

# Course Practical Assignment - 3rd Deliverable (juni 2019)

*Josep Clotet Ginovart*

*Eric Martin Obispo*



## Bank client data

### Description of input variables:

1. age (numeric)
2. job : type of job (categorical: ‘admin’, ‘blue-collar’, ‘entrepreneur’, ‘housemaid’, ‘management’, ‘retired’, ‘self-employed’, ‘services’, ‘student’, ‘technician’, ‘unemployed’, ‘unknown’)
3. marital : marital status (categorical: ‘divorced’,‘married’,‘single’,‘unknown’; note: ‘divorced’ means divorced or widowed)
4. education (categorical:‘basic.4y’, ‘basic.6y’, ‘basic.9y’, ‘high.school’, ‘illiterate’, ‘professional.course’, ‘university.degree’, ‘unknown’)
5. default: has credit in default? (categorical: ‘no’,‘yes’,‘unknown’)
6. housing: has housing loan? (categorical: ‘no’,‘yes’,‘unknown’)
7. loan: has personal loan? (categorical: ‘no’,‘yes’,‘unknown’) # related with the last contact of the current campaign:
8. contact: contact communication type (categorical:‘cellular’,‘telephone’)
9. month: last contact month of year (categorical: ‘jan’, ‘feb’, ‘mar’,..., ‘nov’, ‘dec’)
10. day\_of\_week: last contact day of the week (categorical:‘mon’,‘tue’,‘wed’,‘thu’,‘fri’)
11. duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y=‘no’). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
12. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14. previous: number of contacts performed before this campaign and for this client (numeric)
15. poutcome: outcome of the previous marketing campaign (categorical: ‘failure’,‘nonexistent’,‘success’) # social and economic context attributes
16. emp.var.rate: employment variation rate - quarterly indicator (numeric)
17. cons.price.idx: consumer price index - monthly indicator (numeric)
18. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19. euribor3m: euribor 3 month rate - daily indicator (numeric)
20. nr.employed: number of employees - quarterly indicator (numeric)
21. y - has the client subscribed a term deposit? (binary: ‘yes’,‘no’)



### Loading packages:

### Load validated data from Deliverable 1:

```
#invisible() prevent hte output in console of the function
#dirwd<- "D:/Users/Usuari/Documents/ADEIpractica"
#dirwd<- "//pax/perfils/1173408.CR/Downloads/deliverable"
dirwd<- "D:/Documents/GitHub/ADEI"
setwd(dirwd)

load( paste0(dirwd, "/bank-additional/Bank5000_validated.RData") )
```

```
summary(df)
```

```
##      age          job          marital
## Min.   :18.00   job-admin.    :1246   marital-divorced: 554
## 1st Qu.:32.00   job-blue-collar:1171   marital-married  :3055
## Median :38.00   job-technician : 796   marital-single   :1377
## Mean   :40.07   job-services   : 498
## 3rd Qu.:47.00   job-management: 411
## Max.   :87.00   job-retired   : 205
## (Other)        : 659
##                  education          default
## education-basic.4y     : 533   default-no     :3954
## education-basic.6y     : 289   default-unknown:1032
## education-basic.9y     : 767
## education-high.school :1218
## education-professional.course: 615
## education-university.degree  :1564
##
##      housing          loan          contact
## housing-no :2261   loan-no :4217   contact-cellular :3122
## housing-yes:2725  loan-yes: 769   contact-telephone:1864
##
##      month          day_of_week          duration
## month-may:1741   day_of_week-1mon:1016   Min.   : 5.0
## month-jul: 829   day_of_week-2tue:1043   1st Qu.:101.0
## month-aug: 697   day_of_week-3wed: 971   Median :177.0
## month-jun: 652   day_of_week-4thu:1034   Mean   :250.6
## month-nov: 507   day_of_week-5fri: 922   3rd Qu.:316.0
## month-apr: 310
## (Other)   : 250   Max.   :1580.0
##      campaign          pdays          previous
## Min.   : 1.000   Min.   : 0.00   Min.   :0.0000
## 1st Qu.: 1.000   1st Qu.:19.00   1st Qu.:0.0000
## Median : 2.000   Median :19.00   Median :0.0000
## Mean   : 2.535   Mean   :18.53   Mean   :0.1598
## 3rd Qu.: 3.000   3rd Qu.:19.00   3rd Qu.:0.0000
## Max.   :25.000   Max.   :19.00   Max.   :4.0000
##
##      poutcome          emp.var.rate          cons.price.idx
## poutcome-failure     : 477   Min.   :-3.40000   Min.   :92.20
## poutcome-nonexistent:4353  1st Qu.:-1.80000   1st Qu.:93.08
## poutcome-success     : 156   Median : 1.10000   Median :93.75
##                               Mean   : 0.06446   Mean   :93.57
##                               3rd Qu.: 1.40000   3rd Qu.:93.99
##                               Max.   : 1.40000   Max.   :94.77
##
##      cons.conf.idx          euribor3m          nr.employed          y
## Min.   :-50.80   Min.   :0.635   Min.   :4964   y-no :4429
## 1st Qu.:-42.70   1st Qu.:1.334   1st Qu.:5099   y-yes: 557
## Median :-41.80   Median :4.857   Median :5191
```

```

##  Mean    :-40.43   Mean    :3.614   Mean    :5166
##  3rd Qu.:-36.40   3rd Qu.:4.961   3rd Qu.:5228
##  Max.   :-26.90   Max.   :5.000   Max.   :5228
##
##      num_missings      num_outliers      num_errors
##  Min.   :0.0000   Min.   :0.00000   Min.   :0
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0
##  Median :0.0000   Median :0.00000   Median :0
##  Mean   :0.1111   Mean   :0.00361   Mean   :0
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0
##  Max.   :3.0000   Max.   :2.00000   Max.   :0
##
##          f.season      minutes           f.age
##  season-spring     :2117   Min.   : 0.08333   f.age-[18,32]:1352
##  season-summer     :2178   1st Qu.: 1.68333   f.age-(32,38]:1205
##  season-autumnwinter: 691   Median : 2.95000   f.age-(38,47]:1220
##                                         Mean   : 4.17703   f.age-(47,87]:1209
##                                         3rd Qu.: 5.26667
##                                         Max.   :26.33333
##
##          f.duration           f.campaign
##  f.duration-[5,101]   :1252   f.campaign-[0,2]  :3392
##  f.duration-(101,177]  :1243   f.campaign-(2,5]  :1181
##  f.duration-(177,316]  :1247   f.campaign-(5,25]: 413
##  f.duration-(316,1.58e+03] :1244
##
##      f.pdays           f.previous
##  f.pdays-sometime: 177   f.previous-never:4353
##  f.pdays-never   :4809   f.previous-some  : 633
##
##      f.emp.var.rate           f.cons.price.idx
##  f.emp.var.rate-[-Inf,0]  :2086   f.cons.price.idx-[92.2,93.1]:1409
##  f.emp.var.rate-(0, Inf]  :2900   f.cons.price.idx-(93.1,93.7]:1086
##                                         f.cons.price.idx-(93.7,94]  :1819
##                                         f.cons.price.idx-(94,94.8]  : 672
##
##      f.cons.conf.idx           f.euribor3m
##  f.cons.conf.idx-[-50.8,-42.7] :1856   f.euribor3m-[0.635,1.33]:1254
##  f.cons.conf.idx-(-42.7,-41.8]: 967   f.euribor3m-(1.33,4.86] :1466
##  f.cons.conf.idx-(-41.8,-36.4] :1231   f.euribor3m-(4.86,4.96] :1130
##  f.cons.conf.idx-(-36.4,-26.9]: 932   f.euribor3m-(4.96,5]    :1136
##
##      f.nr.employed
##  f.nr.employed-[4.96e+03,5.1e+03] :1639

```

```

##   f.nr.employed-(5.1e+03,5.19e+03] :1003
##   f.nr.employed-(5.19e+03,5.23e+03] :2344
##
##
```

## Linear Model Building - target numeric “duration” de la trucada

Per tal d'elaborar un model lineal que predigi el valor de la variable numerica target *duration*, primer hem de decidir quines son les variables que utilitzarem en la seva construccio. En altres paraules, trobar quines variables ens aporten informacio i precisio al model predictiu, pero sense sobreparametritzar-lo.

### Variables numeriques explicatives pel target numeric

#### Model inicial amb totes les variables numeriques

Una primera (i dolenta) aproximacio podria ser la d'usar un model lineal inicial que tingui en compte totes les variables numeriques aportades. Veiem com aquestes variables en un model lineal nomes ens expliquen l' 1.3% de la variabilitat de l'output *duration* (Multiple R-squared: 0.01309)! El que vol dir que gairebe un 99% d'questa variabilitat de la duracio queda sense explicar; aixi doncs aquesta primera aproximacio, a part d'estar sobreparametritzada, no prediu gens be.

```
vars_con
```

```

## [1] "age"           "duration"        "campaign"       "pdays"
## [5] "previous"      "emp.var.rate"    "cons.price.idx" "cons.conf.idx"
## [9] "euribor3m"     "nr.employed"

m1<-lm(duration~., data=df[,vars_con]) #es passa nomes el df amb les variables continues perque
#li hem posat un '.' al model que vol dir que les agafi totes, i ara nomes volem variables
#explicatives numeriques; si especificiquem les variables numeriques explicatives del model i volem
#que el calculi per tot el df incloent variables categoriques, "data" el passem com =df.
summary(m1)
```

```

##
## Call:
## lm(formula = duration ~ ., data = df[, vars_con])
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -363.78 -146.18  -73.40   60.34 1336.54
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1509.2349  2363.9473  0.638  0.5232
## age          -0.1649    0.3135 -0.526  0.5988
## campaign     -6.9311    1.3298 -5.212 1.94e-07 ***
## pdays         -3.2592    1.5930 -2.046  0.0408 *
## previous     -20.6929    9.1633 -2.258  0.0240 *
## emp.var.rate  25.7481   11.9837  2.149  0.0317 *
## cons.price.idx  9.8281   14.1899  0.693  0.4886
## cons.conf.idx -0.9407    1.1320 -0.831  0.4060
## euribor3m    -13.2088   14.9069 -0.886  0.3756
## nr.employed   -0.4030    0.2522 -1.598  0.1101
## ---
```

```

## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.1 on 4976 degrees of freedom
## Multiple R-squared: 0.01309, Adjusted R-squared: 0.0113
## F-statistic: 7.333 on 9 and 4976 DF, p-value: 1.097e-10

```

### Model inicial amb noms les variables numeriques rellevants

Una altra opció mes adient seria la d'obtenir un model inicial utilitzant noms les variables numeriques que son rellevants, i a partir d'aquí mirar si es pot reduir la parametrització del model i seguir amb un bon ajust predictiu de la variabilitat del nostre output *duration*. Per a trobar les variables rellevants, podem realitzar tests de Fisher mitjançant la comanda Anova d'R, o bé utilitzar la comanda condes vista en anteriors entregues.

Inferential criteria o Bayesian info criteria Utilitzem la comanda Anova per a realitzar tests de Fisher i detectar i eliminar variables poc explicatives en els models. El test Anova ens diu línia a línia si cada variable es significativa a l'hora d'aportar informació en el model. Cada fila es refereix a un test de models encaixats del model m1 amb el model m1 sense la variable expressada en la fila. Per tant, si el p-valor es <0.05 podem refutar la H0 que deia que els models eren iguals. La podem refutar amb les variables age i indicadors socioeconomics, que vol dir que no ens aporten informació extra al model. En canvi, per les variables campaign, pdays, previous i emp.var.rate, no podem refutar la H0, el que vol dir que si que ens estan aportant informació al model i no les podem eliminar. Podriem contemplar també quedar-nos amb nr.employed, ja que esta prop de la frontera del p-valor teòric vs la flexibilitat a la pràctica. El model m2 es el model obtingut amb aquestes variables rellevants.

**#METODE TESTS FISHER:**

**#remove non significant variables, per a saber quines son fem tests de Fisher amb la comanda Anova d'R**  
**Anova(m1)**

```

## Anova Table (Type II tests)
##
## Response: duration
##          Sum Sq Df F value    Pr(>F)
## age        14529  1  0.2768  0.59883
## campaign  1425963  1 1663 1.942e-07 ***
## pdays      219732  1 1861  0.04081 *
## previous   267683  1  5.0997  0.02397 *
## emp.var.rate 242319  1  4.6165  0.03171 *
## cons.price.idx 25180  1  0.4797  0.48858
## cons.conf.idx 36248  1  0.6906  0.40601
## euribor3m   41212  1  0.7851  0.37562
## nr.employed 134032  1  2.5535  0.11012
## Residuals  261191266 4976
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
m2<-lm(duration~campaign+pdays+previous+emp.var.rate+nr.employed, data=df)
Anova(m2)

```

```

## Anova Table (Type II tests)
##
## Response: duration
##          Sum Sq Df F value    Pr(>F)
## campaign  1354099  1 25.7778 3.970e-07 ***
## pdays     159994  1  3.0458  0.08101 .
## previous   242160  1  4.6100  0.03183 *
## emp.var.rate 894268  1 17.0240 3.751e-05 ***

```

```

## nr.employed      1275709      1 24.2855 8.576e-07 ***
## Residuals      261597980 4980
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Un altre metode pas a pas es el metode Akaike. Va eliminant en models successius les variables que treient-les, obtindriem un AIC mes baix, ja que ens interessa un coeficient d'Akaike que quan no li treiem cap variable es mantingui igual. Un trade-off entre el fitting del model i la sobre parametritzacio. El model m3  $duration \sim campaign + pdays + previous + emp.var.rate + cons.conf.idx + nr.employed$  es el model obtingut amb aquest metode.

#### #AKAIKE:

```
m3<-step(m1)
```

Hi ha un altre metode anomenat Bayesian, el qual es millor per a mostres grans com la nostra, ja que tot i funcionar com l'anterior, solem sortir models mes simplificats. El model m4  $duration \sim campaign + emp.var.rate + nr.employed$  es el model obtingut amb aquest metode.

#### #BAYESIAN (BIC):

