Relació entre les velocitats contractada i real en les línies ADSL.

César Alierta , Francisco Román, Fernando Ballestero

Resum

Objectiu: Determinar quina és la relació entre la velocitat de baixada real (VBR) i la contractada (VBC).

Mètodes: Entre el 23 de març i el 12 d'abril de 2010 es van recollir les dades de 41 alumnes de la FIB referents a les seves velocitats reals i contractades de connexió a Internet. Es va examinar la relació entre la VBR i la VBC mitjançant regressió lineal simple.

Resultats: La relació entre els logaritmes de VBR i VBC es pot expressar mitjançant l'equació:

$$\log(VBR) = 1.35 + 0.80 \cdot \log(VBC)$$

Els errors estàndards de la constant i del coeficient del log(VBC) són 0.59 i 0.07 respectivament.

Conclusió: A més VBC, més baixa el rati entre VBR/VBC. Concretament, doblar la VBC suposa multiplicar la VBR només en 1.74 .

Introducció

La diferència entre la velocitat contractada i la velocitat real en les connexions a Internet és una de les queixes més freqüents entre els usuaris de les línies ADSL. En un inici, les diferents companyies proveïdores d'Internet només garantien un velocitat de baixada del 10% i no va ser fins al juliol de 2008 que es va posar en marxa un decret[1] que obligava a les operadores a garantir un 80% de la velocitat contractada.

Objectiu

L'objectiu d'aquest estudi és determinar si es pot establir una relació entre la velocitat de baixada real (VBR) i la contractada (VBC) i obtenir la naturalesa d'aquesta relació.

Material y mètodes

Es va proposar recollir la informació sobre la connexió d'ADSL a 143 alumnes estudiants del tercer quadrimestre de la FIB. D'aquests, 41 van participar en l'estudi.

Els alumnes van realitzar una única prova de connexió entre els dies 23 de març i 12 d'abril de 2010. No es van incloure les mesures realitzades des de mòdems amb velocitats de 56 kb/s o inferiors, des de dispositius d'Internet mòbils o des de xarxes locals.

Les variables recollides per cada participant van ser: les velocitats contractades de pujada i baixada de dades proporcionades pel proveïdor del servei; les velocitats reals de pujada i baixada mesurades a través d'un únic test de velocitat (http://www.internautas.org/testvelocidad/); la latència de l'ADSL; el dia i l'hora de la prova; la distancia fins a la central (http://www.adslnet.es/distancia-adsl); el proveïdor del servei; el tipus de connexió (cable o wifi); i la ciutat des d'on es realitza la connexió.

La variable resposta en aquest estudi és la VBR i la variable predictora és la VBC.

Anàlisi estadística

La descriptiva de les variables numèriques es realitza a través de les mitjanes i Desviacions Estàndards (DE) i la de les variables categòriques mitjançant els percentatges de cadascuna d'elles. En l'anàlisi principal es van seguir els següents passos:

---

[1] Per respecte a la Competència Transversal *Ús Solvent Dels Recursos D'informació*, aquí hauríem d'indicar la referència

I. Modelatge

Es va realitzar un ajustament de la variable VBR en funció de VBC a través de regressió lineal simple.

II. Comprovació de les premisses i transformació de les variables:

Posteriorment, en l'anàlisi dels residus, es va avaluar:

a. Linealitat i Homoscedasticitat (residus en front de valors esperats),

b. Normalitat. (QQ-Norm),

c. Independència (residus en front dels residus retardats).

Com el model no es pot validar, s'han transformat logarítmicament les dues variables.

Resultats

Descriptiva

La taula 1 descriu les variables numèriques (mitjana i DE) i categòriques (número i percentatge) i mostra que les velocitats reals són inferiors a les contractades. La companyia amb més usuaris dins de la mostra és Telefònica i el percentatge d'usuaris que van optar per la connexió amb cable va ser lleugerament superior. En la mostra, també hi havia un major nombre d'usuaris de fora de Barcelona.

	n	Mitjana (DE)
Velocitat de baixada real	41	4198.5(2677.3)
Velocitat de baixada contractada	39	6406.6(5004.0)
Latència	41	142.6(66.9)
Distància a la central	41	1939.4(1545.2)

		n (%)
Proveïdor	Jazztel	3(7.5%)
	Ono	5(12.5%)
	Orange	2(5.0%)
	Telefónica	26(65.0%)
	Vodafone	1(2.5%)
	Ya.com	3(7.5%)
Connexió	Cable	23(56.1%)
	Wifi	18(43.9%)
Localització	Barcelona	10(24.4%)
	Altres	31(75.6%)

Taula 1: Descriptiva de les variables. Les velocitats estan expressades en Kb/s i les distàncies en metres.

