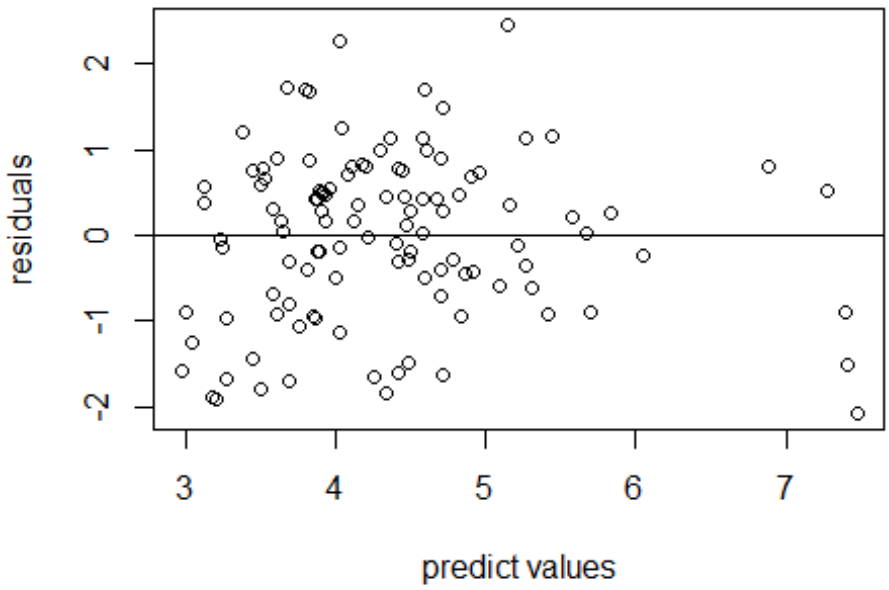
```
## [1] 0.9850368
```

```
## [1] -2.71473e-16
```

Observamos que el valor medio de los residuos es casi cero. Se espera que los residuos sean independientes de la variable explicativa y del modelo ajustado, que la varianza sea constante, y que sigan una distribución normal. De no ser así podríamos suponer que el modelo no está bien ajustado o que faltan predictores.

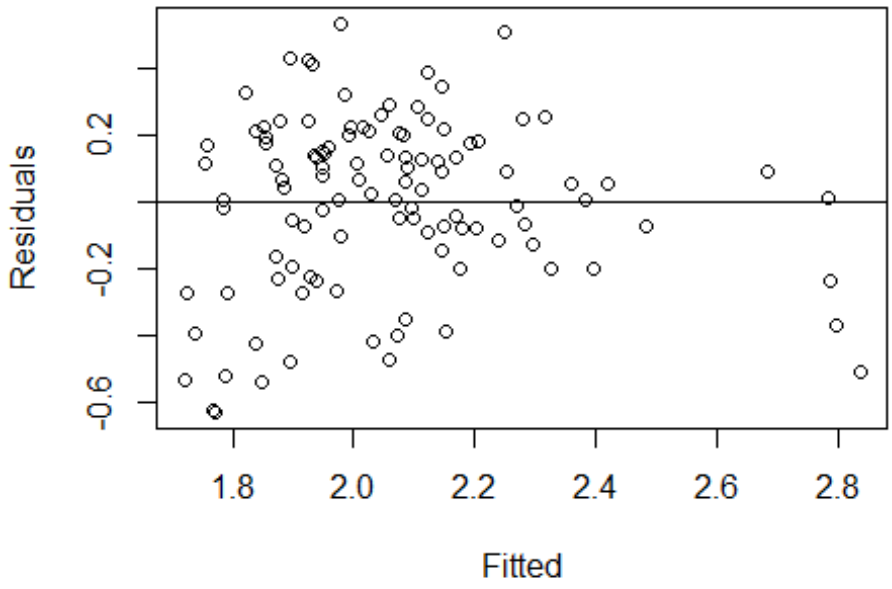
#Homocedasticidad: igualdad en la varianza de los errores.

```
plot(fitted(lmodc),residuals(lmodc), xlab="predict values",
ylab="residuals")
abline(h=0)
```