```
m4<-step(m1, k=log(nrow(df)))
```

```

## Start:  AIC=54264.89
## duration ~ age + campaign + pdays + previous + emp.var.rate +
##           cons.price.idx + cons.conf.idx + euribor3m + nr.employed
##
##              Df Sum of Sq      RSS      AIC
## - age             1    14529 261205795 54257
## - cons.price.idx  1    25180 261216446 54257
## - cons.conf.idx   1    36248 261227514 54257
## - euribor3m       1    41212 261232478 54257
## - nr.employed     1    134032 261325298 54259
## - pdays           1    219732 261410997 54261
## - emp.var.rate    1    242319 261433584 54261
## - previous         1    267683 261458948 54261
## <none>                  261191266 54265
## - campaign        1    1425963 262617229 54284
##
## Step:  AIC=54256.65
## duration ~ campaign + pdays + previous + emp.var.rate + cons.price.idx +
##           cons.conf.idx + euribor3m + nr.employed
##
##              Df Sum of Sq      RSS      AIC
## - cons.price.idx  1    25242 261231037 54249
## - cons.conf.idx   1    39406 261245201 54249
## - euribor3m       1    42216 261248011 54249
## - nr.employed     1    131927 261337722 54251
## - pdays           1    218978 261424773 54252
## - emp.var.rate    1    243151 261448946 54253
## - previous         1    270229 261476024 54253
## <none>                  261205795 54257
## - campaign        1    1426321 262632115 54275
##
## Step:  AIC=54248.62
## duration ~ campaign + pdays + previous + emp.var.rate + cons.conf.idx +
##           euribor3m + nr.employed
##

```

```

##                                     Df Sum of Sq      RSS      AIC
## - euribor3m          1    25603 261256640 54241
## - cons.conf.idx     1    103360 261334397 54242
## - pdays             1    230438 261461474 54244
## - previous          1    275181 261506218 54245
## - nr.employed       1    438597 261669634 54248
## <none>                  261231037 54249
## - emp.var.rate      1    447006 261678042 54249
## - campaign          1    1413437 262644474 54267
##
## Step:  AIC=54240.59
## duration ~ campaign + pdays + previous + emp.var.rate + cons.conf.idx +
##           nr.employed
##
##                                     Df Sum of Sq      RSS      AIC
## - pdays             1    226621 261483260 54236
## - previous          1    271870 261528510 54237
## - cons.conf.idx     1    341340 261597980 54239
## <none>                  261256640 54241
## - emp.var.rate      1    1109140 262365780 54253
## - campaign          1    1392303 262648943 54259
## - nr.employed       1    1458307 262714947 54260
##
## Step:  AIC=54236.4
## duration ~ campaign + previous + emp.var.rate + cons.conf.idx +
##           nr.employed
##
##                                     Df Sum of Sq      RSS      AIC
## - previous          1    106252 261589512 54230
## - cons.conf.idx     1    274714 261757974 54233
## <none>                  261483260 54236
## - emp.var.rate      1    1158906 262642166 54250
## - campaign          1    1389083 262872343 54254
## - nr.employed       1    1566560 263049820 54258
##
## Step:  AIC=54229.91
## duration ~ campaign + emp.var.rate + cons.conf.idx + nr.employed
##
##                                     Df Sum of Sq      RSS      AIC
## - cons.conf.idx     1    281879 261871391 54227
## <none>                  261589512 54230
## - emp.var.rate      1    1125903 262715415 54243
## - campaign          1    1382960 262972472 54248
## - nr.employed       1    1460309 263049821 54249
##
## Step:  AIC=54226.77
## duration ~ campaign + emp.var.rate + nr.employed
##
##                                     Df Sum of Sq      RSS      AIC
## <none>                  261871391 54227
## - emp.var.rate      1    927321 262798712 54236
## - nr.employed       1   1277524 263148915 54243
## - campaign          1   1348382 263219773 54244

```

```

summary(m4)

##
## Call:
## lm(formula = duration ~ campaign + emp.var.rate + nr.employed,
##      data = df[, vars_con])
##
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -342.28 -147.72 - 73.63   62.68 1318.28 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 3091.9562   572.8442   5.398 7.07e-08 ***
## campaign     -6.7107    1.3250  -5.065 4.24e-07 ***
## emp.var.rate  21.1243    5.0293   4.200 2.71e-05 ***
## nr.employed   -0.5469    0.1109  -4.930 8.49e-07 *** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.3 on 4982 degrees of freedom
## Multiple R-squared:  0.01052,   Adjusted R-squared:  0.009924 
## F-statistic: 17.66 on 3 and 4982 DF,  p-value: 2.131e-11

```

Conedes per a obtenir les variables numeriques rellevants Utilitzant la comanda condes vista en anteriors entregues tambe podem trobar les variables que son rellevants pel nostre model. El model que obtenim així es el m5, i la comanda Anova ens diu que pdays no esta aportant res de nou (p-value=0.38). La treiem i ens queda el model m6, que correspon a *duration ~ euribor3m + nr.employed + campaign*. Si apliquem un metode BIC en aquest model obtenim un model m7 amb nomes un sol parametre aportant informacio que es campaign. Aquest model es massa simple i no ens va be per a treballar, així que ens quedem amb el model m4 obtingut anteriorment tambe pel metode Bayesian.

```

condes(df, 11)
#variable target: 11 (duration)

#Agafem com a variables explicatives les $quanti del condes:
m5<-lm(duration~pdays+euribor3m+nr.employed+campaign, data=df); Anova(m5)
m6<-lm(duration~euribor3m+nr.employed+campaign, data=df); Anova(m6)
#totes les variables ens aporten informacio nova

#BIC
m7<-step(m6, k=log(nrow(df)))

```

La comanda vif d'R ens diu les variables utilitzades en el model tenen redundancies. Si el seu valor esta per sota de 3 es valid; i si dos valors son iguals vol dir que d'aquelles dues variables ens n'hem de quedar nomes una! En el cas del model m4 ens quedarem amb la variable campaign i nr.employed (triada d'entre les dues amb valor similar), i obtindrem així el model m44.

```

vif(m4)

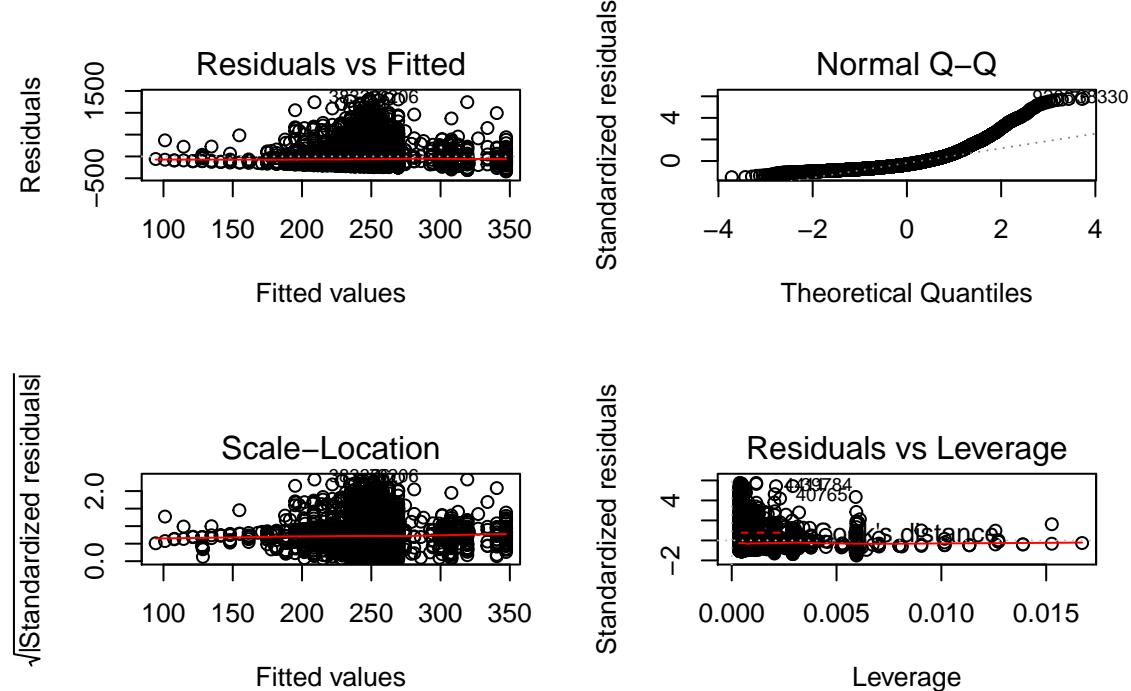
##          campaign emp.var.rate  nr.employed
##          1.024848    6.027929    6.013152
m44<-lm(duration~campaign+nr.employed, data=df)

```

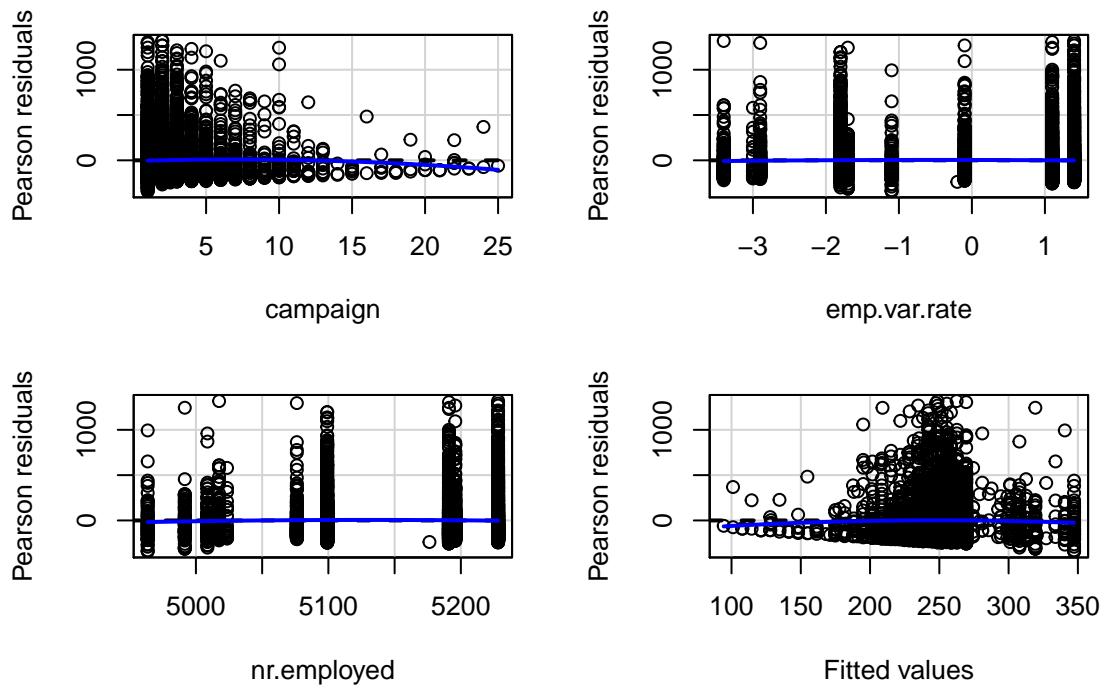
## Plots del model m4

Podem veure en el diagnostic del model que no es gens bo, a continuació doncs, farem un seguit de transformacions i hi afegirem variables factor explicatives. D'aquesta manera arribarem al nostre model definitiu, el qual diagnosticarem mes en profunditat.

```
par(mfrow=c(2,2))
plot(m4) #models forca dolents
```



```
par(mfrow=c(1,1))
residualPlots(m4)
```

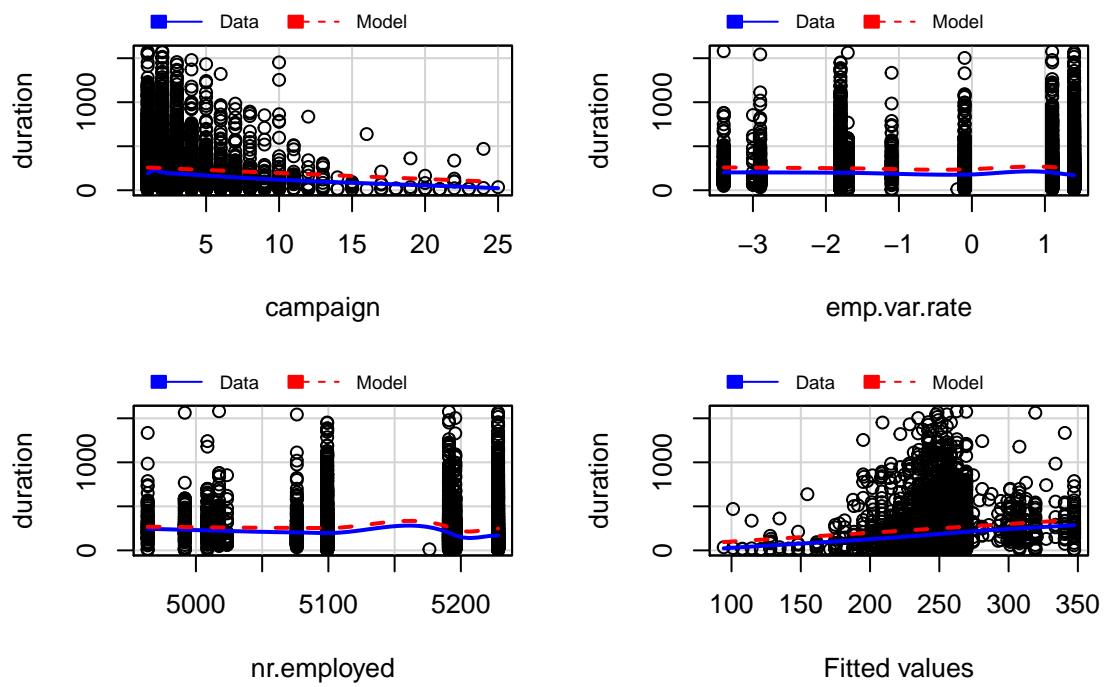


```

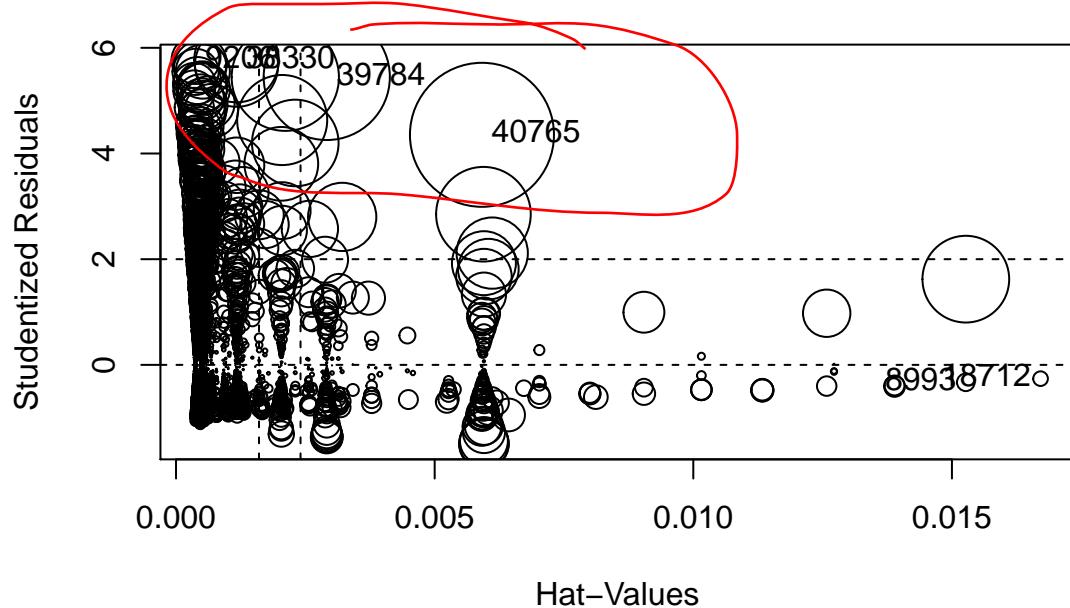
##           Test stat Pr(>|Test stat|)
## campaign      -1.8909    0.05870 .
## emp.var.rate   -0.7423    0.45794
## nr.employed    -1.6329    0.10255
## Tukey test     -2.4163    0.01568 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
marginalModelPlots(m4)

```

## Marginal Model Plots



