Normalidad y datos atípicos

J. Ramajo

2020

library(readr)  
library(car)

library(MASS)  
library(effects)

library(tidyverse)

library(RcmdrMisc)

library(sfsmisc)

library(quantreg)

library(lmtest)

library(sandwich)  
library(dynlm)  
library(moments)  
library(tseries)

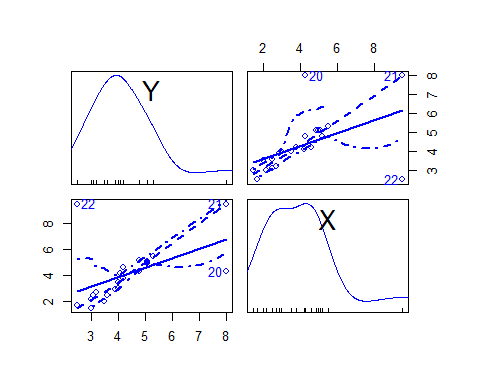
#  
ATIP <- read\_csv("ATIP.csv")

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## X = col\_double(),  
## Y = col\_double()  
## )

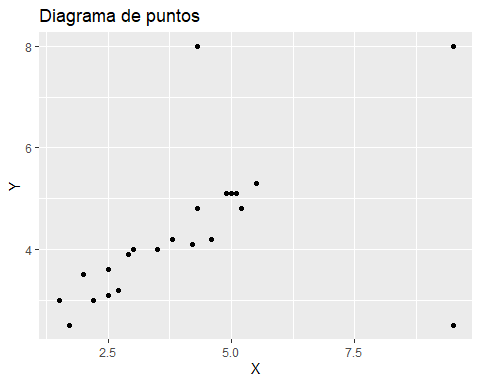
View(ATIP)  
summary(ATIP)

## X Y   
## Min. :1.500 Min. :2.500   
## 1st Qu.:2.550 1st Qu.:3.275   
## Median :4.000 Median :4.050   
## Mean :4.109 Mean :4.318   
## 3rd Qu.:4.975 3rd Qu.:5.025   
## Max. :9.500 Max. :8.000

#  
# Diagrama de puntos  
#  
scatterplotMatrix(~ Y + X, id=list(n=3), smooth=list(span=0.7), data=ATIP)



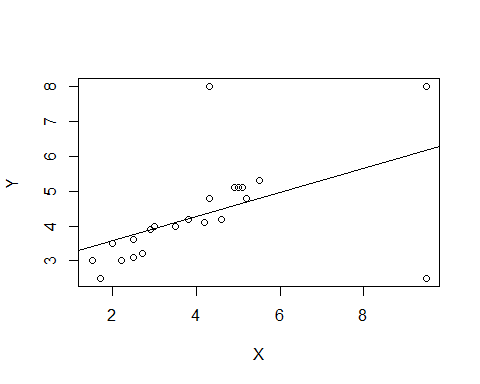
ggplot(ATIP, aes(x=X, y=Y)) + geom\_point() + labs(title="Diagrama de puntos", x="X", y="Y")