Anàlisi principal

La figura 1 mostra la distribució de les dues variables que intervenen en la regressió, així com de la seva transformació logarítmica

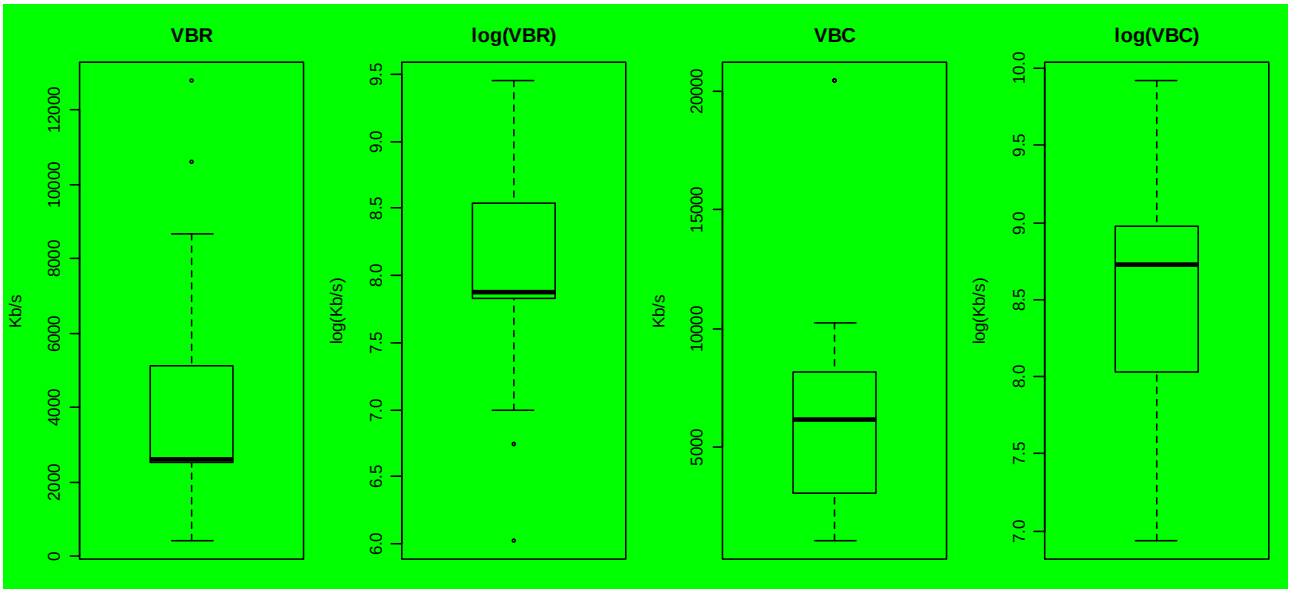


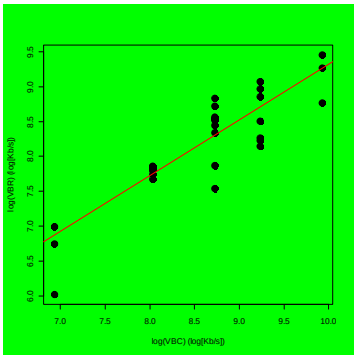
Figura 1: Distribució de les variables resposta i predictora i de les seves transformacions logarítmiques

El fet que en la VBR la mediana estigui situada en la part inferior de la capsa i que la longitud dels bigotis superiors sigui més gran indica una asimetria amb cues a la dreta. La transformació logarítmica millora la simetria. No obstant, la premissa de Normalitat s'ha d'estudiar amb l'anàlisi dels residus.