```
influencePlot(m4)
```

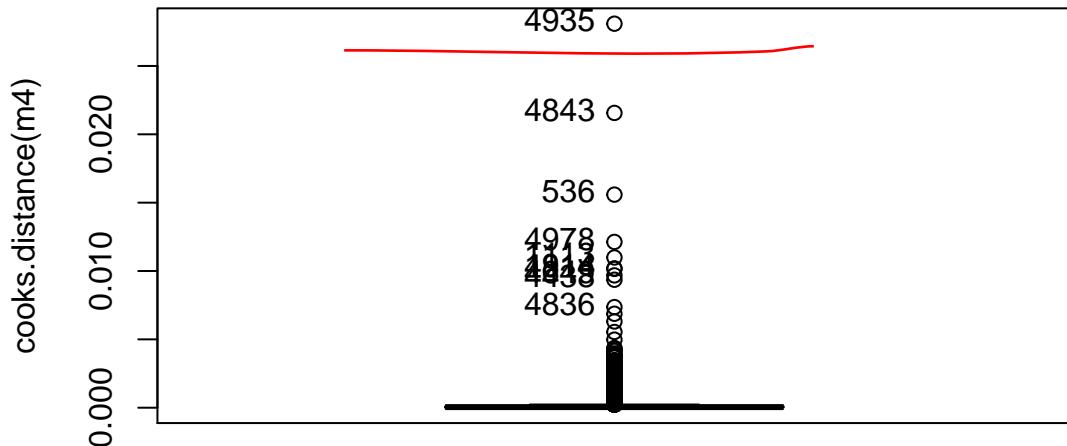


```
##          StudRes      Hat      CookD
## 8993   -0.3300087 0.0152684634 0.0004222271
## 9206    5.7696545 0.0003776632 0.0031239403
```

```

## 18712 -0.2611508 0.0167134662 0.0002898615
## 38330  5.7684911 0.0011697218 0.0096794548
## 39784  5.4482303 0.0029140821 0.0215638797
## 40765  4.3537452 0.0059141513 0.0280913224
Boxplot(cooks.distance(m4))

```



```

## [1] 4935 4843 536 4978 1113 1914 4814 4649 4438 4836

```

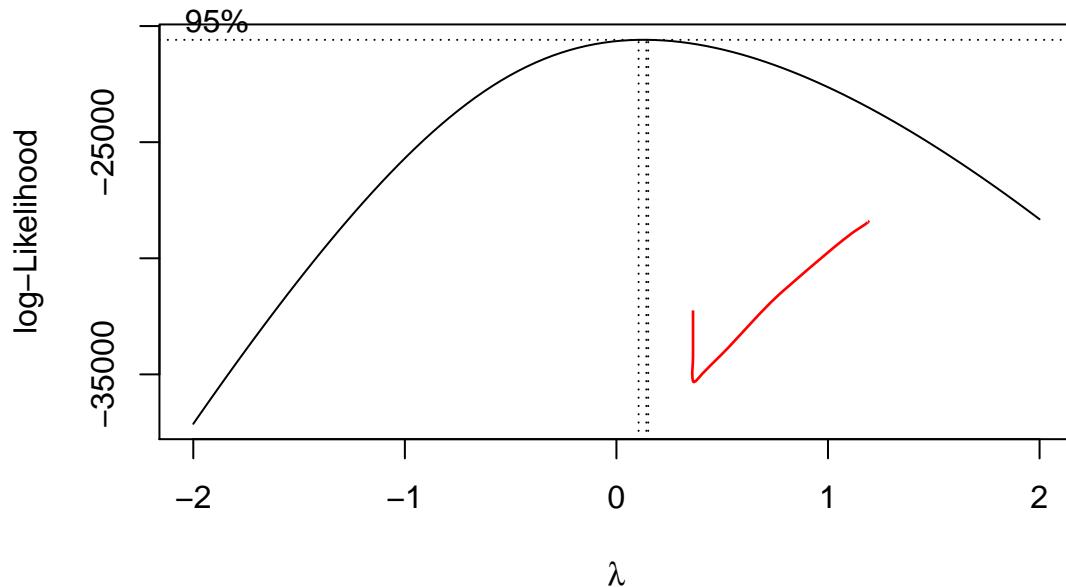
### Transformacio de la variable target numerica

A vegades una transformacio de la variable target numerica pot millorar el model. La comanda Bo-Cox ens mostra com el valor de lambda estimat es proper a 0, vol dir que hem d'elevar a 0 el target, com que això no es pot fer i amb les grafiques anteriors no em pogut indentificar cap patró que ens induis a elevar alguna de les variables, la transformacio estadistica del nostre target sera el logaritme, i la farem a partir del model m4 obtingut anteriorment (així donarem marge a possibles reduccions del mateix). A través de la grafica Normal Q-Q, podem observar com el nou model s'ajusta molt mes que l'anterior a una distribució normal i en podem identificar unes cues amb tendència inferior respecte a la línia de la normal. Per altra banda, també podem observar com en "Residuals vs Leverage" la majoria dels punts es concentra a la part esquerra de la grafica, ens indica que no tenim valors influents, cap dels valors es troba més enlla dels marges del leverage (les línies no es dibuixen).

```

#Box-Cox
boxcox(m4, data=df)

```



```

#TRANSFORMACIO LOGARITMICA Y(m4) -> logY
m8<-lm(log(duration)~campaign+emp.var.rate+nr.employed, data=df); Anova(m8)

## Anova Table (Type II tests)
##
## Response: log(duration)
##             Sum Sq   Df F value    Pr(>F)
## campaign      139.3     1 175.543 < 2.2e-16 ***
## emp.var.rate   21.8     1  27.484 1.650e-07 ***
## nr.employed    31.0     1  39.109 4.343e-10 ***
## Residuals   3953.3 4982
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#BIC
m10<-step(m8, k=log(nrow(df)))

## Start:  AIC=-1123.1
## log(duration) ~ campaign + emp.var.rate + nr.employed
##
##             Df Sum of Sq   RSS       AIC
## <none>              3953.3 -1123.10
## - emp.var.rate  1     21.809 3975.1 -1104.18
## - nr.employed   1     31.034 3984.3 -1092.62
## - campaign      1    139.296 4092.6  -958.95

vif(m10)

##      campaign emp.var.rate  nr.employed
## 1.024848     6.027929     6.013152

```

```

#emp.var.rate i nr.employed mostren molta colinearitat, ens quedem amb emp.var.rate
#per a ser una variable mes entenedora
m11<-lm(log(duration)~emp.var.rate+campaign, data=df)
vif(m11)

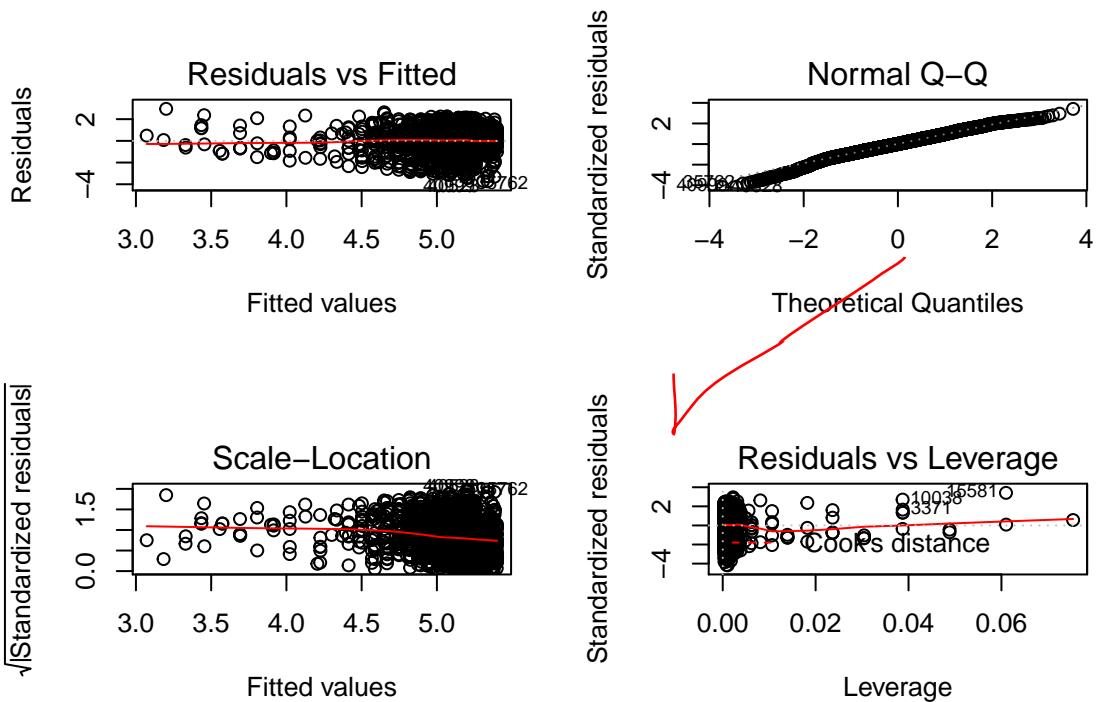
## emp.var.rate      campaign
##       1.024654    1.024654
summary(m11)

##
## Call:
## lm(formula = log(duration) ~ emp.var.rate + campaign, data = df)
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -3.6643 -0.5519  0.0113  0.5914  2.6497
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.327250  0.018169 293.205 <2e-16 ***
## emp.var.rate -0.008891  0.008087 -1.099   0.272
## campaign     -0.068651  0.005167 -13.286 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8942 on 4983 degrees of freedom
## Multiple R-squared:  0.03612, Adjusted R-squared:  0.03574
## F-statistic: 93.38 on 2 and 4983 DF, p-value: < 2.2e-16

#POLINOMIC REGRESSION
#com hi ha poques variables provem de transformar-les totes amb regressio polinomica
m20<-lm(log(duration)~poly(euribor3m, 2)+poly(campaign, 2), data=df)
summary(m20)
Anova(m20)
#veiem com amb ambdos termes quadratics(2) tenen un p-value <0.05.
#Es millor que el terme lineal(1) en el cas d l'euribor3m, podriem fer per tant aquesta
#transformacio quadratica.

par(mfrow=c(2,2))
plot(m20)

```



```
par(mfrow=c(1,1))
```

### Variables discretes explicatives pel target numeric

Mitjançant la comanda condes intentarem trobar variables discretes que estiguin relacionades amb la variable target numèrica duration. D'aquesta manera sabrem quines variables discretes podem utilitzar en el model predictiu per a que ens aportin informació. A partir del millor model m11 anterior de variables continues, hem d'obtenir un nou model afegint les variables discretes i factoritzades (que no estiguin ja en el model de forma numèrica). En el nostre cas agafem campaign i nr.employed com a variables continues, i afegim f.cons.conf.idx+f.cons.price.idx+month+f.euribor3m+poutcome com a variables discretes. Com que el condens anterior ens ha donat com a variables factor significatives algunes que ja tenim en el model com a continues, hem de triar una o altra versió. Per a saber si agafar una variable com a continua o factoritzada, hem de fer el següent i veiem com en ambdues variables obtenim que es millor usar la seva versió numèrica (no factoritzada).

```
condes(df[, c("duration", vars_dis)], 1, proba=0.01)
```

#a partir del millor model anterior (m11) amb variables continues afegim factors  
~~m60<-lm(log(duration)~campaign+nr.employed+f.cons.conf.idx+f.cons.price.idx+month~~  
~~+f.euribor3m+poutcome, data=df)~~  
summary(m60)

```
##  
## Call:  
## lm(formula = log(duration) ~ campaign + nr.employed + f.cons.conf.idx +  
##       f.cons.price.idx + month + f.euribor3m + poutcome, data = df)  
##  
## Residuals:  
##      Min        1Q    Median        3Q       Max  
## -3.7849 -0.5447  0.0007  0.5763  2.5903
```

```

##
## Coefficients:
##                               Estimate Std. Error t value
## (Intercept)           14.1263162  4.8821924  2.893
## campaign            -0.0685567  0.0051782 -13.240
## nr.employed         -0.0017450  0.0009734 -1.793
## f.cons.conf.idxf.cons.conf.idx-(-42.7,-41.8] -0.2215250  0.1385032 -1.599
## f.cons.conf.idxf.cons.conf.idx-(-41.8,-36.4] -0.1054326  0.0924315 -1.141
## f.cons.conf.idxf.cons.conf.idx-(-36.4,-26.9]  0.0841657  0.0844295  0.997
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]  0.1778539  0.0932539  1.907
## f.cons.price.idxf.cons.price.idx-(93.7,94]   0.2395880  0.1201263  1.994
## f.cons.price.idxf.cons.price.idx-(94,94.8]   0.1626503  0.0996614  1.632
## monthmonth-aug      -0.1611314  0.1410615 -1.142
## monthmonth-dec      -0.2665829  0.2010504 -1.326
## monthmonth-jul      -0.0104232  0.1309917 -0.080
## monthmonth-jun      0.0601159  0.1626995  0.369
## monthmonth-mar      -0.1319787  0.1316415 -1.003
## monthmonth-may      0.0062834  0.1106291  0.057
## monthmonth-nov      -0.0381862  0.1182949 -0.323
## monthmonth-oct      -0.2155702  0.1281603 -1.682
## monthmonth-sep      -0.1174620  0.1507313 -0.779
## f.euribor3mf.euribor3m-(1.33,4.86]       0.1413905  0.1056155  1.339
## f.euribor3mf.euribor3m-(4.86,4.96]       0.1738018  0.1201916  1.446
## f.euribor3mf.euribor3m-(4.96,5]        0.0598234  0.1297290  0.461
## poutcomepoutcome-nonexistent    0.0512680  0.0476555  1.076
## poutcomepoutcome-success     0.2738577  0.0862378  3.176
## Pr(>|t|)
## (Intercept)          0.00383 **
## campaign             < 2e-16 ***
## nr.employed          0.07308 .
## f.cons.conf.idxf.cons.conf.idx-(-42.7,-41.8] 0.10979
## f.cons.conf.idxf.cons.conf.idx-(-41.8,-36.4]  0.25407
## f.cons.conf.idxf.cons.conf.idx-(-36.4,-26.9]  0.31887
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]  0.05655 .
## f.cons.price.idxf.cons.price.idx-(93.7,94]   0.04616 *
## f.cons.price.idxf.cons.price.idx-(94,94.8]   0.10274
## monthmonth-aug       0.25339
## monthmonth-dec      0.18492
## monthmonth-jul      0.93658
## monthmonth-jun      0.71178
## monthmonth-mar      0.31612
## monthmonth-may      0.95471
## monthmonth-nov      0.74686
## monthmonth-oct      0.09262 .
## monthmonth-sep      0.43585
## f.euribor3mf.euribor3m-(1.33,4.86]       0.18072
## f.euribor3mf.euribor3m-(4.86,4.96]       0.14823
## f.euribor3mf.euribor3m-(4.96,5]        0.64472
## poutcomepoutcome-nonexistent    0.28207
## poutcomepoutcome-success     0.00150 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8878 on 4963 degrees of freedom

```