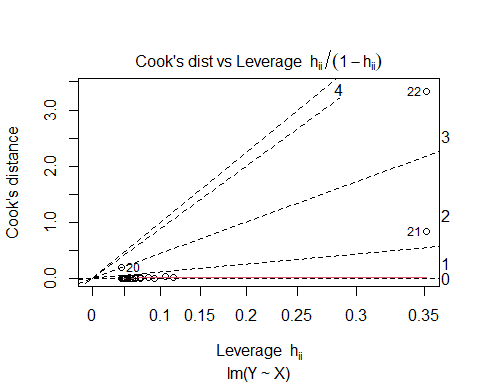
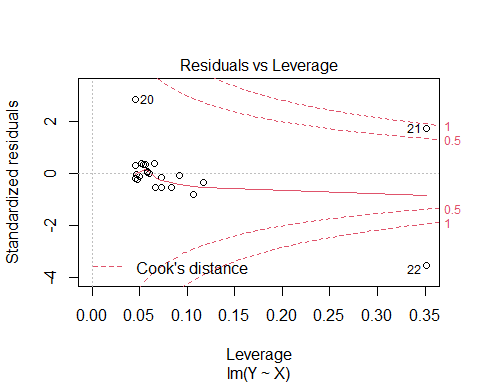
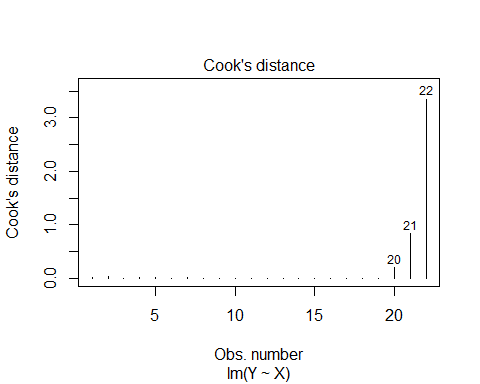
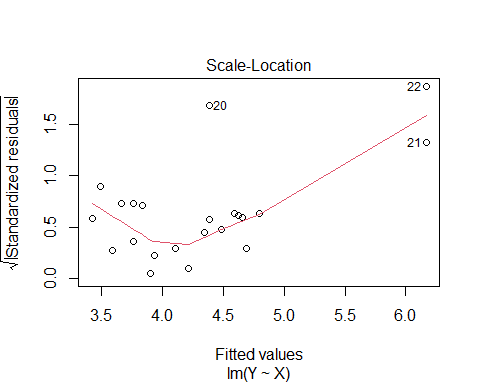
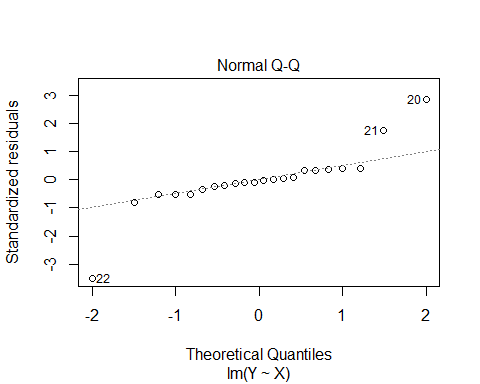
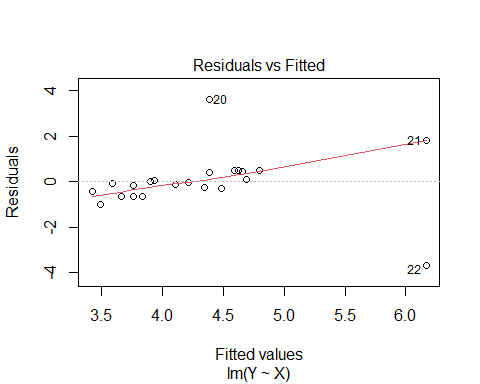
#  
# Modelo de regresión lineal  
#  
S(lm\_YX <- lm(Y ~ X, data = ATIP))

## Call: lm(formula = Y ~ X, data = ATIP)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.9087 0.6128 4.746 0.000123 \*\*\*  
## X 0.3430 0.1331 2.577 0.017987 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard deviation: 1.297 on 20 degrees of freedom  
## Multiple R-squared: 0.2493  
## F-statistic: 6.643 on 1 and 20 DF, p-value: 0.01799   
## AIC BIC   
## 77.79 81.06

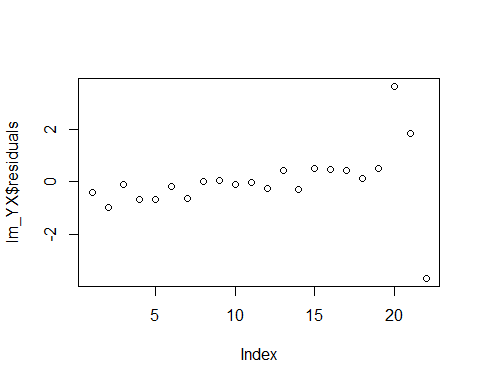
plot(Y ~ X , data=ATIP)  
abline(lm\_YX)



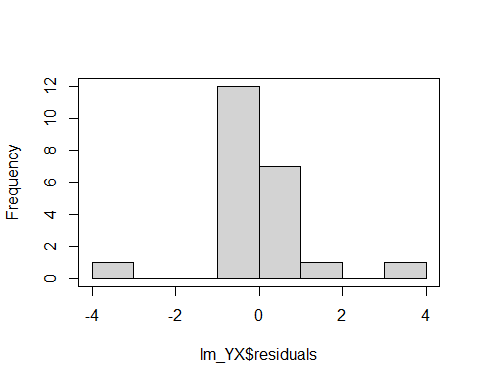
plot(lm\_YX, which = 1:6)



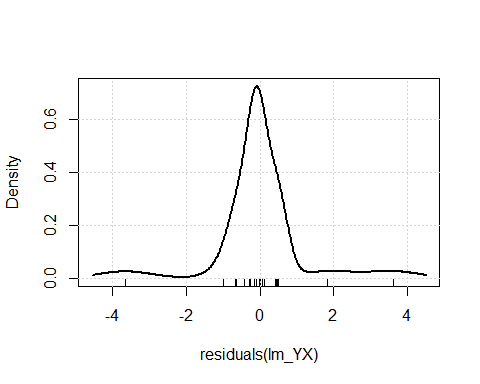
#  
# Distribución de los errores del modelo  
#  
plot(lm\_YX$residuals)