No se ven muy uniformes. Puede ser útil hacer la transformación de la raíz cuadrada de la respuesta:

```
lmodc2 <- lm(sqrt(infrisk) ~stay+culratio+region, data=senic)
plot(fitted(lmodc2),residuals(lmodc2),xlab="Fitted",ylab="Residuals")
abline(h=0)
```

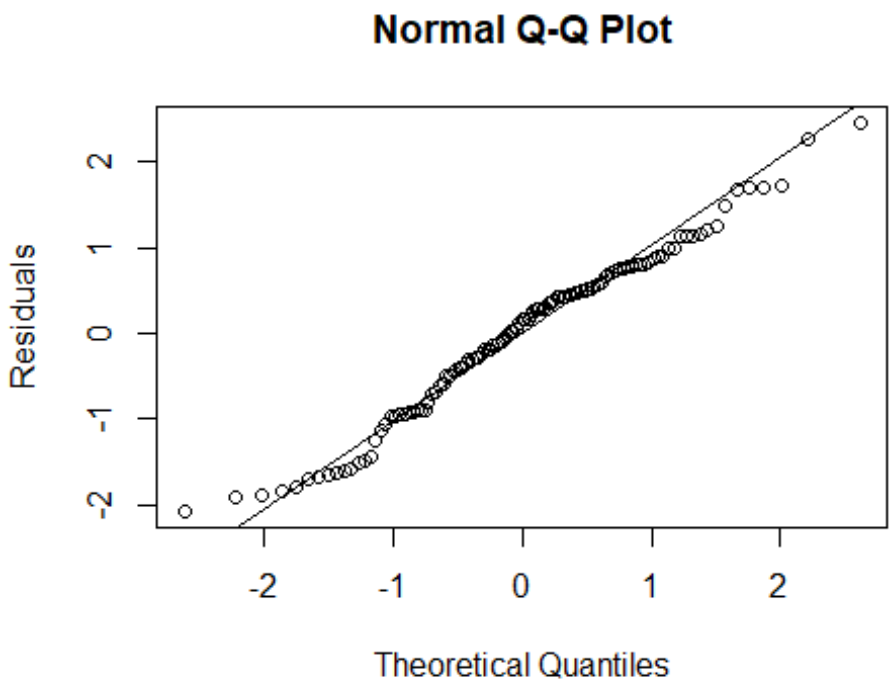



```
library(faraway)

sumary(lm(sqrt(abs(residuals(lmodc)))~fitted(lmodc)))

##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    0.8655047  0.1495400  5.7878 6.711e-08
## fitted(lmodc) -0.0090241  0.0335973 -0.2686    0.7887
##
## n = 113, p = 2, Residual SE = 0.32853, R-Squared = 0
```

El gráfico nos muestra una varianza no-lineal. El valor de p(>0.05) indica que no existe una relación significativa.

#Normalidad:

```
qqnorm(residuals(lmodc),ylab="Residuals")
qqline(residuals(lmodc))
```




```
shapiro.test(residuals(lmodc))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(lmodc)
## W = 0.98082, p-value = 0.1047
```

Ni gráficamente ni por el test de contraste, vemos que la distribución de los residuos se aparte mucho de la normal.La hipótesis del test es que los residuos son normales, y por el valor de p no podemos rechazarla.

#Leverage:

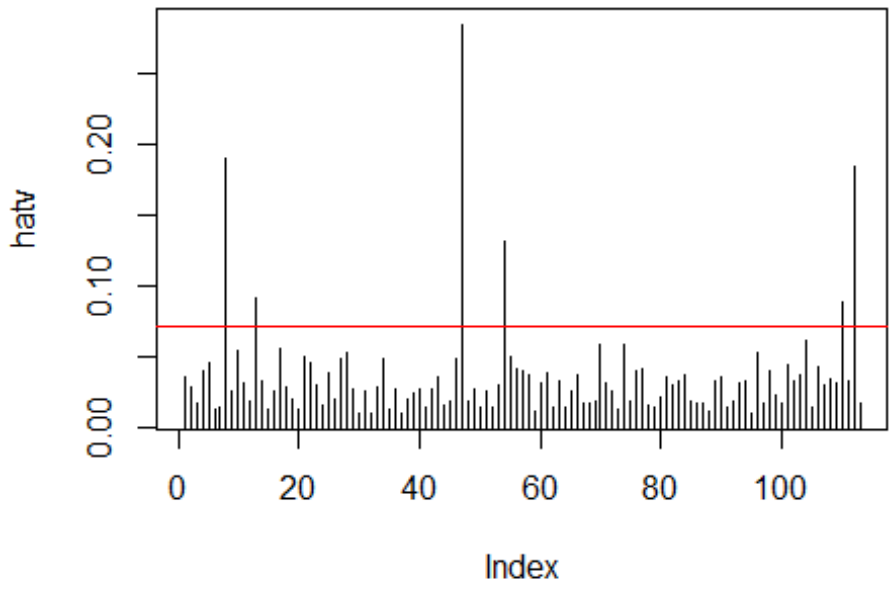
observaciones con el leverage más alto:

```
hatv <- hatvalues(lmodc  )
head(sort(hatv,decreasing=T)); sum(hatv)

##         47          8        112         54         13        110
## 0.28400703 0.18931763 0.18465286 0.13069866 0.09142453 0.08840986

## [1] 4
```

Gráficamente:

```
pc <- length(lmodc$coefficients)
nc <- length(lmodc$fitted.values)
```

```r
leverage.mean <- pc/nc
plot(hatv, type="h")
abline(h=2*leverage.mean, col="red")
```



Comprobamos de ambas maneras que las observaciones con un leverage más alto son la 47, 8 y 112.

#Son outliers? consideramos el valor crítico de la t de Student y la corrección de Bonferroni. Nivel de significación 5%.

```r
stud<- rstandard(lmodc)
grlib <- nc-pc-1
head(sort(abs(stud), decreasing = T)); grlib
```

```
##       53        8       35       93       40       96
## 2.533920 2.339587 2.317419 1.964497 1.930958 1.913295
```

```
## [1] 108
```

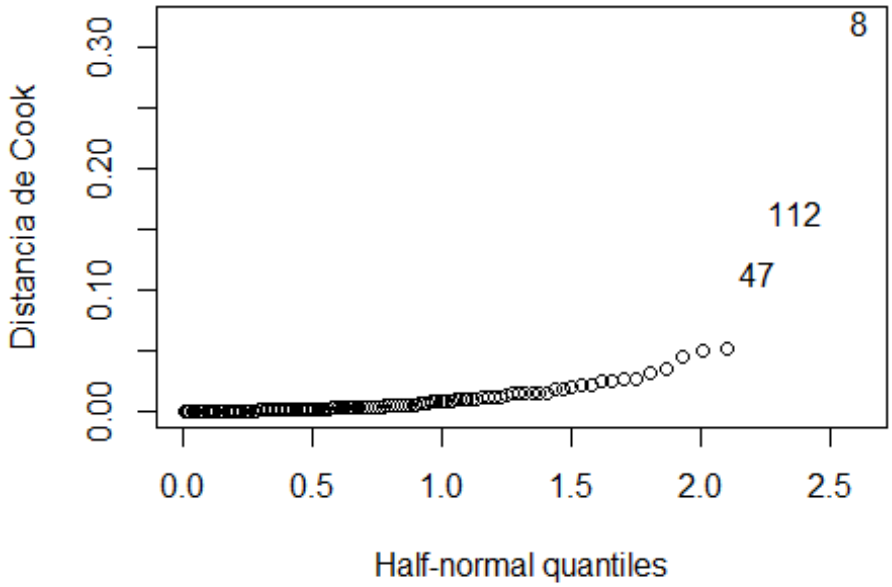
```r
which(abs(stud) > abs(qt(0.05/(2*nc),grlib)))
```

```
## named integer(0)
```

Con este último criterio, todos los residuos quedan por debajo del valor crítico, no hallamos ningún valor atípico. Queda la duda de si puede haber grupos de valores atípicos que no hayamos sabido encontrar.