```

## Multiple R-squared:  0.05375,    Adjusted R-squared:  0.04955
## F-statistic: 12.81 on 22 and 4963 DF,  p-value: < 2.2e-16
Anova(m60)

## Anova Table (Type II tests)
##
## Response: log(duration)
##             Sum Sq   Df  F value    Pr(>F)
## campaign        138.1    1 175.2846 < 2.2e-16 ***
## nr.employed     2.5    1   3.2138  0.073079 .
## f.cons.conf.idx 4.2    3   1.7649  0.151593
## f.cons.price.idx 5.7    3   2.4208  0.064119 .
## month          5.8    9   0.8134  0.603791
## f.euribor3m      4.9    3   2.0732  0.101547
## poutcome        8.0    2   5.0541  0.006416 **
## Residuals     3911.5 4963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#per campaign, mateix model pero amb f.campaign en lloc de campaign:
maux<-lm(log(duration)~f.campaign+nr.employed+f.cons.conf.idx+f.cons.price.idx+month
         +f.euribor3m+poutcome, data=df)

#com que m60 i maux no son anidats, no els podem comparar amb un test de Fisher --> BIC
BIC(m60, maux)
#choose option with minimum BIC -> better
#      df      BIC
# m60 24 13143.82 --> millor model amb campaign com a numerica
# maux 25 13199.11

maux<-lm(log(duration)~campaign+f.nr.employed+f.cons.conf.idx+f.cons.price.idx+month
         +f.euribor3m+poutcome, data=df)
BIC(m60, maux)
#      df      BIC
# m60 24 13143.82 --> millor model amb nr.employed com a numerica
# maux 25 13151.61

#si haguessim hagut de fer el mateix amb pdays, ojo!!!! pq la continua ha estat majoritariament
#imputada, per tant en aquest cas, tot i el test, hauriem d'agafar la factoritzada!

```

Mirem si podem simplificar el model eliminant variables poc significatives mitjancant la comanda step i veiem com ens podem quedar amb les variables numeriques nr.employed i campaign, i la variable factor f.cons.price.idx. Si ho fem mitjancant la comanda Anova, veiem com ens surt el mateix resultat pero agafant també la variable discreta poutcome. Així doncs també li farem cas i aquest model sera el m62, que explica el 5% ( $R^2=0.04959$ ) de la variabilitat.

```

m61<-step(m60, k=log(nrow(df)))

m62<-lm(log(duration)~campaign+nr.employed+f.cons.price.idx+poutcome, data=df); summary(m62)

```

## Interaccions

Partint del model anterior, li afegim interaccions 2 a 2 entre totes les seves variables, simplifiquem i veiem com hi ha dues interaccions significatives: campaign:nr.employed i campaign:f.cons.price.idx. En el nostre model nomes podem tenir en compte interaccions entre dos factors o entre un factor i una variable numerica,

així que amb els tests d'Anova mirarem manualment quines interaccions ens quedem, el que ens porta a un model m73, que explica el 5.5% ( $R^2=0.05534$ ) de la variabilitat de l'output del logaritme de *duration*.

```
#interacció entre 2 variables:
m70<-lm(log(duration)~(campaign+nr.employed+f.cons.price.idx+poutcome)^2, data=df)
summary(m70)
#coef(m70)
invisible(
  m71<-step(m70, k=log(nrow(df)))
)#el criteri Anova(Fisher) reafirma el step(BIC) en aquest cas!
# log(duration) ~ campaign+nr.employed+f.cons.price.idx+campaign:nr.employed+campaign:f.cons.price.idx
#
#                               Df Sum of Sq    RSS      AIC
# <none>                      3907.1 -1130.7
# - campaign:nr.employed      1     13.192 3920.3 -1122.4
# - campaign:f.cons.price.idx 3     30.057 3937.1 -1118.0
invisible(
  Anova(m71)
)
anova(m71, m70) #Pr(>F) = 0.03967 * --> els models no són equivalents

Anova(m70)
#                                     Pr(>F)
# campaign                                < 2.2e-16 ***
# nr.employed                            0.012590 *
# f.cons.price.idx                       1.162e-08 ***
# poutcome                                0.003345 **
# campaign:nr.employed                  0.001721 ** --> entre dos numériques
# campaign:f.cons.price.idx              4.635e-07 *** --> entre numérica i factor --> AGAFEM
# campaign:poutcome                      0.873389   --> entre numérica i factor
# nr.employed:f.cons.price.idx           0.058763 .  --> entre numérica i factor
# nr.employed:poutcome                   0.309191   --> entre numérica i factor
# f.cons.price.idx:poutcome             0.121019   --> entre factors --> AGAFEM aquest per l'entrega

m73<-lm(log(duration)~campaign+nr.employed+f.cons.price.idx+
          poutcome+campaign:f.cons.price.idx+f.cons.price.idx:poutcome, data=df)
anova(m73, m70) #p-value 0.003286 ** -> models no són equivalents -> H0 rejected -> m73 es millor

## Analysis of Variance Table
##
## Model 1: log(duration) ~ campaign + nr.employed + f.cons.price.idx + poutcome +
##           campaign:f.cons.price.idx + f.cons.price.idx:poutcome
## Model 2: log(duration) ~ (campaign + nr.employed + f.cons.price.idx +
##           poutcome)^2
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1    4969 3904.9
## 2    4961 3886.8  8     18.099 2.8877 0.003286 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
summary(m73)

##
## Call:
## lm(formula = log(duration) ~ campaign + nr.employed + f.cons.price.idx +
##       poutcome + campaign:f.cons.price.idx + f.cons.price.idx:poutcome,
```

```

##      data = df)
##
## Residuals:
##      Min    1Q Median    3Q   Max
## -3.8187 -0.5570  0.0047  0.5735  2.6169
##
## Coefficients:
##                               Estimate
## (Intercept)                9.4669014
## campaign            -0.1222066
## nr.employed        -0.0008058
## f.cons.price.idxf.cons.price.idx-(93.1,93.7] -0.1204673
## f.cons.price.idxf.cons.price.idx-(93.7,94]  0.5608898
## f.cons.price.idxf.cons.price.idx-(94,94.8]  0.0523764
## poutcomepoutcome-nonexistent  0.0875966
## poutcomepoutcome-success   0.1580875
## campaign:f.cons.price.idxf.cons.price.idx-(93.1,93.7]  0.0414029
## campaign:f.cons.price.idxf.cons.price.idx-(93.7,94]  0.0737345
## campaign:f.cons.price.idxf.cons.price.idx-(94,94.8]  0.0531081
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-nonexistent  0.0103004
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-nonexistent -0.5762804
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-nonexistent -0.1861021
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-success  0.2510572
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-success -0.3243955
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-success  0.2715234
##
##                               Std. Error
## (Intercept)                1.6422347
## campaign            0.0145243
## nr.employed        0.0003230
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]  0.1100012
## f.cons.price.idxf.cons.price.idx-(93.7,94]  0.3416381
## f.cons.price.idxf.cons.price.idx-(94,94.8]  0.1619483
## poutcomepoutcome-nonexistent  0.0561413
## poutcomepoutcome-success   0.1096234
## campaign:f.cons.price.idxf.cons.price.idx-(93.1,93.7]  0.0184993
## campaign:f.cons.price.idxf.cons.price.idx-(93.7,94]  0.0163059
## campaign:f.cons.price.idxf.cons.price.idx-(94,94.8]  0.0184387
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-nonexistent  0.1106143
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-nonexistent  0.3461481
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-nonexistent  0.1753390
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-success  0.2305015
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-success  0.4506457
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-success  0.2301328
##
##                               t value
## (Intercept)                5.765
## campaign            -8.414
## nr.employed        -2.495
## f.cons.price.idxf.cons.price.idx-(93.1,93.7] -1.095
## f.cons.price.idxf.cons.price.idx-(93.7,94]  1.642
## f.cons.price.idxf.cons.price.idx-(94,94.8]  0.323
## poutcomepoutcome-nonexistent  1.560
## poutcomepoutcome-success   1.442
## campaign:f.cons.price.idxf.cons.price.idx-(93.1,93.7]  2.238
## campaign:f.cons.price.idxf.cons.price.idx-(93.7,94]  4.522

```

```

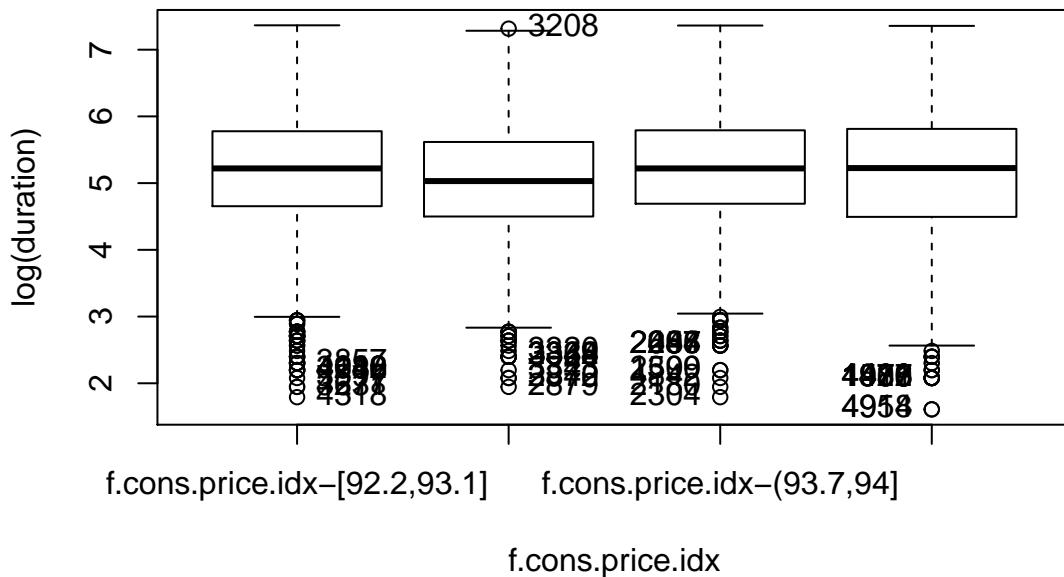
## campaign:f.cons.price.idxf.cons.price.idx-(94,94.8]                      2.880
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-nonexistent   0.093
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-nonexistent     -1.665
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-nonexistent     -1.061
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-success       1.089
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-success       -0.720
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-success       1.180
##
## Pr(>|t|)
## (Intercept)                                         8.68e-09
## campaign                                              < 2e-16
## nr.employed                                           0.01263
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]          0.27351
## f.cons.price.idxf.cons.price.idx-(93.7,94]           0.10070
## f.cons.price.idxf.cons.price.idx-(94,94.8]           0.74640
## poutcomepoutcome-nonexistent                         0.11876
## poutcomepoutcome-success                           0.14934
## campaign:f.cons.price.idxf.cons.price.idx-(93.1,93.7] 0.02526
## campaign:f.cons.price.idxf.cons.price.idx-(93.7,94] 6.27e-06
## campaign:f.cons.price.idxf.cons.price.idx-(94,94.8] 0.00399
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-nonexistent 0.92581
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-nonexistent 0.09601
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-nonexistent 0.28857
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-success    0.27613
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-success    0.47165
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-success    0.23811
##
## (Intercept)                                         ***
## campaign                                              ***
## nr.employed                                           *
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]          *
## f.cons.price.idxf.cons.price.idx-(93.7,94]           ***
## f.cons.price.idxf.cons.price.idx-(94,94.8]           **
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-nonexistent .
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-nonexistent .
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-nonexistent .
## f.cons.price.idxf.cons.price.idx-(93.1,93.7]:poutcomepoutcome-success
## f.cons.price.idxf.cons.price.idx-(93.7,94]:poutcomepoutcome-success
## f.cons.price.idxf.cons.price.idx-(94,94.8]:poutcomepoutcome-success
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8865 on 4969 degrees of freedom
## Multiple R-squared:  0.05534,  Adjusted R-squared:  0.0523
## F-statistic: 18.19 on 16 and 4969 DF,  p-value: < 2.2e-16

```

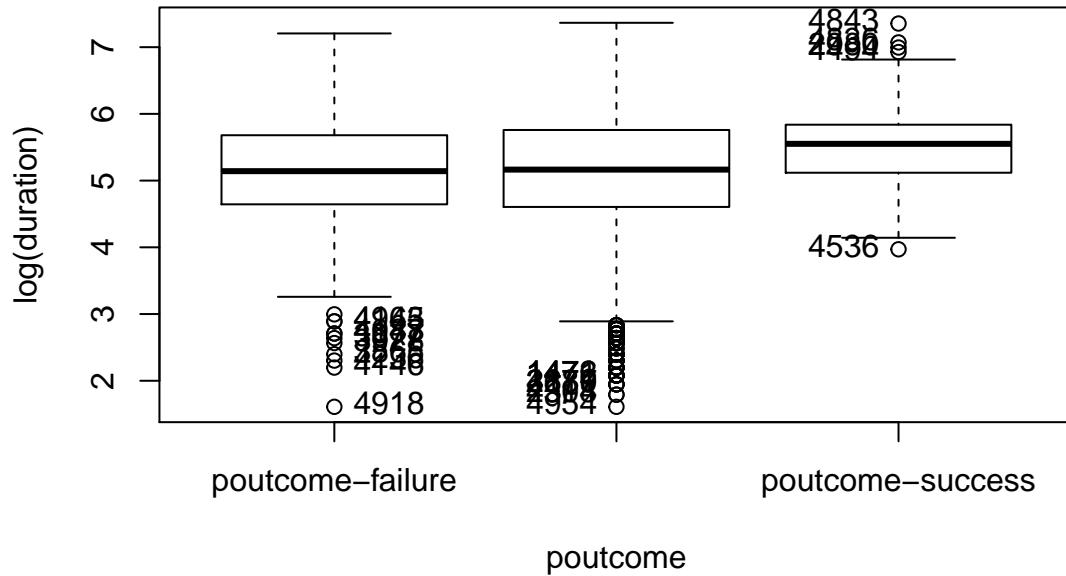
### Interaction between a couple of factors in our model m73

El model escollit m73 considera una interacció entre dos factors *f.cons.price.idx:poutcome*.