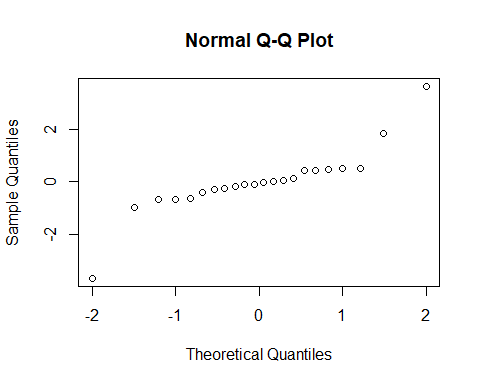
hist(lm\_YX$residuals, main = "")  
box()



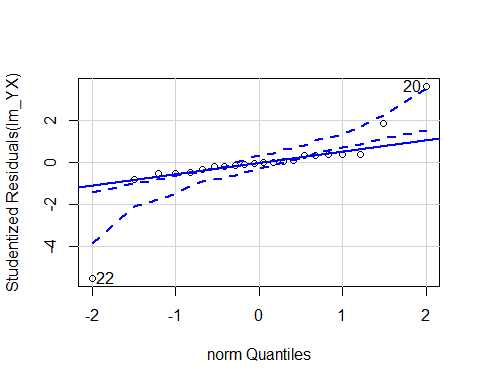
densityPlot(residuals(lm\_YX))



qqnorm(residuals(lm\_YX))

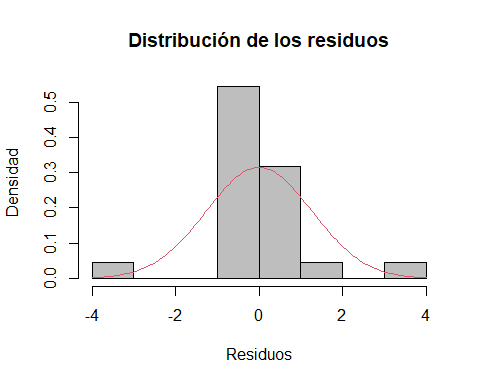


qqPlot(lm\_YX, distribution="norm")



## [1] 20 22

#  
# Contrastes de normalidad  
#  
r <- resid(lm\_YX)  
rbar <- mean(r)  
sdr <- sd(r)  
hist(lm\_YX$residuals, col="grey", freq=FALSE, main="Distribución de los residuos", ylab="Densidad", xlab="Residuos")  
curve(dnorm(x, rbar, sdr), col=2, add=TRUE, ylab="Densidad", xlab="r")



#  
# Librería moments

#  
skewness(lm\_YX$residuals)

## [1] 0.06517037

kurtosis(lm\_YX$residuals)

## [1] 7.089934

agostino.test(lm\_YX$residuals)

##   
## D'Agostino skewness test  
##   
## data: lm\_YX$residuals  
## skew = 0.06517, z = 0.15162, p-value = 0.8795  
## alternative hypothesis: data have a skewness

anscombe.test(lm\_YX$residuals)

##   
## Anscombe-Glynn kurtosis test  
##   
## data: lm\_YX$residuals  
## kurt = 7.0899, z = 3.0186, p-value = 0.002539  
## alternative hypothesis: kurtosis is not equal to 3

jarque.test(lm\_YX$residuals)

##   
## Jarque-Bera Normality Test  
##   
## data: lm\_YX$residuals  
## JB = 15.349, p-value = 0.0004645  
## alternative hypothesis: greater

# librería tseries  
jarque.bera.test(lm\_YX$residuals)

##   
## Jarque Bera Test  
##   
## data: lm\_YX$residuals  
## X-squared = 15.349, df = 2, p-value = 0.0004645

shapiro.test(lm\_YX$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: lm\_YX$residuals  
## W = 0.81712, p-value = 0.0009388

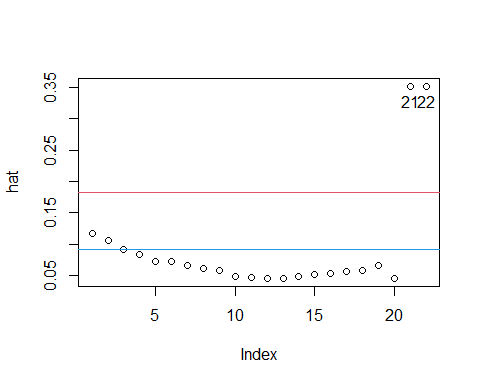
#  
# Detección de observaciones atípicas  
#  
# Observaciones atípicas en las variables explicativas (leverages <-> apalancamiento)  
#  
hat <- hatvalues(lm\_YX)  
hat

## 1 2 3 4 5 6 7   
## 0.11711229 0.10654749 0.09227928 0.08381979 0.07270953 0.07270953 0.06635534   
## 8 9 10 11 12 13 14   
## 0.06084327 0.05840303 0.04935980 0.04646022 0.04554154 0.04583820 0.04799135   
## 15 16 17 18 19 20 21   
## 0.05203927 0.05380964 0.05579054 0.05798197 0.06581944 0.04583820 0.35137515   
## 22   
## 0.35137515

which(hat > 2 \* mean(hat))