L'annex II mostra que l'ajust realitzat amb les variables VBR i VBC sense transformar no compleix la premissa d'homoscedasticitat. Com aplicar el logaritme natural únicament a la variable resposta no va donar resultats satisfactoris, es va aplicar aquesta transformació a les dues variables.

L'ajust del model lineal de log(VBR) en funció de log(VBC) és el següent:

```
Call:
lm(formula = adsl$log.obs.down ~ log(veloc.cont.down), data =
adsl)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85634 -0.04399  0.09101  0.16892  0.52709

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.34587    0.58609   2.296   0.0274 *
log(veloc.cont.down)  0.79768    0.06867  11.616 6.64e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3156 on 37 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.7848,     Adjusted R-squared: 0.779
F-statistic: 134.9 on 1 and 37 DF,  p-value: 6.639e-14
```



La transformació logarítmica ha aumentat la R^2 de 0.70 a 0.78. El model és:

$$\log(VBR) = 1.35 + 0.80 \cdot \log(VBC)$$

La figura 2 mostra els 4 gràfics de validació un cop fetes la transformacions. Els residus en front dels valors predits recolcen raonablement les premisses de linealitat i d'homogeneïtat de la variància, però el QQ-Norm i l'histograma mostren una cua allargada a l'esquerra. La premissa d'independència entre observacions consecutives sembla correcte, ja que no s'observa cap patró que indiqui relació entre els residus i els residus retardats (per altra banda, el mètode emprat per recollir les dades sembla robust per mantenir independència entre les dades).