```
scatterplot(log(duration)~f.cons.price.idx, data=df)
```



```
## [1] "4318" "3671" "4237" "3597" "3682" "4088" "4146" "4210" "4280" "3857"  
## [11] "2879" "2870" "3342" "3343" "2868" "3329" "3332" "3344" "3345" "3320"  
## [21] "3208" "2304" "2180" "4842" "1599" "2300" "487" "666" "2048" "2254"  
## [31] "2297" "4918" "4954" "1471" "1472" "1476" "1486" "1489" "1507" "4916"  
## [41] "4928"  
scatterplot(log(duration)~poutcome, data=df)
```

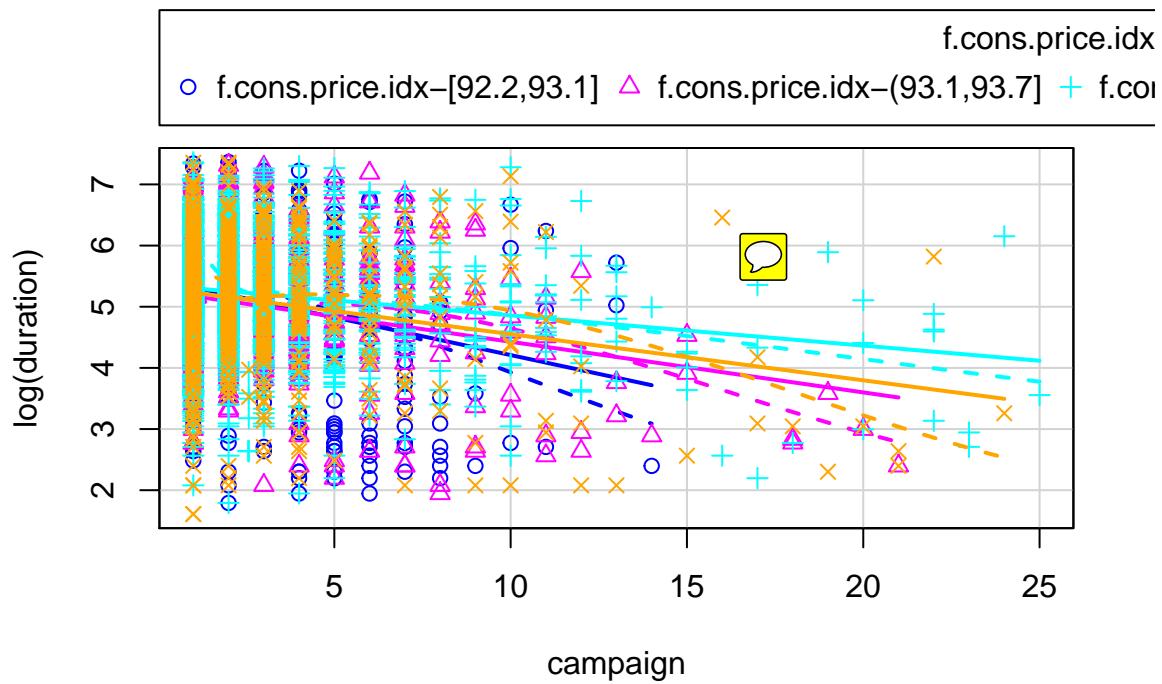


```
## [1] "4918" "4146" "4236" "3865" "3928" "3877" "4083" "4147" "4143" "4965"  
## [11] "4954" "2304" "4318" "2180" "2879" "3671" "4237" "1471" "1472" "1476"  
## [21] "4536" "2980" "4494" "4836" "4843"
```

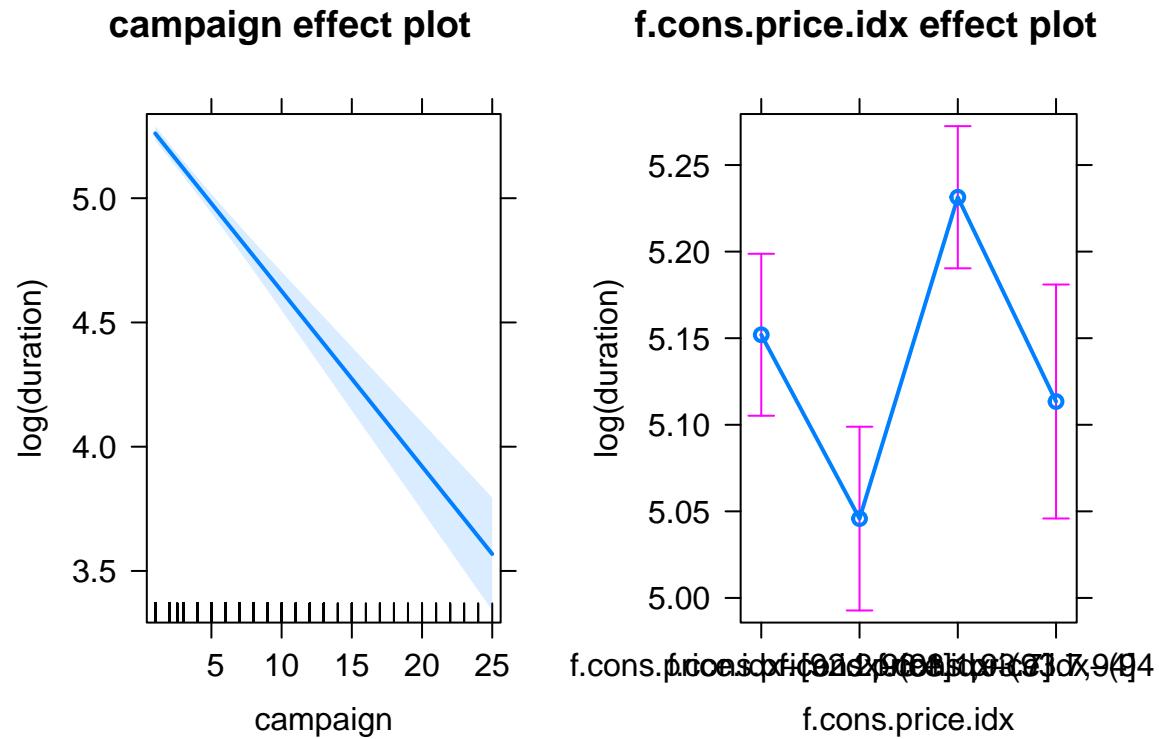
## Interaction between a factor and a covariate in our model m73

El model escollit m73 també considera una interacció entre un factor i una variable numèrica *campaign:f.cons.price*.

```
#model petit sense interaccions
m85<-lm(log(duration)~campaign+f.cons.price.idx, data=df)
scatterplot(log(duration)~campaign|f.cons.price.idx, data=df) #Suport visual
```



```
plot(allEffects(m85)) #effects library
```



```
#model gran amb interaccions: 3 parametres-> campaign, f.cons.price.idx, campaign:f.cons.price.idx
m855<-lm(log(duration)~campaign*f.cons.price.idx, data=df)
#are interactions significant?
```

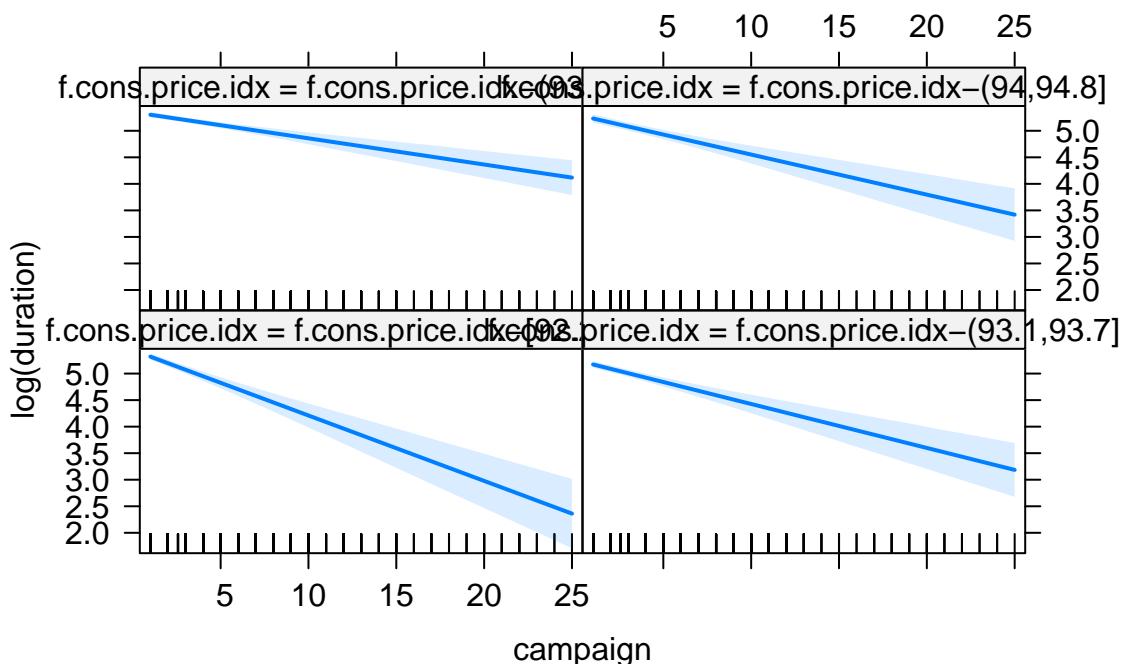
```

anova(m855, m855) #Pr(>F) 5.152e-05 *** --> H0 rejected --> m855 amb la interaccio es millor

## Analysis of Variance Table
##
## Model 1: log(duration) ~ campaign + f.cons.price.idx
## Model 2: log(duration) ~ campaign * f.cons.price.idx
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1    4981 3960.6
## 2    4978 3942.8  3     17.853 7.5136 5.152e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
plot(allEffects(m855))

```

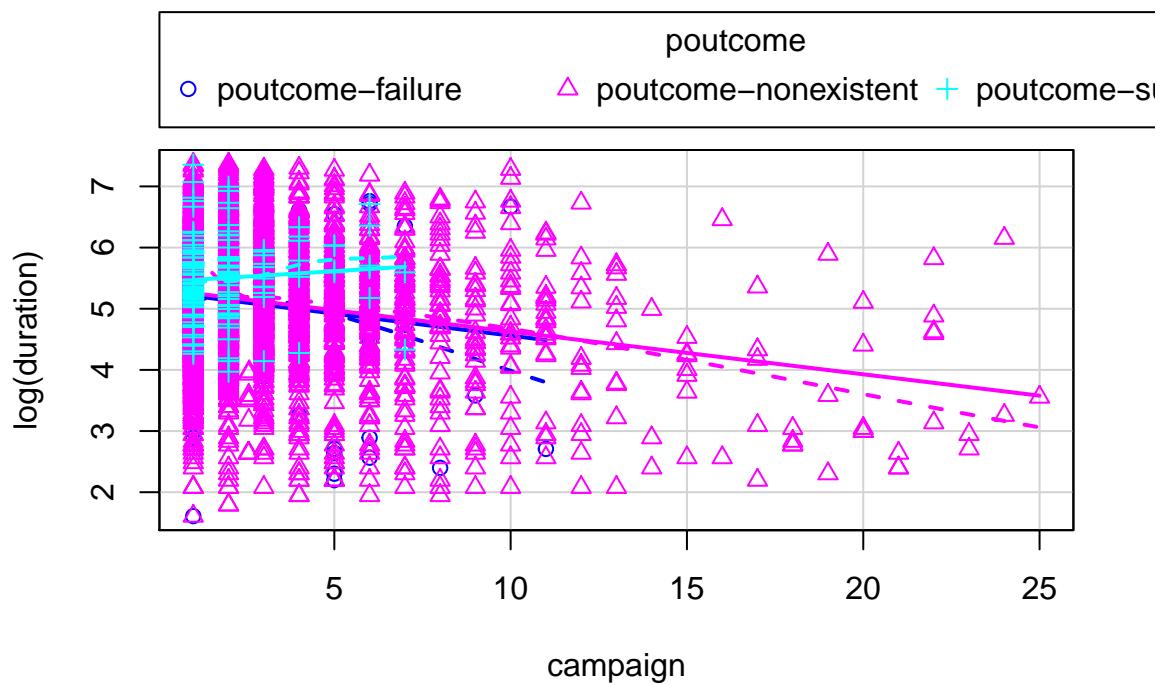
### campaign\*f.cons.price.idx effect plot