## 21 22   
## 21 22

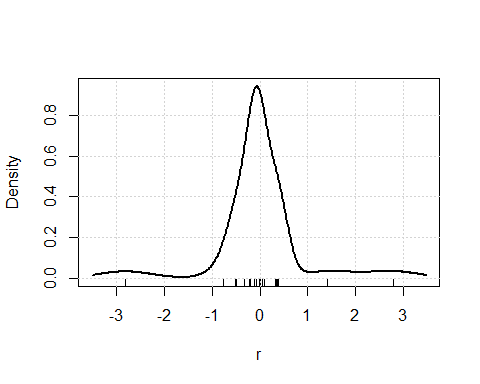
plot(hat)  
abline(h = mean(hat), col = 4)  
abline(h = 2 \* mean(hat), col = 2)  
id <- which(hat > 2 \* mean(hat))  
text(id, hat[id], rownames(ATIP)[id], pos = 1, xpd = TRUE)



#  
# Observaciones atípicas en la variable dependiente (outliers)  
#  
slm\_YX <- summary(lm\_YX)  
# Residuos estandarizados  
r <- lm\_YX$residuals/slm\_YX$sigma  
r

## 1 2 3 4 5 6   
## -0.326246342 -0.764575891 -0.073018040 -0.511347589 -0.513587999 -0.128144520   
## 7 8 9 10 11 12   
## -0.489385374 -0.002650574 0.047995086 -0.084220090 -0.009371804 -0.192232641   
## 13 14 15 16 17 18   
## 0.320945194 -0.220916086 0.393553070 0.367110034 0.340666999 0.082957877   
## 19 20 21 22   
## 0.389072250 2.787783458 1.412745624 -2.827132642

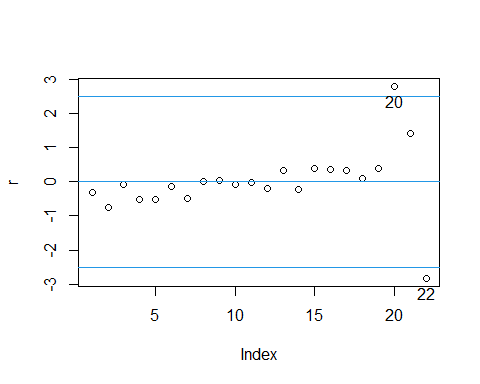
densityPlot(r)



which(abs(r) > 2.5)

## 20 22   
## 20 22

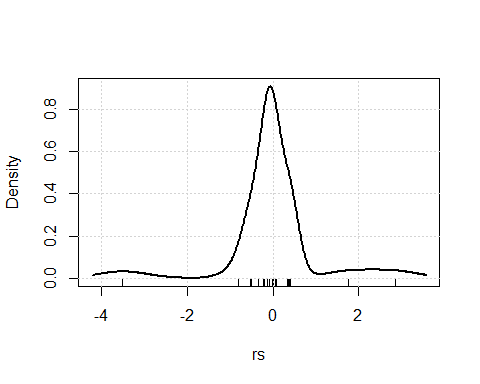
plot(r)  
abline(h = c(0,-2.5, 2.5), col = 4)  
id <- which(abs(r) > 2.5)  
text(id, r[id], rownames(ATIP)[id], pos = 1, xpd = TRUE)



# Residuos estudentizados (internamente)  
rs <- rstandard(lm\_YX)  
rs

## 1 2 3 4 5 6   
## -0.347210552 -0.808881428 -0.076639744 -0.534226907 -0.533343452 -0.133073672   
## 7 8 9 10 11 12   
## -0.506477522 -0.002735085 0.049461155 -0.086378889 -0.009597404 -0.196765348   
## 13 14 15 16 17 18   
## 0.328563914 -0.226415885 0.404211003 0.377404433 0.350587067 0.085472817   
## 19 20 21 22   
## 0.402545377 2.853960931 1.754151818 -3.510341691

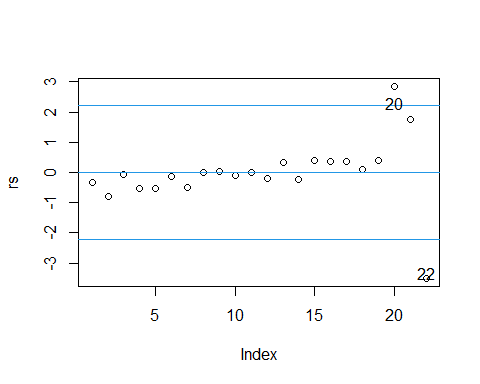
densityPlot(rs)



which(abs(rs) > 2)