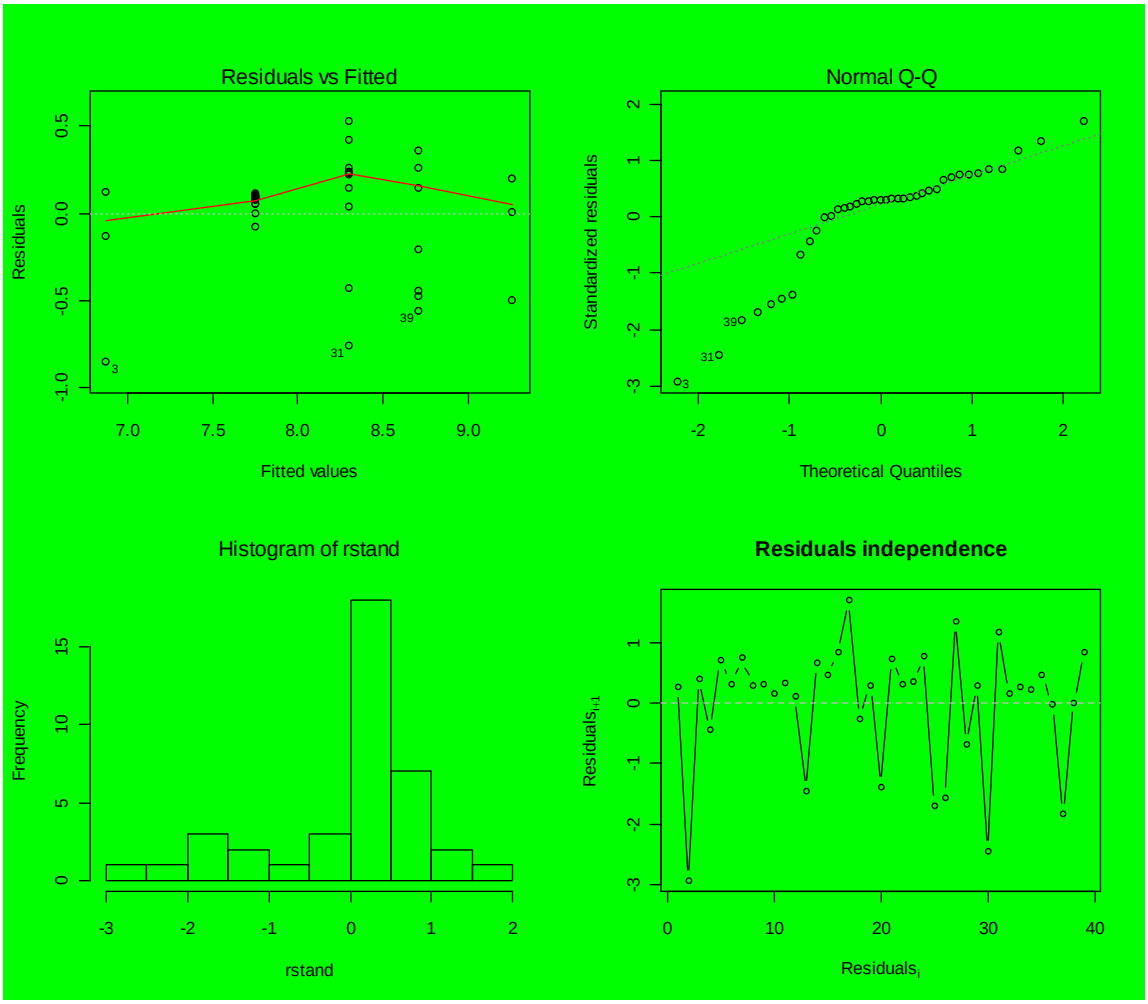


Figura 2: Gràfics de residus per avaluar les premisses amb les dades transformades

Discussió

En el model transformat, el coeficient de 0.80, $IC_{95\%}$ = [0.66 a 0.94] de la variable explicativa en el segon model s'ha d'interpretar com que la VBR s'incrementa en un 74% ($2^{0.8}$ = 1.74) $IC_{95\%}$: de 66% a 82% al incrementar un 100% la VBC. Això es dedueix com segueix:

$$\log(VBR) = 1.35 + \log(VBC) \Rightarrow VBR = e^{1.35} \cdot VBC^{0.8}$$

$$VBC' = 2 \cdot VBC \Rightarrow VBR' = e^{1.35} \cdot VBC'^{0.8} = e^{1.35} \cdot (2 \cdot VBC)^{0.8} = 2^{0.8} \cdot VBR$$

Dit en unes altres paraules, al augmentar la VBC està disminuint el rati VBR/VBC.

La R^2 representa la proporció de variabilitat explicada pel model. En els models ajustats aquesta oscil·la entre 0.70 i 0.78 que és un valor força alt i que indica que es pot predir amb força exactitud la VBR sabent la VBC. Així, la predicció de la VBR per a un individu concret que tingui una VBC igual a la mediana de les VBC's de la mostra (6144 Kb/s) serà de 4041 Kb/s amb un $IC_{95\%}$ de 2113

a 7727 Kb/s. És a dir, un usuari amb 6 Mb/s contractats, tindrà puntualment —amb un 95% de confiança— entre 2 i 7.5 Mb/s de VBR, aproximadament.

Per altra banda, el valor mig esperat de VBR per tots els usuaris amb aquesta velocitat contractada serà de 4040 Kb/s amb un $IC_{95\%}$ de 3631 a 4497 Kb/s. Això és equivalent a dir que la mitjana de la VBR per als individus que tenen 6 Mb/s contractats estarà amb una confiança del 95% entre 3.5 Mb/s i 4.4 Mb/s

Aquests resultats s'han de prendre amb molta cura ja que el model escollit no ha passat totes les proves de validació. Al forçar el compliment de la premissa d'homoscedasticitat, s'ha perdut la Normalitat dels residus

Un altre limitació és el baix nombre d'observacions per fer una validació gràfica acurada.

La principal limitació de l'anàlisi és que les dades provenen d'aproximadament el 40% dels voluntaris, el que provoca una incertesa addicional que no està contemplada en les mesures estadístiques de l'error aleatori. Això comporta que pot haver una infra-estimació de l'EE dels estimadors i, en conseqüència una amplada infra-estimada dels IC calculats.

Sobre la generalització, desconeixem si els voluntaris són una mostra representativa de tots els alumnes als quals se li va oferir participar en el estudi. La nostra mostra està composada per un grup molt homogeni d'alumnes universitaris de la mateixa carrera en la mateixa universitat compresos en una franja d'edat molt estreta. S'hauria de verificar si els resultats són extrapolables a altres poblacions en altres mostres.

Per la recerca futura, recomanem perfeccionar el modelatge, tot tenint en compte les altres variables explicatives que conté la base de dades per ser incloses en l'ajustament.