```

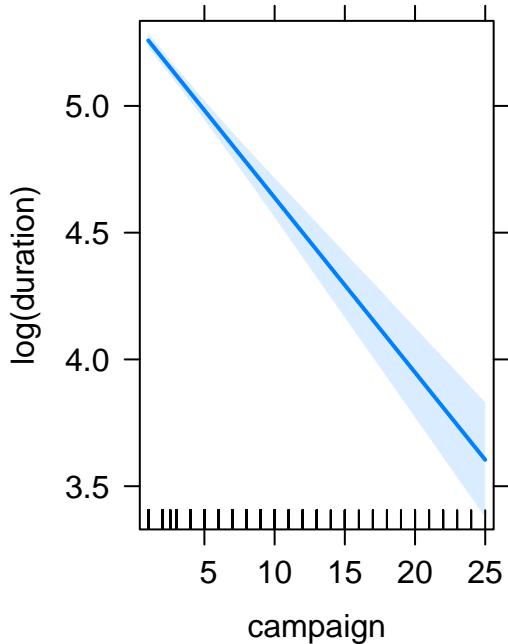
#campaign:poutcome --> segons el test anterior, la interaccio no ha sortit gaire
#significativa i no s'inclou en el model definitiu, pero s'interpreta millor visualment.
m86<-lm(log(duration)~campaign+poutcome, data=df)
scatterplot(log(duration)~campaign|poutcome, data=df)

```

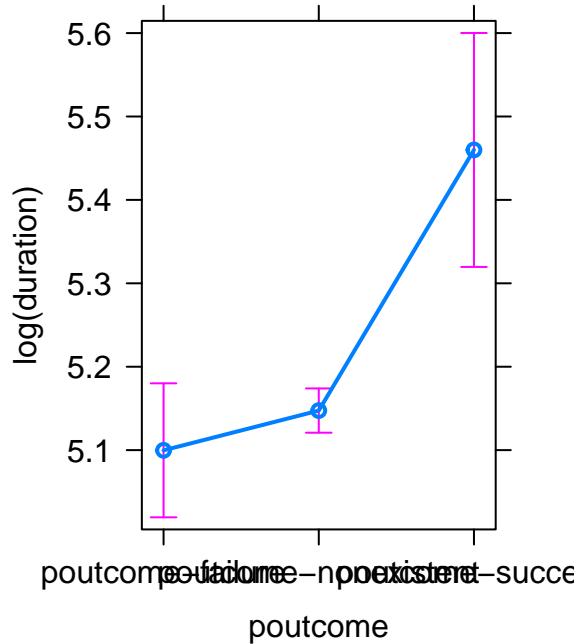


```
plot(allEffects(m86))
```

**campaign effect plot**



**poutcome effect plot**



```
m866<-lm(log(duration)~campaign*poutcome, data=df)
```

```
anova(m86, m866) #Pr(>F) 0.1435--> H0 accepted --> els models son iguals, per tant no cal
```

```

## Analysis of Variance Table
##
## Model 1: log(duration) ~ campaign + poutcome
## Model 2: log(duration) ~ campaign * poutcome
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1     4982 3969.1
## 2     4980 3966.1  2      3.0931 1.9419 0.1435
#el model gran amb interaccions

```

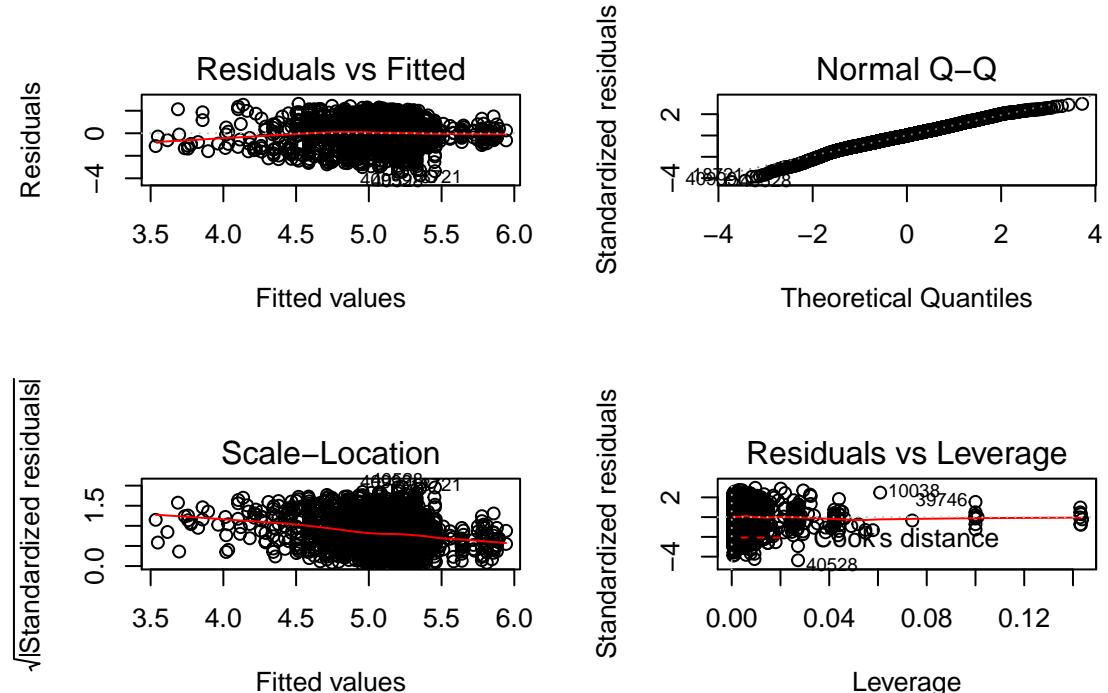
## Diagnostics del model definitiu m73

Residual vs Fitted En aquesta grafica podem veure els valors dels residus en l'eix Y i els valors de duracio en l'eix X. El nuvol de punts s'hauria de trobar encaixat dins un marge delimitat per dues línies paral·leles reconeixibles a la vista, fet que no podem veure en el nostre grafic i ens indica la incorrectesa del model. A més, també podem apreciar com els residus no es troben distribuïts de manera aleatoria, sino que es concentren en la part central de la grafica de manera molt significativa. De totes maneres, ha millorat respecte els models inicials. Normal Q-Q Aquesta grafica ens mostra els residus envers als seus valors esperats en el suposat d'una distribució normal. Podem dir que els residus no segueixen ben una distribució normal donat que s'allunyen significativament de la recta de punts en els seus extrems. Scale-Location L'objectiu d'aquesta grafica es visualitzar la primera grafica però amb valors de residus estandarditzats per tal de poder demostrar-ne homoscedasticitat. Es pot apreciar que no n'hi ha, es a dir, que la variància dels errors en les diferents observacions no és constant. Residual vs Leverage Aquesta grafica ens ajuda a identificar "influent data" en els nostres residus basant-se en la distància de Cook. En la nostra grafica no s'hi aprecia cap observació realment influent en el model amb gran distància de Cook, però si que s'observen forces observacions amb gran residu. Tot i així, no podem extreure grans conclusions mes enllà de dir que el model no es gaire bo.

```

par(mfrow=c(2,2))
plot(m73)

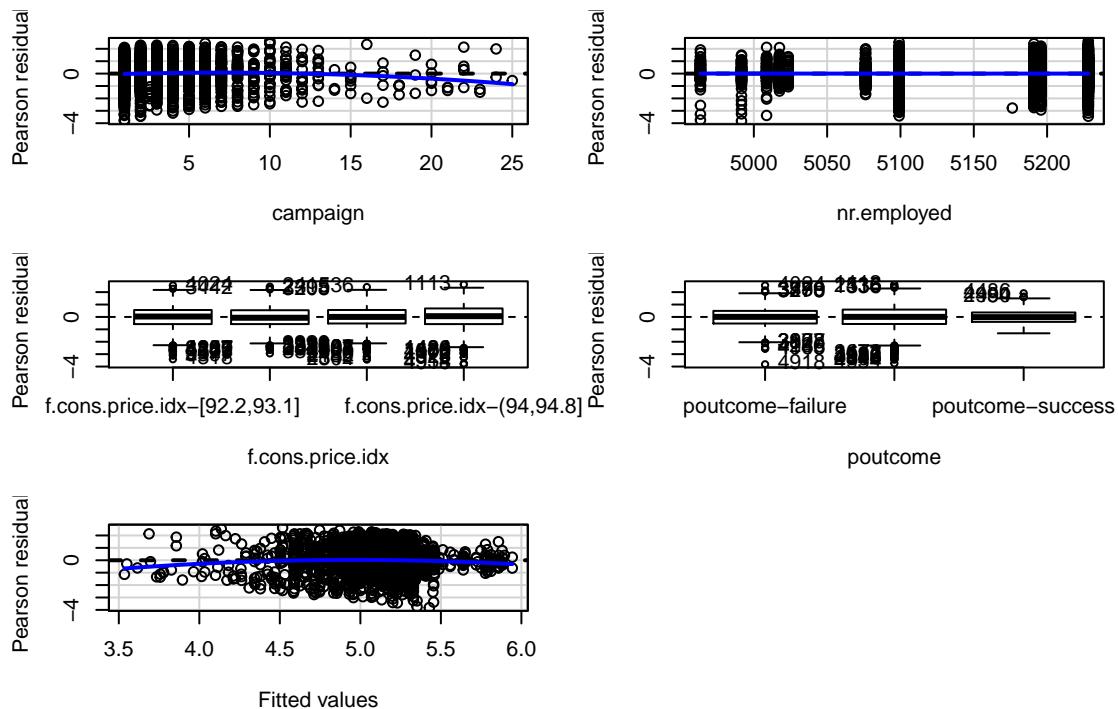
```



```
par(mfrow=c(1,1))
```

ResidualPlots Aquesta comanda ens mostra els residus de Pearson per a cada variable explicativa del model. En cap d'elles podem indentificar un patro que ens suggereixi un canvi o transformacio de la variable en el nostre model. marginalModelPlots Mitjançant un plot dels valors de log(duration) de les nostres dades, als qual hi afegeix un smoother blau, podem veure com el nostre model, que esta representat amb un de vermell, no esta predint gaire acuradament (tret del cas nr.employed). Això es veu perque no es sobreposen ambdues línies de suavitzacio, aixi doncs, el model obtingut no es gaire apropiat, sobretot en valors de la part esquerra de la grafica. influencePlot Aquesta comanda crea un grafic de bombolles amb el valor dels residus "Studentized", on el diametre de les bombolles mante una relacio proporcional a la seva distancia de Cook de l'individu (es veu tambe en el boxplot). Amb l'ajuda de la taula que s'ens mostra podem identificar com les mostres 10038 i 40528 son les mes influents, les que tenen una distancia de Cook mes elevada, es a dir, que treient-les el model canviaria més. Les observacions 39584 i 39699 son les que tenen mes leverage (hat value), es a dir, que cauen mes fora de la majoria de valors predicts del model. Tot i que les observacions amb mes leverage poden tenir mes habilitat per a afectar al model, en aquest cas no ho fan, ja que han han caigut en el patro del model i tenen una distancia de Cook baixa, es a dir, no son influents. Per ultim, també tenim algunes observacions amb un residu força elevat, com son les 10038, 40528 i 40999.

```
residualPlots(m73)
```

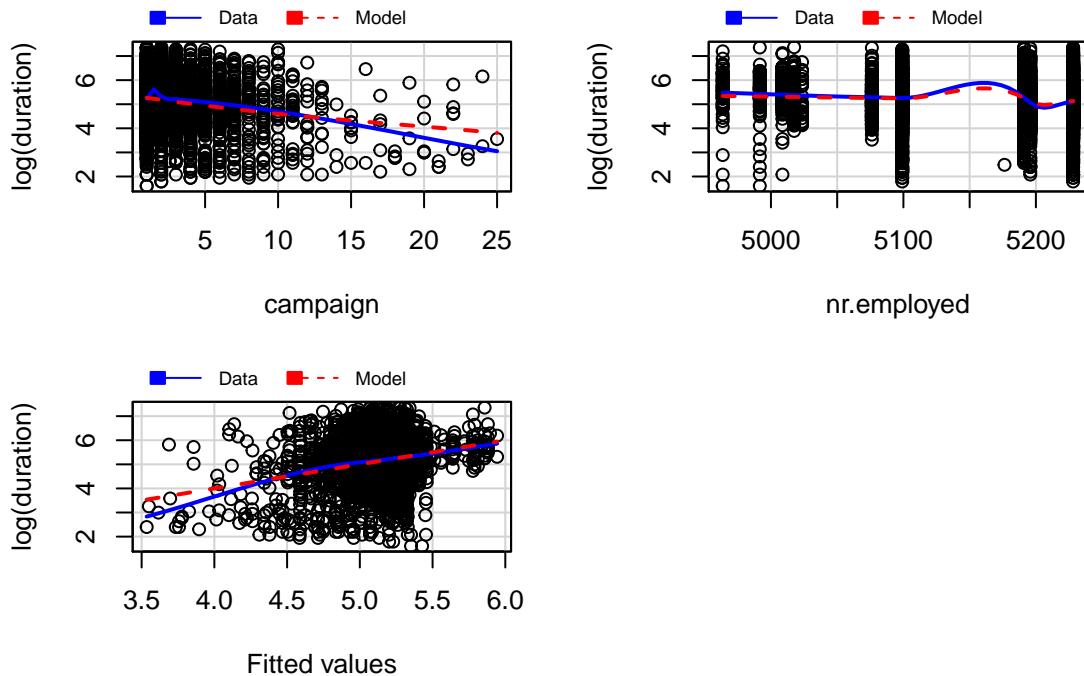


```
##           Test stat Pr(>|Test stat|)  
## campaign      -4.2245    2.438e-05 ***  
## nr.employed   0.1803     0.8569  
## f.cons.price.idx  
## poutcome  
## Tukey test     -5.1074    3.266e-07 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

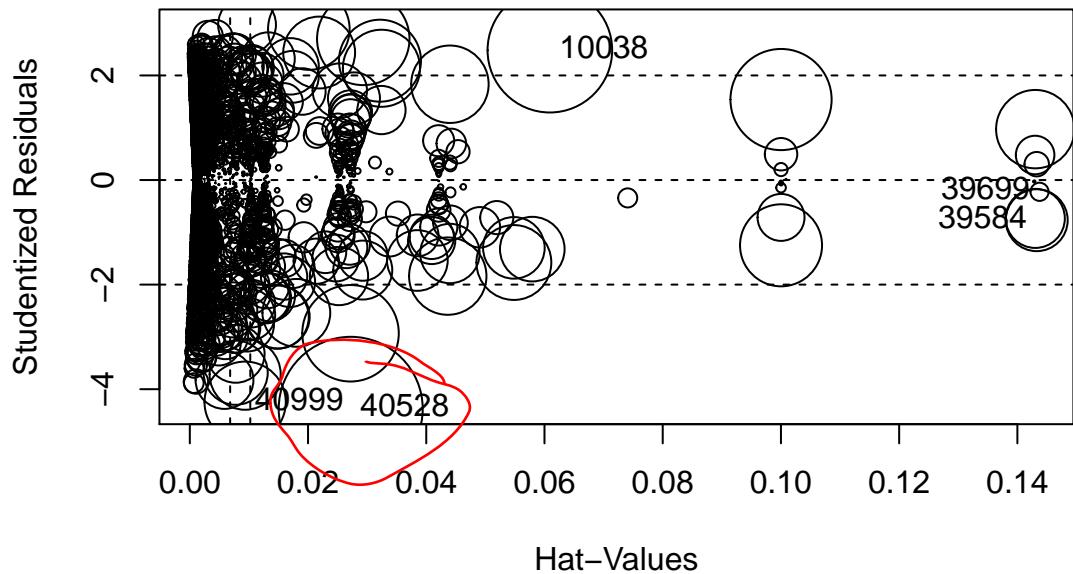
```
marginalModelPlots(m73)
```

```
## Warning in mmpls(...): Interactions and/or factors skipped
```