## 20 22   
## 20 22

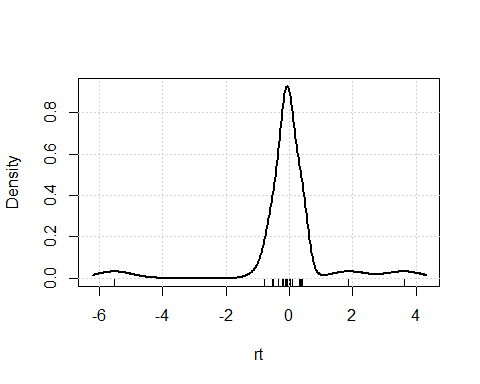
plot(rs)  
abline(h = c(0,-2, 2)\*sd(rs), col = 4)  
id <- which(abs(r) > 2\*sd(rs))  
text(id, r[id], rownames(ATIP)[id], pos = 1, xpd = TRUE)



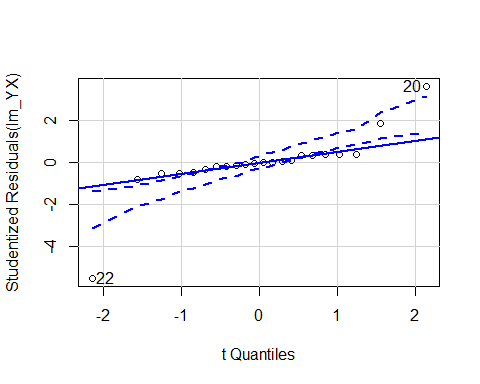
# Residuos estudentizados (externamente)  
rt <- rstudent(lm\_YX)  
rt

## 1 2 3 4 5 6   
## -0.339443573 -0.801621429 -0.074710153 -0.524455394 -0.523575574 -0.129761631   
## 7 8 9 10 11 12   
## -0.496849799 -0.002665832 0.048211719 -0.084207435 -0.009354414 -0.191969037   
## 13 14 15 16 17 18   
## 0.321112299 -0.220966280 0.395595340 0.369165231 0.342764870 0.083323817   
## 19 20 21 22   
## 0.393951873 3.613064780 1.858685559 -5.522252312

densityPlot(rt)



qqPlot(lm\_YX)



## [1] 20 22

outlierTest(lm\_YX)

## rstudent unadjusted p-value Bonferroni p  
## 22 -5.522252 2.5102e-05 0.00055224  
## 20 3.613065 1.8525e-03 0.04075600

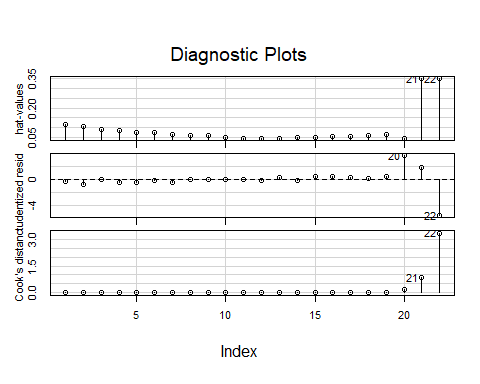
#  
# Diagnósticos  
#  
# Observaciones influyentes (cálculo de DFBETAS\_i, DFFITS\_i, COVRATIO\_i, DCOOK\_i y h\_i ; inf -> señalan obs. inusuales para al menos una medida)  
#  
influence.measures(lm\_YX)

## Influence measures of  
## lm(formula = Y ~ X, data = ATIP) :  
##   
## dfb.1\_ dfb.X dffit cov.r cook.d hat inf  
## 1 -0.121055 0.096704 -0.123628 1.240 8.00e-03 0.1171   
## 2 -0.268656 0.209618 -0.276825 1.160 3.90e-02 0.1065   
## 3 -0.022687 0.016968 -0.023821 1.220 2.99e-04 0.0923   
## 4 -0.148489 0.107322 -0.158632 1.175 1.31e-02 0.0838   
## 5 -0.132415 0.089762 -0.146611 1.161 1.12e-02 0.0727   
## 6 -0.032817 0.022246 -0.036336 1.193 6.94e-04 0.0727   
## 7 -0.115812 0.074339 -0.132456 1.157 9.12e-03 0.0664   
## 8 -0.000569 0.000341 -0.000679 1.180 2.42e-07 0.0608   
## 9 0.009826 -0.005654 0.012007 1.176 7.59e-05 0.0584   
## 10 -0.013126 0.005397 -0.019188 1.165 1.94e-04 0.0494   
## 11 -0.001193 0.000304 -0.002065 1.162 2.24e-06 0.0465   
## 12 -0.017270 -0.001833 -0.041933 1.156 9.24e-04 0.0455   
## 13 0.025883 0.006439 0.070382 1.149 2.59e-03 0.0458   
## 14 -0.011611 -0.011406 -0.049612 1.158 1.29e-03 0.0480   
## 15 0.009670 0.032970 0.092688 1.150 4.48e-03 0.0520   
## 16 0.005558 0.034690 0.088036 1.154 4.05e-03 0.0538   
## 17 0.001936 0.035862 0.083319 1.159 3.63e-03 0.0558   
## 18 -0.000315 0.009609 0.020672 1.175 2.25e-04 0.0580   
## 19 -0.012690 0.058166 0.104569 1.167 5.71e-03 0.0658   
## 20 0.291228 0.072449 0.791915 0.408 1.96e-01 0.0458 \*  
## 21 -0.917055 1.276478 1.368027 1.223 8.33e-01 0.3514 \*  
## 22 2.724619 -3.792484 -4.064480 0.252 3.34e+00 0.3514 \*