ANNEX I. SCRIPT EN R

```
#############################################################################
# Lectura de les dades
#############################################################################
colClasses <-
c(rep("factor",2),rep("numeric",3),rep("factor",2),"numeric","factor",rep("numer
ic",2),rep("factor",3),rep("numeric",4))
adsl <- read.table(url("http://www-
eio.upc.es/teaching/pe/DADES/dades_ADSL.txt"), header=TRUE,
na.strings='00',colClasses=colClasses,dec=",")


#############################################################################
# Depuració i recodificació
#############################################################################
adsl$dia <- as.Date(adsl$dia,"%d/%m/%y")          # Donar format de data
adsl$veloc.cont.down <- adsl$veloc.cont.down*1024 # Es passa de Mb/s a Kb/s
adsl$log.cont.down <- log(adsl$veloc.cont.down)   # Es treu el logaritme


#############################################################################
### Inspecció de les dades
#############################################################################
names(adsl)
head(adsl)
summary(adsl)


#############################################################################
### Descriptiva
#############################################################################

###### Matrix - plot
install.packages('car')
library(car)
varnum <-
c("down.speed","veloc.cont.down","up.speed","veloc.cont.up","latency","dist.cent
ral","Ratio.down","Ratio.up")
varfact <- c("proveedor","cable.o.wifi","is.BCN")
scatterplotMatrix(adsl[,varnum],diagonal="boxplot",smooth =
FALSE,pch=19,cex=0.8)

###### Descriptiva univariant variables numériques
summary(adsl[,varnum])
apply(adsl[,varnum],2,sd,na.rm=TRUE)
par(mfrow=c(2,4))
apply(adsl[,varnum],2,boxplot)

###### Taula descriptiva Global
Desc1 <- matrix(NA,nrow=6,ncol=3)
namesvarnum <- c("Velocitat de baixada real","Velocitat de baixada contractada",
"Velocitat de pujada real","Velocitat de pujada contractada",
"Latència","Distància a la central")
namesDes  <- c("n","mitjana","desviació")
rownames(Desc1) <- namesvarnum
colnames(Desc1) <- namesDes
for (i in 1:6){
      Desc1[i,1] <- sum(!is.na(adsl[,varnum[i]]))
      Desc1[i,2] <- mean(adsl[,varnum[i]],na.rm=TRUE)
      Desc1[i,3] <- sd(adsl[,varnum[i]],na.rm=TRUE)
}
```

```
###### Descriptiva univariant variables categòriques
summary(adsl[,varfact])
Table <- apply(adsl[,varfact],2,table)
P <- list()
par(mfrow=c(2,2))

for (i in 1:3){
      P[[i]] <- prop.table(Table[[i]])*100
      barplot(P[[i]],las=2)
}
print(P)


################################################################################
### Anàlisi principal
################################################################################

### Boxplot de les dues variables que intervenen en la regressió
par(mfrow=c(2,2),mar=c(1.5,4,3,1))
boxplot(adsl$down.speed,main="VBR",xlab="",ylab="Kb/s")
boxplot(log(adsl$down.speed),main="log(VBR)",xlab="",ylab="log(Kb/s)")

boxplot(adsl$veloc.cont.down,main="VBC",xlab="",ylab="Kb/s")
boxplot(log(adsl$veloc.cont.down),main="log(VBC)",xlab="",ylab="log(Kb/s)")

### Histograma per la versió dolenta
par(mfrow=c(1,2),mar=c(3.5,4,3,1))
hist(adsl$down.speed,main="",xlab="",ylab="")
hist(adsl$veloc.cont.down,main="",xlab="",ylab="")

################### Regressió lineal #############################################
mod.lm1 <- lm(down.speed~veloc.cont.down,data=adsl)
summary(mod.lm1)
confint(mod.lm1)

### Gràfic
plot(adsl$veloc.cont.down,adsl$down.speed,pch=19,cex=1.8,xlab="VBC
(Kb/s)",ylab="VBR (Kb/s)")
abline(mod.lm1$coef[1],mod.lm1$coef[2],col=2,lwd=2)

### Residus
par(mfrow=c(2,2))
plot(mod.lm1,which = c(2,1))
rstand <- rstandard(mod.lm1)
hist(rstand,font.main=1)
plot(1:length(rstand),rstand,cex=0.8,xlab=expression(Residuals[i]),ylab=expressi
on(Residuals[i+1]),type="b",main="Residuals independence")
abline(h=0,col="grey",lty=2)


##### Regressió lineal amb transformació logarítmica a les dues variables ######
mod.lm2 <- lm(log.obs.down~log.cont.down,data=adsl)
summary(mod.lm2)
confint(mod.lm2)

### Gràfic
par(mfrow=c(1,1))
plot(adsl$log.cont.down,adsl$log.obs.down,pch=19,cex=1.8,xlab="log(VBC)
(log[Kb/s])",ylab="log(VBR) (log[Kb/s])")
abline(mod.lm2$coef[1],mod.lm2$coef[2],col=2,lwd=2)
```