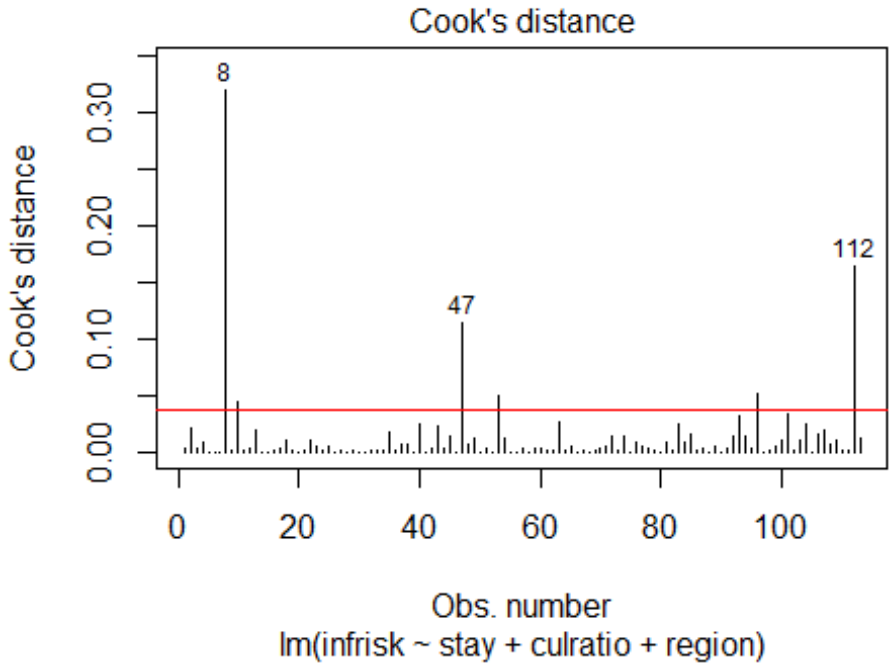
#Observaciones influyentes: Calculamos la distancia de Cook como medida de la influencia de los puntos y la representamos contra los cuartiles de una distribución seminormal.

```
cook <- cooks.distance(lmodc)
halfnorm(cook,nlab=3,ylab="Distancia de Cook")
```



#Criterio de selección:

```
plot(lmodc, which=4)
abline(h=4/((nc-pc-2)), col="red")
```



Vemos que las observaciones 8, 112 y 47 son las que tienen más influencia.

Predicción del riesgo de infección (intervalo 90%) lmodc

stay= 9.6, cultratio= 15.5, region=NE, lmodc

```
class(senic$region)
```

```
## [1] "numeric"
```

```
p<- predict(lmodc, newdata = data.frame(stay= 9.6, culratio= 15.5,
region=1), interval="prediction", level=0.9)
p
```

```
##        fit      lwr      upr
## 1 3.990901 2.331176 5.650626
```

NE está codificado como valor numérico (1).

```
xc<- model.matrix(~stay+culratio+region,data=senic)
yc<- senic$infrisk
dim(xc)
```

```
## [1] 113   4
```

```
head(xc[,2:4])
```

```
##    stay culratio region
## 1  7.13      9.0      4
## 2  8.82      3.8      2
## 3  8.34      8.1      3
## 4  8.95     18.9      4
## 5 11.20     34.5      1
## 6  9.76     21.9      2
```

1=NE, 2=NC, 3=S, 4=W

```
summary(lmodc)
```

```
##
## Call:
## lm(formula = infrisk ~ stay + culratio + region, data = senic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0750 -0.6897  0.1616  0.6959  2.4579
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.370464   0.708804  -0.523   0.6023
## stay         0.331548   0.057229   5.793 6.77e-08 ***
## culratio     0.060398   0.009781   6.175 1.16e-08 ***
## region       0.242329   0.107666   2.251   0.0264 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.985 on 109 degrees of freedom
## Multiple R-squared:  0.4748, Adjusted R-squared:  0.4604
## F-statistic: 32.85 on 3 and 109 DF,  p-value: 3.337e-15
```

Predicción manual stay y culratio (en este orden) 90%

```
0.331548+c(-1,1)*qt(0.9,109)*0.057229
```

```
## [1] 0.2577588 0.4053372
```