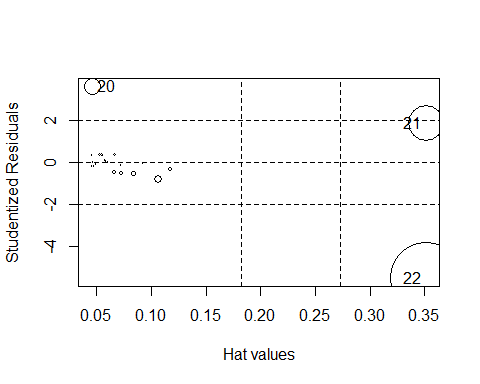
S(influence.measures(lm\_YX))

## Potentially influential observations of  
## lm(formula = Y ~ X, data = ATIP) :  
##   
## dfb.1\_ dfb.X dffit cov.r cook.d hat   
## 20 0.29 0.07 0.79 0.41\_\* 0.20 0.05   
## 21 -0.92 1.28\_\* 1.37\_\* 1.22 0.83\_\* 0.35\_\*  
## 22 2.72\_\* -3.79\_\* -4.06\_\* 0.25\_\* 3.34\_\* 0.35\_\*

influenceIndexPlot(lm\_YX, vars=c("Hat", "Studentized","Cook"))



influencePlot(lm\_YX, xlab="Hat values")



## StudRes Hat CookD  
## 20 3.613065 0.0458382 0.1956463  
## 21 1.858686 0.3513751 0.8334543  
## 22 -5.522252 0.3513751 3.3376919

# Medidas individuales  
hat <- hatvalues(lm\_YX)  
dfbetas <- dfbetas(lm\_YX)  
dffits <- dffits(lm\_YX)  
dcook <- cooks.distance(lm\_YX)  
hat ; dfbetas ; dffits; dcook

## 1 2 3 4 5 6 7   
## 0.11711229 0.10654749 0.09227928 0.08381979 0.07270953 0.07270953 0.06635534   
## 8 9 10 11 12 13 14   
## 0.06084327 0.05840303 0.04935980 0.04646022 0.04554154 0.04583820 0.04799135   
## 15 16 17 18 19 20 21   
## 0.05203927 0.05380964 0.05579054 0.05798197 0.06581944 0.04583820 0.35137515   
## 22   
## 0.35137515

## (Intercept) X  
## 1 -0.1210550547 0.0967044410  
## 2 -0.2686558904 0.2096182664  
## 3 -0.0226869829 0.0169684103  
## 4 -0.1484894369 0.1073215817  
## 5 -0.1324152881 0.0897623911  
## 6 -0.0328174664 0.0222464815  
## 7 -0.1158120753 0.0743388221  
## 8 -0.0005691882 0.0003412440  
## 9 0.0098255572 -0.0056536588  
## 10 -0.0131259608 0.0053971790  
## 11 -0.0011927906 0.0003037916  
## 12 -0.0172701507 -0.0018327463  
## 13 0.0258829797 0.0064389476  
## 14 -0.0116105156 -0.0114064102  
## 15 0.0096704719 0.0329704531  
## 16 0.0055580652 0.0346902377  
## 17 0.0019363561 0.0358623130  
## 18 -0.0003147545 0.0096088290  
## 19 -0.0126898190 0.0581659541  
## 20 0.2912279682 0.0724492173  
## 21 -0.9170553069 1.2764784676  
## 22 2.7246194301 -3.7924844977

## 1 2 3 4 5   
## -0.1236277852 -0.2768249321 -0.0238207591 -0.1586322104 -0.1466111950   
## 6 7 8 9 10   
## -0.0363357436 -0.1324561839 -0.0006785315 0.0120070906 -0.0191879700   
## 11 12 13 14 15   
## -0.0020648467 -0.0419330580 0.0703816904 -0.0496120344 0.0926875935   
## 16 17 18 19 20   
## 0.0880362600 0.0833186873 0.0206721685 0.1045694572 0.7919148784   
## 21 22   
## 1.3680269444 -4.0644798251