### Marginal Model Plots



```
influencePlot(m73)
```

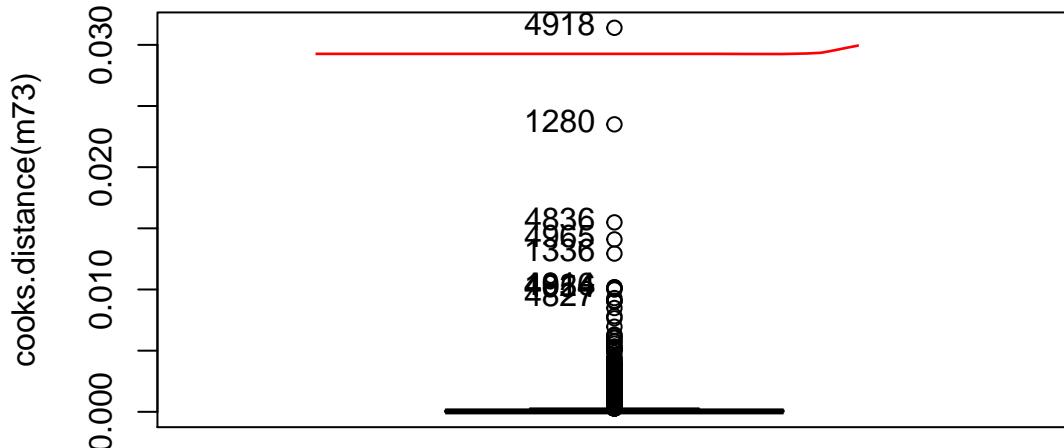


```

##          StudRes      Hat      CookD
## 10038    2.4832006 0.060909838 0.0235019298
## 39584   -0.7680869 0.143271206 0.0058039438
## 39699   -0.2249883 0.143752609 0.0005000004
## 40528   -4.3754701 0.027217651 0.0313944221
## 40999   -4.2491549 0.009357579 0.0099980359

```

```
Boxplot(cooks.distance(m73))
```



```
## [1] 4918 1280 4836 4965 1336 4816 1914 4024 4954 4827
```

## Binary Regression Models - target binari d'acceptació del producte financer “y”

Per tal d'elaborar un model lineal que predigui el valor de la variable binaria target  $y$ , primer hem de decidir quines són les variables (columnes) que utilitzarem en la seva construcció. En altres paraules, trobar quines variables ens aporten informació i precisió al model predictiu, però sense sobreparametritzar-lo.

### Work and test samples division

Dividim la nostra mostra en dues submostres: el 75% de la mostra inicial serà per a treballar amb les dades (dataframe work -  $dfw$ ), i el 25% restant serà per a testejar-les (dataframe test -  $dft$ ).

```
set.seed(69)
sam<-sample(1:nrow(df), 0.75*nrow(df)) #random sample without replacement
dfw<-df[sam,] #work75%
dft<-df[-sam,] #test25%
```

### Variables numèriques explicatives pel target binari

A partir de la informació del catdes, que ens diu les variables numèriques més explicatives, generem un model inicial del tipus binomial, i a partir d'aquí el simplificarem per a que no quedi sobreparametritzat (model gm2).

```
#numeric variables
catdes( dfw[,c("y", vars_con)], 1)
```

```

#glm amb les variables continues significatives
gm1<-glm(y~nr.employed+euribor3m+emp.var.rate+pdays+previous+cons.price.idx+
           cons.conf.idx+campaign, family=binomial, data=dfw); summary(gm1)

## 
## Call:
## glm(formula = y ~ nr.employed + euribor3m + emp.var.rate + pdays +
##       previous + cons.price.idx + cons.conf.idx + campaign, family = binomial,
##       data = dfw)
## 
## Deviance Residuals:
##    Min      1Q   Median      3Q      Max
## -1.6348 -0.3897 -0.3574 -0.2725  2.7048
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -86.895389  41.451889 -2.096 0.036056 *
## nr.employed -0.000228  0.003897 -0.058 0.953354
## euribor3m   -0.343097  0.238396 -1.439 0.150096
## emp.var.rate -0.413371  0.192678 -2.145 0.031921 *
## pdays        -0.106215  0.018245 -5.822 5.82e-09 ***
## previous     -0.362876  0.124945 -2.904 0.003681 **
## cons.price.idx  0.974082  0.265630  3.667 0.000245 ***
## cons.conf.idx   0.051647  0.016582  3.115 0.001842 **
## campaign      -0.044702  0.030582 -1.462 0.143822
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
## Null deviance: 2619.2 on 3738 degrees of freedom
## Residual deviance: 2196.2 on 3730 degrees of freedom
## AIC: 2214.2
## 
## Number of Fisher Scoring iterations: 5
Anova(gm1)

```

```

## Analysis of Deviance Table (Type II tests)
## 
## Response: y
##             LR Chisq Df Pr(>Chisq)
## nr.employed      0.003  1  0.9533446
## euribor3m        2.088  1  0.1484697
## emp.var.rate      4.572  1  0.0324990 *
## pdays            35.065  1  3.189e-09 ***
## previous          8.879  1  0.0028853 **
## cons.price.idx   12.730  1  0.0003599 ***
## cons.conf.idx     9.757  1  0.0017864 **
## campaign          2.323  1  0.1274845
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
vif(gm1) #check colinear variables

```

```

##      nr.employed      euribor3m    emp.var.rate      pdays      previous
##      32.326246      60.227711     33.905107      1.750267     1.944268
## cons.price.idx  cons.conf.idx      campaign
##      10.091478      2.980404     1.025755

#our strategy: remove 2 colinear variables and campaign
gm2<-glm(y~emp.var.rate+pdays+previous+cons.price.idx+cons.conf.idx,
           family=binomial, data=dfw); summary(gm2)

##
## Call:
## glm(formula = y ~ emp.var.rate + pdays + previous + cons.price.idx +
##       cons.conf.idx, family = binomial, data = dfw)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q        Max
## -1.7093  -0.3918  -0.3725  -0.2786   2.6749
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.127e+02  1.312e+01 -8.590 < 2e-16 ***
## emp.var.rate -8.590e-01  5.815e-02 -14.771 < 2e-16 ***
## pdays        -1.045e-01  1.820e-02 -5.744 9.26e-09 ***
## previous     -3.158e-01  1.227e-01 -2.573  0.0101 *
## cons.price.idx 1.218e+00  1.407e-01  8.653 < 2e-16 ***
## cons.conf.idx  3.905e-02  9.953e-03  3.924 8.72e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2619.2 on 3738 degrees of freedom
## Residual deviance: 2204.1 on 3733 degrees of freedom
## AIC: 2216.1
##
## Number of Fisher Scoring iterations: 5

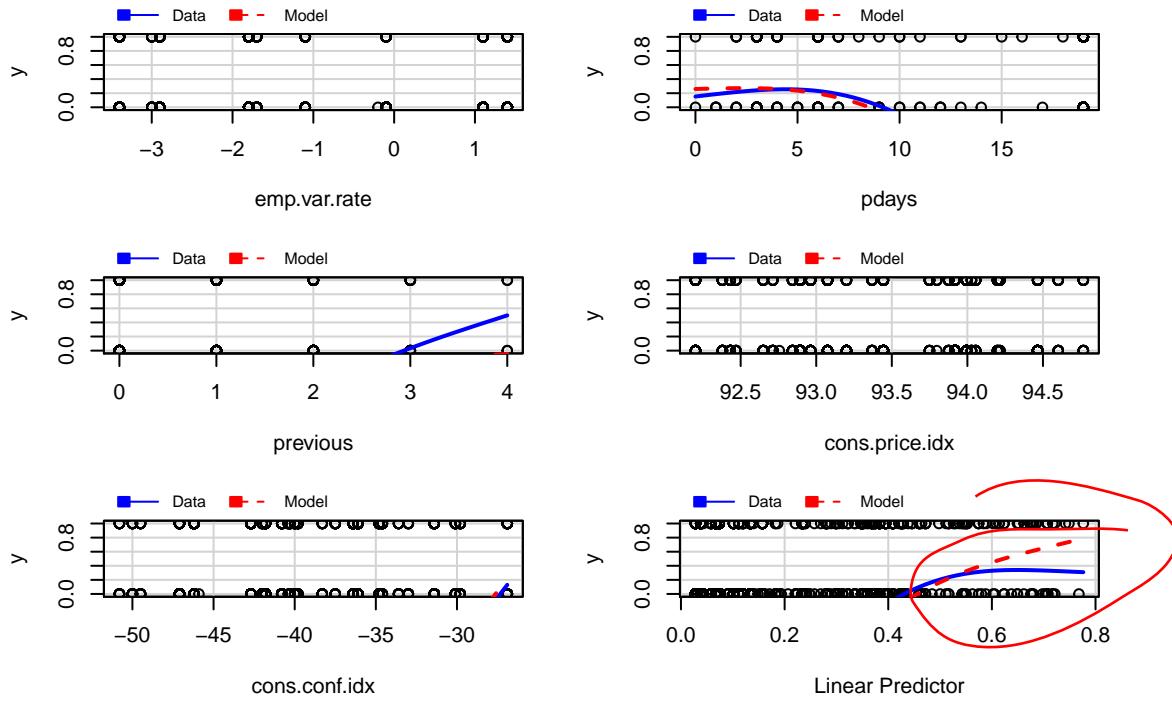
vif(gm2)

##      emp.var.rate      pdays      previous cons.price.idx  cons.conf.idx
##      3.120624      1.738325     1.878031      2.877660     1.060744

marginalModelPlots(gm2) #some missfit data vs model

```

## Marginal Model Plots



### Transforming variables

No veiem cap patró en els marginal plots que ens pugui ajudar a l'hora de seleccionar una transformació de variable en el model. Alguns que hem provat no milloraven el model.

```
#gm3<-glm(y~emp.var.rate+pdays+poly(previous, 2)+cons.price.idx+cons.conf.idx,
#           family=binomial, data=dfw); summary(gm3)
#Anova(gm3)
#marginalModelPlots(gm3)
```

### Variables discretes explicatives pel target binari

A partir del millor model en aquest punt (gm2), comprovem per a cada variable numèrica si es millor la seva utilització com a factor o com a numèrica. Ens quedem doncs amb un model gm2f, el qual té les variables pdays i f.cons.conf.idx com a factors i la resta com a numèriques: y~emp.var.rate+f.pdays+previous+cons.price.idx+cons.conf.idx.

```
#amb pdays, ojo!!!! pq la continua ha estat majoritariament imputada, per tant en aquest cas,
#no fem cas del test, agafem la factoritzada!
gm2<-glm(y~emp.var.rate+f.pdays+previous+cons.price.idx+cons.conf.idx, family=binomial, data=dfw)

#f.emp.var.rate or emp.var.rate?
gm2f<-glm(y~f.emp.var.rate+f.pdays+previous+cons.price.idx+cons.conf.idx, family=binomial, data=dfw)
BIC(gm2, gm2f)
#f.emp.var.rate dona pitjor (mes baix) BIC, ens quedem amb la variable numèrica

gm2f<-glm(y~emp.var.rate+f.pdays+f.previous+cons.price.idx+cons.conf.idx, family=binomial, data=dfw)
BIC(gm2, gm2f)
#previous com a numèrica
```

```

gm2f<-glm(y~emp.var.rate+f.pdays+previous+f.cons.price.idx+cons.conf.idx, family=binomial, data=dfw)
BIC(gm2, gm2f)
#cons.price.idx com a numerica

gm2f<-glm(y~emp.var.rate+f.pdays+previous+cons.price.idx+f.cons.conf.idx, family=binomial, data=dfw)
BIC(gm2, gm2f)
#f.cons.conf.idx com a factor

```

Ara afegirem les variables discretes explicatives que siguin significatives, ho farem a partir de les que ens indiqui el catdes, sempre sense repetir una variable si ja esta representada com a numerica. Mirem les colinearitats i eliminem les variables que en tinguin, per a continuaciO simplifcar el model amb la comanda step. Veiem com finalment el model gm44 es un model sense colinearitats, on tots els efectes de les variables explicatives son significatius.

```

#discrete variables
catdes( dfw[,c("y", vars_dis)], 1)

#assumim gm2f el millor model en aquest punt i li afegim les significatives
#(entre month i f.season triem la primera):
gm4<-glm(y~emp.var.rate+f.pdays+previous+cons.price.idx+f.cons.conf.idx+f.nr.employed+poutcome
          +f.euribor3m+contact+default+f.age+education+month+marital+f.campaign, family=binomial, data=dfw)
summary(gm4)

## 
## Call:
## glm(formula = y ~ emp.var.rate + f.pdays + previous + cons.price.idx +
##       f.cons.conf.idx + f.nr.employed + poutcome + f.euribor3m +
##       contact + default + f.age + education + month + marital +
##       f.campaign, family = binomial, data = dfw)
## 
## Deviance Residuals:
##    Min      1Q   Median      3Q     Max
## -2.0209 -0.4060 -0.3271 -0.2619  2.8284
## 
## Coefficients:
##                               Estimate Std. Error
## (Intercept)                -1.172e+02  4.731e+01
## emp.var.rate                 -8.825e-01  4.864e-01
## f.pdaysf.pdays-never        -3.195e-01  6.806e-01
## previous                      -1.861e-01  2.263e-01
## cons.price.idx                  1.226e+00  4.978e-01
## f.cons.conf.idxf.cons.conf.idx-(-42.7,-41.8] -1.839e-01  5.606e-01
## f.cons.conf.idxf.cons.conf.idx-(-41.8,-36.4]  8.917e-01  4.222e-01
## f.cons.conf.idxf.cons.conf.idx-(-36.4,-26.9]  6.494e-01  3.654e-01
## f.nr.employedf.nr.employed-(5.1e+03,5.19e+03] -5.090e-01  1.319e+00
## f.nr.employedf.nr.employed-(5.19e+03,5.23e+03]  3.442e-02  1.460e+00
## poutcomepoutcome-nonexistent  2.584e-01  3.287e-01
## poutcomepoutcome-success      1.107e+00  6.593e-01
## f.euribor3mf.euribor3m-(1.33,4.86]  1.395e-01  3.972e-01
## f.euribor3mf.euribor3m-(4.86,4.96]  8.233e-01  4.758e-01
## f.euribor3mf.euribor3m-(4.96,5]  5.810e-01  5.496e-01
## contactcontact-telephone     -4.058e-01  2.195e-01
## defaultdefault-unknown       -1.550e-01  1.864e-01
## f.agef.age-(32,38]           -3.159e-01  1.692e-01