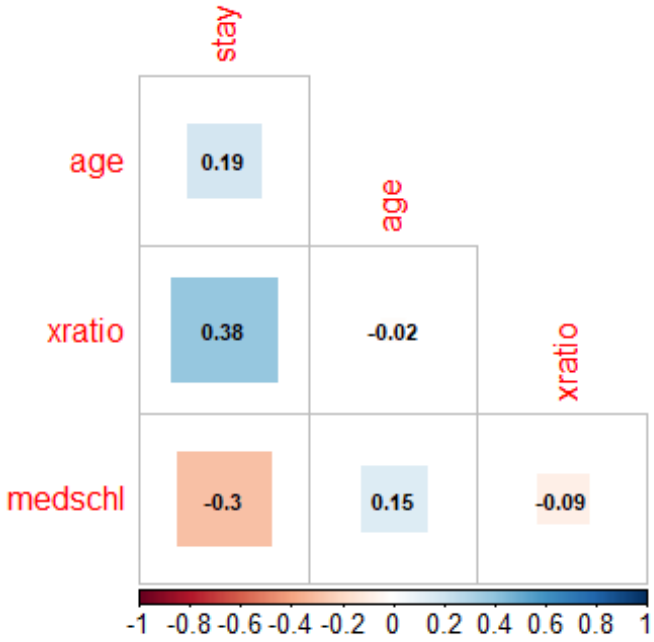
```
0.06+c(-1,1)*qt(0.9,109)*0.009781
```

```
## [1] 0.0473887 0.0726113
```

Modelo stay+age+xratio+medschl (modd), alpha=0.1

```
modd<- lm(infrisk~stay+ age+ xratio+ medschl, data=senic)
summodd<- summary(modd)
corsenicb <- cor(senic[,-c(1,4,5,7,9,10,11,12)])
corsenicb; corrplot(corsenicb, method= "square",type="lower", diag=F,
addCoef.col = "black",number.cex = 0.7)
```

```
##               stay        age      xratio     medschl
## stay     1.0000000  0.1889140  0.38248193 -0.29695100
## age      0.1889140  1.0000000 -0.01885490  0.14512637
## xratio   0.3824819 -0.0188549  1.00000000 -0.08669664
## medschl -0.2969510  0.1451264 -0.08669664  1.00000000
```



No existe una gran correlación con age, y menos aún con xratio.

```
cor.test(senic$medschl, senic$age); cor.test(senic$medschl, senic$xratio)
```

```
##
##  Pearson's product-moment correlation
##
## data:  senic$medschl and senic$age
## t = 1.5454, df = 111, p-value = 0.1251
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04069444  0.32124391
## sample estimates:
##       cor
## 0.1451264

##
##  Pearson's product-moment correlation
##
## data:  senic$medschl and senic$xratio
## t = -0.91686, df = 111, p-value = 0.3612
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.26714797  0.09962878
## sample estimates:
##         cor
## -0.08669664
```

Comparamos este modelo (modd), con el mismo modelo pero: 1)considerando que hay interacción entre medschl y age, y 2) hay interacción entre medschl y xratio.

```
modd1<- lm(infrisk~stay+age+xratio+medschl+medschl:age, data=senic)
anova(modd, modd1)

## Analysis of Variance Table
##
## Model 1: infrisk ~ stay + age + xratio + medschl
## Model 2: infrisk ~ stay + age + xratio + medschl + medschl:age
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    108 127.24
## 2    107 125.50  1    1.7436 1.4866 0.2254

modd2<- lm(infrisk~stay+age+xratio*medschl, data=senic)
anova(modd2,modd)

## Analysis of Variance Table
##
## Model 1: infrisk ~ stay + age + xratio * medschl
## Model 2: infrisk ~ stay + age + xratio + medschl
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    107 125.30
## 2    108 127.24 -1   -1.9442 1.6603 0.2003
```

La hipótesis nula es que los términos de interacción son 0.A un nivel de significación alpha=0.1<p, no rechazamos la hipótesis nula. Se puede prescindir de los términos de interacción.