## 1 2 3 4 5 6   
## 7.995633e-03 3.901319e-02 2.985589e-04 1.305531e-02 1.115217e-02 6.942722e-04   
## 7 8 9 10 11 12   
## 9.115586e-03 2.423183e-07 7.586978e-05 1.937057e-04 2.243985e-06 9.236723e-04   
## 13 14 15 16 17 18   
## 2.593076e-03 1.292129e-03 4.484635e-03 4.050098e-03 3.631232e-03 2.248329e-04   
## 19 20 21 22   
## 5.708514e-03 1.956463e-01 8.334543e-01 3.337692e+00

#  
max(hatvalues(lm\_YX))

## [1] 0.3513751

which.max(hatvalues(lm\_YX))

## 21   
## 21

#  
max(abs(dffits(lm\_YX)))

## [1] 4.06448

which.max(abs(dffits(lm\_YX)))

## 22   
## 22

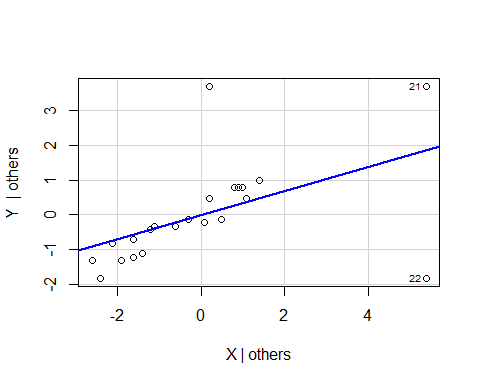
#  
max(cooks.distance(lm\_YX))

## [1] 3.337692

which.max(cooks.distance(lm\_YX))

## 22   
## 22

#  
# Gráficos de variable añadida, buscando casos influyentes  
avPlots(lm\_YX, id=list(cex=0.60, method="mahal"))



#  
# Regresión cuartilítica  
#  
S(lm\_YX <- lm(Y ~ X, data = ATIP))

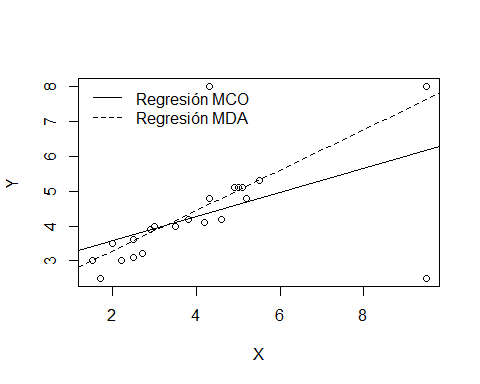
## Call: lm(formula = Y ~ X, data = ATIP)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.9087 0.6128 4.746 0.000123 \*\*\*  
## X 0.3430 0.1331 2.577 0.017987 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard deviation: 1.297 on 20 degrees of freedom  
## Multiple R-squared: 0.2493  
## F-statistic: 6.643 on 1 and 20 DF, p-value: 0.01799   
## AIC BIC   
## 77.79 81.06

S(qr\_YX <- rq(Y ~ X, data = ATIP)) # tau=0.5

## Warning in rq.fit.br(x, y, tau = tau, ci = TRUE, ...): Solution may be nonunique

##   
## Call: rq(formula = Y ~ X, data = ATIP)  
##   
## tau: [1] 0.5  
##   
## Coefficients:  
## coefficients lower bd upper bd  
## (Intercept) 2.13750 1.47445 2.67331   
## X 0.57500 0.32363 0.74276

plot(Y ~ X , data=ATIP)  
abline(lm\_YX)  
abline(qr\_YX, lty=2)  
legend("topleft", c("Regresión MCO", "Regresión MDA"), lty = c(1, 2), bty = "n")



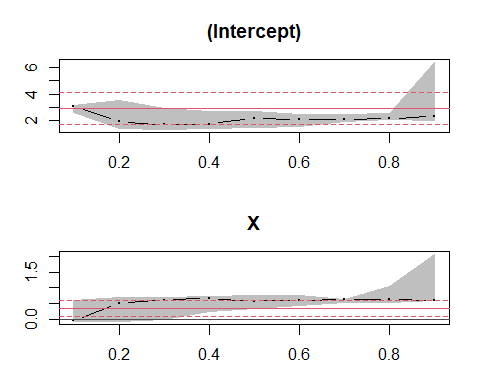
#  
# Secuencial  
S(qr\_YX <- rq(Y ~ X, data = ATIP, tau=seq(0.1,0.9,0.1)))