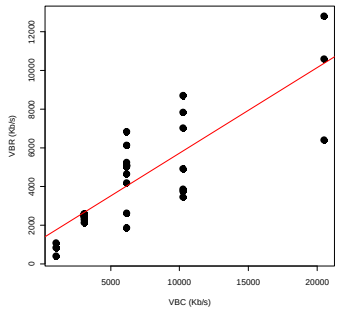
```
### Residus
par(mfrow=c(2,2))
plot(mod.lm2,which = c(2,1))
rstand <- rstandard(mod.lm2)
hist(rstand,font.main=1)
plot(1:length(rstand),rstand,cex=0.8,xlab=expression(Residuals[i]),ylab=expressi
on(Residuals[i+1]),type="b",main="Residuals independence")
abline(h=0,col="grey",lty=2)

### Predicció
med <- data.frame(log.cont.down=median(adsl$log.cont.down,na.rm=T))
predlog1 <- predict(mod.lm2,med,interval="prediction") # Predicció 1 pel log
predlog2 <- predict(mod.lm2,med,interval="confidence") # Predicció 2 pel log

exp(predlog1)        # Predicció per un individu
exp(predlog2)        # Predicció pel valor esperat
```

ANNEX II. MODELATGE SENSE TRANSFORMAR

Modelatge

Al ajustar un model lineal de VBR en funció de VBC s'obté la següent sortida en R:

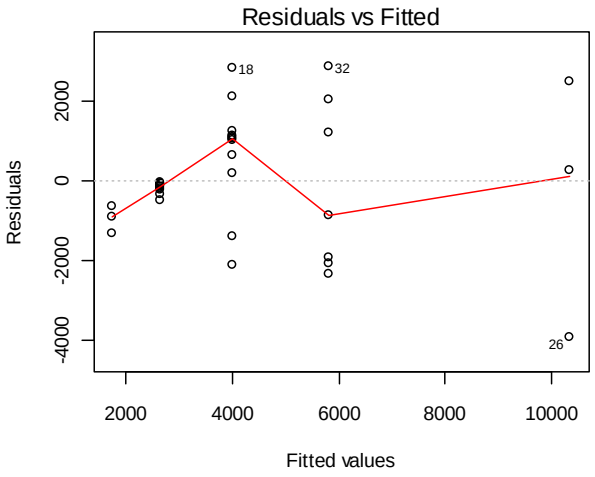
```
Call:
lm(formula = down.speed ~ veloc.cont.down, data = adsl)

Residuals:
    Min      1Q  Median      3Q     Max
-3945.9  -586.6  -108.6  1038.3  2869.2

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.299e+03  3.833e+02   3.388  0.00168 **
veloc.cont.down 4.414e-01  4.738e-02   9.317 3.04e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1461 on 37 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.7012,     Adjusted R-squared: 0.6931
F-statistic: 86.81 on 1 and 37 DF,  p-value: 3.042e-11
```