## Warning in rq.fit.br(x, y, tau = tau, ci = TRUE, ...): Solution may be nonunique

##   
## Call: rq(formula = Y ~ X, tau = seq(0.1, 0.9, 0.1), data = ATIP)  
##   
## tau: [1] 0.1  
##   
## Coefficients:  
## coefficients lower bd upper bd  
## (Intercept) 3.09375 2.61079 3.14417  
## X -0.06250 -0.06703 0.57699  
##   
## Call: rq(formula = Y ~ X, tau = seq(0.1, 0.9, 0.1), data = ATIP)  
##   
## tau: [1] 0.2  
##   
## Coefficients:  
## coefficients lower bd upper bd  
## (Intercept) 1.90000 1.40106 3.50461  
## X 0.50000 -0.08789 0.70048  
##   
## Call: rq(formula = Y ~ X, tau = seq(0.1, 0.9, 0.1), data = ATIP)  
##   
## tau: [1] 0.3  
##   
## Coefficients:  
## coefficients lower bd upper bd  
## (Intercept) 1.68000 1.29958 2.92467  
## X 0.60000 -0.00927 0.70201  
##   
## Call: rq(formula = Y ~ X, tau = seq(0.1, 0.9, 0.1), data = ATIP)  
##   
## tau: [1] 0.4  
##   
## Coefficients:  
## coefficients lower bd upper bd  
## (Intercept) 1.72500 1.37760 2.70125   
## X 0.65000 0.24673 0.70275   
##   
## Call: rq(formula = Y ~ X, tau = seq(0.1, 0.9, 0.1), data = ATIP)  
##   
## tau: [1] 0.5  
##   
## Coefficients:  
## coefficients lower bd upper bd  
## (Intercept) 2.13750 1.47445 2.67331   
## X 0.57500 0.32363 0.74276   
##   
## Call: rq(formula = Y ~ X, tau = seq(0.1, 0.9, 0.1), data = ATIP)  
##   
## tau: [1] 0.6  
##   
## Coefficients:  
## coefficients lower bd upper bd  
## (Intercept) 2.10000 1.53990 2.44895   
## X 0.60000 0.43072 0.74284   
##   
## Call: rq(formula = Y ~ X, tau = seq(0.1, 0.9, 0.1), data = ATIP)  
##   
## tau: [1] 0.7  
##   
## Coefficients:  
## coefficients lower bd upper bd  
## (Intercept) 2.07353 1.93579 2.41902   
## X 0.61765 0.51567 0.63299   
##   
## Call: rq(formula = Y ~ X, tau = seq(0.1, 0.9, 0.1), data = ATIP)  
##   
## tau: [1] 0.8  
##   
## Coefficients:  
## coefficients lower bd upper bd  
## (Intercept) 2.15385 2.06883 2.49759   
## X 0.61538 0.53634 1.04264   
##   
## Call: rq(formula = Y ~ X, tau = seq(0.1, 0.9, 0.1), data = ATIP)  
##   
## tau: [1] 0.9  
##   
## Coefficients:  
## coefficients lower bd upper bd  
## (Intercept) 2.30000 2.03634 6.40464   
## X 0.60000 0.60000 2.06745

plot(summary(qr\_YX), level=0.95)

## Warning in rq.fit.br(x, y, tau = tau, ci = TRUE, ...): Solution may be nonunique

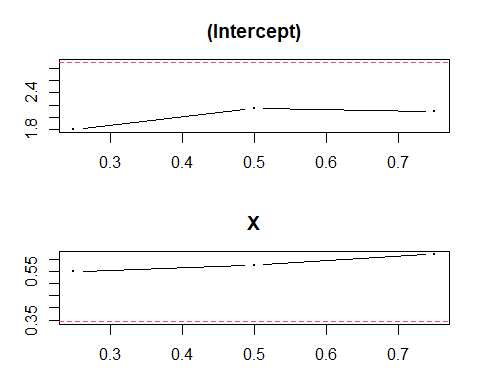


# Discreta  
S(qr\_YX <- rq(Y ~ X, tau = c(0.25, 0.50, 0.75), data = ATIP))

## Warning in rq.fit.br(x, y, tau = tau, ci = TRUE, ...): Solution may be nonunique  
  
## Warning in rq.fit.br(x, y, tau = tau, ci = TRUE, ...): Solution may be nonunique

##   
## Call: rq(formula = Y ~ X, tau = c(0.25, 0.5, 0.75), data = ATIP)  
##   
## tau: [1] 0.25  
##   
## Coefficients:  
## coefficients lower bd upper bd  
## (Intercept) 1.79000 1.39508 3.01749  
## X 0.55000 -0.10318 0.70372  
##   
## Call: rq(formula = Y ~ X, tau = c(0.25, 0.5, 0.75), data = ATIP)  
##   
## tau: [1] 0.5  
##   
## Coefficients:  
## coefficients lower bd upper bd  
## (Intercept) 2.13750 1.47445 2.67331   
## X 0.57500 0.32363 0.74276   
##   
## Call: rq(formula = Y ~ X, tau = c(0.25, 0.5, 0.75), data = ATIP)  
##   
## tau: [1] 0.75  
##   
## Coefficients:  
## coefficients lower bd upper bd  
## (Intercept) 2.09848 2.03503 2.39653   
## X 0.62121 0.52358 0.65685

plot(qr\_YX)



#