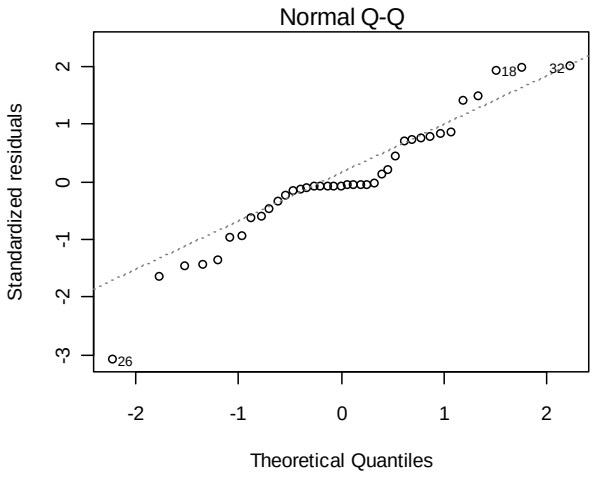


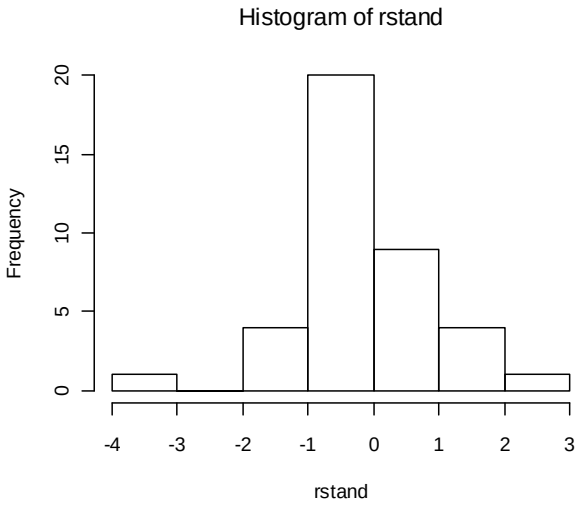
El coeficient de 0.44, $IC_{95\%}$ = [0.35 a 0.54] de la variable explicativa en el model s'ha d'interpretar com el guany de VBR per cada increment en una unitat de VBC. Aquest valor queda molt per sota en el millor dels casos del 80% d'eficiència exigit pel decret de l'any 2008. De totes formes, a continuació s'ha de validar a través de l'anàlisi dels residus.

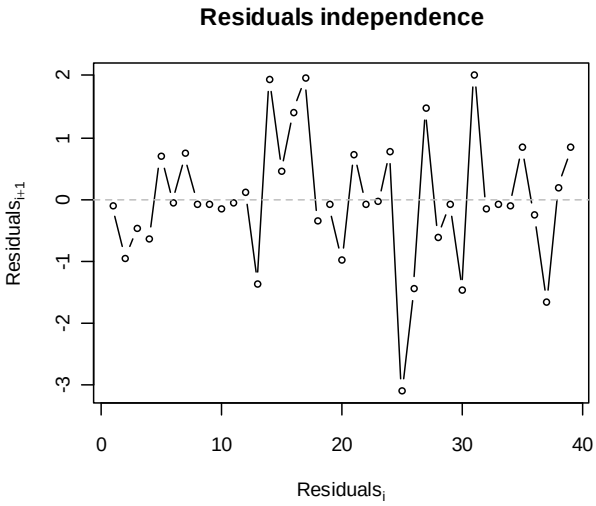
II. Comprovació de les premisses

La següent figura mostra els quatre gràfics de residus necessaris per fer l'anàlisi de les premisses. El baix nombre d'observacions i el fet que molts valors de la variable explicativa estiguin repetits complica extreure conclusions. El primer gràfic serveix per avaluar la linealitat i la homoscedasticitat. La linealitat sembla complir-se ja que el núvol de punts es distribueix simètricament respecte al 0 pels diferents valors predits. Altrament, la variància augmenta a mesura que augmenta el valor predit, fenomen que es coneix com heteroscedasticitats.


Residuals vs Fitted
Residuals
Fitted values
18
32
26
2000
4000
6000
8000
10000
-4000
-2000
0
2000


Normal Q-Q
Standardized residuals
Theoretical Quantiles
18
32
26
-2
-1
0
1
2
-3


Histogram of rstand
Frequency
rstand
-4
-3
-2
-1
0
1
2
3
0
5
10
15
20


Residuals independence
Residuals$_{i+1}$
Residuals$_i$
0
10
20
30
40
-3
-2
-1
0
1
2