```

```

## f.agef.age-(38,47]          -2.719e-01  1.817e-01
## f.agef.age-(47,87]          -1.099e-01  1.777e-01
## educationeducation-basic.6y 1.258e-03  3.308e-01
## educationeducation-basic.9y -9.690e-02  2.532e-01
## educationeducation-high.school -1.759e-02  2.276e-01
## educationeducation-professional.course 2.463e-01  2.474e-01
## educationeducation-university.degree 2.076e-01  2.153e-01
## monthmonth-aug             -8.638e-02  4.435e-01
## monthmonth-dec             -7.972e-01  6.361e-01
## monthmonth-jul             -6.037e-02  4.813e-01
## monthmonth-jun             -6.067e-01  4.825e-01
## monthmonth-mar             5.768e-01  4.024e-01
## monthmonth-may             -3.841e-01  4.299e-01
## monthmonth-nov             -5.517e-01  5.415e-01
## monthmonth-oct             -5.562e-02  4.900e-01
## monthmonth-sep             -2.901e-01  4.559e-01
## maritalmarital-married    7.928e-02  1.972e-01
## maritalmarital-single     7.256e-02  2.238e-01
## f.campaignf.campaign-(2,5] 8.189e-02  1.405e-01
## f.campaignf.campaign-(5,25] -2.359e-01  2.540e-01
##
## (Intercept)                  -2.477   0.0133 *
## emp.var.rate                 -1.814   0.0696 .
## f.pdaysf.pdays-never        -0.469   0.6387
## previous                      -0.822   0.4110
## cons.price.idx                2.464   0.0137 *
## f.cons.conf.idxf.cons.conf.idx(-42.7,-41.8] -0.328   0.7428
## f.cons.conf.idxf.cons.conf.idx(-41.8,-36.4]  2.112   0.0347 *
## f.cons.conf.idxf.cons.conf.idx(-36.4,-26.9]  1.777   0.0755 .
## f.nr.employedf.nr.employed-(5.1e+03,5.19e+03] -0.386   0.6996
## f.nr.employedf.nr.employed-(5.19e+03,5.23e+03]  0.024   0.9812
## poutcomepoutcome-nonexistent 0.786   0.4319
## poutcomepoutcome-success     1.680   0.0930 .
## f.euribor3mf.euribor3m-(1.33,4.86]  0.351   0.7254
## f.euribor3mf.euribor3m-(4.86,4.96]  1.730   0.0836 .
## f.euribor3mf.euribor3m-(4.96,5]      1.057   0.2904
## contactcontact-telephone     -1.849   0.0645 .
## defaultdefault-unknown       -0.831   0.4057
## f.agef.age-(32,38]           -1.867   0.0619 .
## f.agef.age-(38,47]           -1.496   0.1346
## f.agef.age-(47,87]           -0.619   0.5362
## educationeducation-basic.6y  0.004   0.9970
## educationeducation-basic.9y  -0.383   0.7020
## educationeducation-high.school -0.077   0.9384
## educationeducation-professional.course 0.995   0.3195
## educationeducation-university.degree 0.964   0.3348
## monthmonth-aug               -0.195   0.8456
## monthmonth-dec               -1.253   0.2101
## monthmonth-jul               -0.125   0.9002
## monthmonth-jun               -1.257   0.2086
## monthmonth-mar               1.433   0.1518
## monthmonth-may               -0.893   0.3717
## monthmonth-nov               -1.019   0.3082
## monthmonth-oct               -0.114   0.9096

```

```

## monthmonth-sep -0.636 0.5246
## maritalmarital-married 0.402 0.6877
## maritalmarital-single 0.324 0.7457
## f.campaignf.campaign-(2,5] 0.583 0.5600
## f.campaignf.campaign-(5,25] -0.929 0.3528
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2619.2 on 3738 degrees of freedom
## Residual deviance: 2116.9 on 3701 degrees of freedom
## AIC: 2192.9
##
## Number of Fisher Scoring iterations: 6
vif(gm4) #valor de colinealitat en la 2a columna (eliminem fins a obtenir tots els valors <3)

##          GVIF Df GVIF^(1/(2*Df))
## emp.var.rate 221.469476 1 14.881851
## f.pdays      11.753022 1 3.428268
## previous     6.098233 1 2.469460
## cons.price.idx 31.890396 1 5.647158
## f.cons.conf.idx 219.801351 3 2.456622
## f.nr.employed 1148.647706 2 5.821658
## poutcome     30.415086 2 2.348401
## f.euribor3m   104.117842 3 2.168973
## contact       2.676215 1 1.635914
## default       1.156703 1 1.075501
## f.age          1.574669 3 1.078611
## education      1.320436 5 1.028186
## month         1483.148770 9 1.500296
## marital        1.422431 2 1.092088
## f.campaign    1.078035 2 1.018962

gm4<-glm(y~previous+cons.price.idx+f.cons.conf.idx+poutcome+f.euribor3m+contact+default+f.age
           +education+month+marital+f.campaign, family=binomial, data=dfw); summary(gm4); vif(gm4)

##
## Call:
## glm(formula = y ~ previous + cons.price.idx + f.cons.conf.idx +
##       poutcome + f.euribor3m + contact + default + f.age + education +
##       month + marital + f.campaign, family = binomial, data = dfw)
##
## Deviance Residuals:
##      Min        1Q        Median        3Q        Max 
## -1.8595   -0.4196   -0.3331   -0.2717    2.8330 
##
## Coefficients:
##                               Estimate Std. Error z value
## (Intercept)                 -8.025e+00  1.504e+01 -0.534
## previous                   -9.403e-02  2.023e-01 -0.465
## cons.price.idx                8.483e-02  1.609e-01  0.527
## f.cons.conf.idxf.cons.conf.idxx(-42.7,-41.8] -5.047e-01  4.076e-01 -1.238
## f.cons.conf.idxf.cons.conf.idxx(-41.8,-36.4]  9.311e-01  2.464e-01  3.778

```

```

## f.cons.conf.idxf.cons.conf.idx-(-36.4,-26.9] 1.214e-01 2.709e-01 0.448
## poutcomepoutcome-nonexistent 2.804e-01 3.186e-01 0.880
## poutcomepoutcome-success 1.447e+00 2.680e-01 5.397
## f.euribor3mf.euribor3m-(1.33,4.86] -1.606e+00 2.688e-01 -5.973
## f.euribor3mf.euribor3m-(4.86,4.96] -1.212e+00 2.632e-01 -4.606
## f.euribor3mf.euribor3m-(4.96,5] -1.688e+00 2.763e-01 -6.109
## contactcontact-telephone -3.760e-01 2.046e-01 -1.838
## defaultdefault-unknown -2.144e-01 1.852e-01 -1.158
## f.agef.age-(32,38] -3.252e-01 1.677e-01 -1.939
## f.agef.age-(38,47] -2.509e-01 1.801e-01 -1.394
## f.agef.age-(47,87] -7.597e-02 1.764e-01 -0.431
## educationeducation-basic.6y -3.253e-02 3.277e-01 -0.099
## educationeducation-basic.9y -1.432e-01 2.513e-01 -0.570
## educationeducation-high.school -6.174e-02 2.253e-01 -0.274
## educationeducation-professional.course 1.846e-01 2.454e-01 0.752
## educationeducation-university.degree 1.908e-01 2.127e-01 0.897
## monthmonth-aug -1.283e+00 3.746e-01 -3.424
## monthmonth-dec -1.380e+00 6.047e-01 -2.282
## monthmonth-jul -1.297e+00 3.379e-01 -3.839
## monthmonth-jun -1.321e+00 4.690e-01 -2.817
## monthmonth-mar -9.427e-05 4.013e-01 0.000
## monthmonth-may -2.156e+00 3.174e-01 -6.792
## monthmonth-nov -1.203e+00 3.909e-01 -3.078
## monthmonth-oct -5.709e-01 3.857e-01 -1.480
## monthmonth-sep -1.081e+00 4.439e-01 -2.435
## maritalmarital-married 1.148e-01 1.974e-01 0.582
## maritalmarital-single 1.178e-01 2.233e-01 0.527
## f.campaignf.campaign-(2,5] 1.108e-01 1.390e-01 0.797
## f.campaignf.campaign-(5,25] -2.182e-01 2.521e-01 -0.866
## Pr(>|z|)
## (Intercept) 0.593560
## previous 0.642085
## cons.price.idx 0.597954
## f.cons.conf.idxf.cons.conf.idx-(-42.7,-41.8] 0.215647
## f.cons.conf.idxf.cons.conf.idx-(-41.8,-36.4] 0.000158 ***
## f.cons.conf.idxf.cons.conf.idx-(-36.4,-26.9] 0.653985
## poutcomepoutcome-nonexistent 0.378757
## poutcomepoutcome-success 6.77e-08 ***
## f.euribor3mf.euribor3m-(1.33,4.86] 2.33e-09 ***
## f.euribor3mf.euribor3m-(4.86,4.96] 4.10e-06 ***
## f.euribor3mf.euribor3m-(4.96,5] 1.00e-09 ***
## contactcontact-telephone 0.066101 .
## defaultdefault-unknown 0.246929
## f.agef.age-(32,38] 0.052509 .
## f.agef.age-(38,47] 0.163427
## f.agef.age-(47,87] 0.666642
## educationeducation-basic.6y 0.920933
## educationeducation-basic.9y 0.568902
## educationeducation-high.school 0.784055
## educationeducation-professional.course 0.451778
## educationeducation-university.degree 0.369771
## monthmonth-aug 0.000617 ***
## monthmonth-dec 0.022475 *
## monthmonth-jul 0.000123 ***

```

```

## monthmonth-jun          0.004853 **
## monthmonth-mar          0.999813
## monthmonth-may          1.10e-11 ***
## monthmonth-nov          0.002087 **
## monthmonth-oct          0.138834
## monthmonth-sep          0.014906 *
## maritalmarital-married 0.560792
## maritalmarital-single   0.597989
## f.campaignf.campaign-(2,5] 0.425166
## f.campaignf.campaign-(5,25] 0.386737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2619.2  on 3738  degrees of freedom
## Residual deviance: 2140.3  on 3705  degrees of freedom
## AIC: 2208.3
##
## Number of Fisher Scoring iterations: 6
##                               GVIF Df GVIF^(1/(2*Df))
## previous           4.846187 1    2.201406
## cons.price.idx    3.254165 1    1.803930
## f.cons.conf.idx   28.988886 3    1.752691
## poutcome          5.029068 2    1.497517
## f.euribor3m      16.970692 3    1.603061
## contact           2.340668 1    1.529924
## default            1.147276 1    1.071110
## f.age              1.545620 3    1.075269
## education          1.300784 5    1.026646
## month              64.251018 9    1.260195
## marital            1.411570 2    1.089998
## f.campaign         1.070570 2    1.017194

```



#### Anova(gm4)

```

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##             LR Chisq Df Pr(>Chisq)
## previous       0.216  1   0.64181
## cons.price.idx 0.278  1   0.59809
## f.cons.conf.idx 35.173  3  1.120e-07 ***
## poutcome        30.291  2  2.645e-07 ***
## f.euribor3m    69.337  3  1.8e-15 ***
## contact         3.516  1   0.06078 .
## default          1.383  1   0.23951
## f.age            4.815  3   0.18588
## education        4.885  5   0.43004
## month            81.877  9  6.843e-14 ***
## marital          0.361  2   0.83497
## f.campaign       1.687  2   0.43025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

gm44<-step(gm4, k=log(nrow(dfw)))
vif(gm44)
summary(gm44)

#provem si f.season va millor que la variable month
gm44season<-glm(y~f.cons.conf.idx+poutcome+f.euribor3m+f.season, family=binomial, data=dfw)
#equivalent al test de Fisher pero per al target factor (binari):
anova(gm44, gm44season, test="Chisq") #Pr(>Chi) < 2.2e-1 -> no son equivalents, per tant el model

## Analysis of Deviance Table
##
## Model 1: y ~ f.cons.conf.idx + poutcome + f.euribor3m + month
## Model 2: y ~ f.cons.conf.idx + poutcome + f.euribor3m + f.season
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3721    2158.4
## 2      3728    2259.4 -7   -100.97 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#mes gran (amb month que te mes categories) hauria de ser millor

#podriem crear una nova variable a partir de months que sigui months bons i dolents,
#podria simplificar el model!! ✓

```

## Interaccions

En aquest apartat intentarem millorar el nostre model actual (gm44) a traves de les interaccions. Establirem un model inicial (gmint) que anira acumulant les millores en el cas de que l'aportacio de les interaccions sigui positiva per al model. En el primer cas provarem la interaccio f.euribor3m amb f.cons.conf.idx, la qual veiem que ens aporta informacio al model.

```

gmint<-glm(y~f.cons.conf.idx+poutcome+f.euribor3m+month, family=binomial, data=dfw)
gmint1<-glm(y~f.cons.conf.idx*f.euribor3m+poutcome+month, family=binomial, data=dfw)

```

```
Anova(gmint1); anova(gmint, gmint1, test="Chisq") #H0 rej -> models no equivalents
```

```

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##                         LR Chisq Df Pr(>Chisq)
## f.cons.conf.idx          43.105  3  2.337e-09 ***
## f.euribor3m              93.292  3  < 2.2e-16 ***
## poutcome                  32.667  2  8.062e-08 ***
## month                      25.674  9  0.002309 **
## f.cons.conf.idx:f.euribor3m 30.062  6  3.825e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Deviance Table
##
## Model 1: y ~ f.cons.conf.idx + poutcome + f.euribor3m + month
## Model 2: y ~ f.cons.conf.idx * f.euribor3m + poutcome + month
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3721    2158.4
## 2      3715    2128.3  6   30.062 3.825e-05 ***
## ---

```

```
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

A continuacio, tot i que segons tests Anova anteriors cap variable numerica ens aportava informacio significativa, provarem la interaccion entre una variable factor i una numerica, tal i com se'sns diu a la practica. Sera el cas de la interaccio entre cons.price.idx i f.euribor3m, la qual l'affegirem al model gmint1. Veiem doncs com si que aquesta variable numerica en forma d'interaccio ens aporta infomacio significativa i ens quedem amb el model gmint2. Si mitjancant la comanda step simplifiquem el model, obtenim el model gmint3, que trobem no ser equivalent al model gmint2, pel qual no acabarem de fer-li cas a la comanda step i ens quedarem amb el model anterior gmint2.
```

```
gmint2<-glm(y~cons.price.idx:f.euribor3m+poutcome+month+f.cons.conf.idx*f.euribor3m,
              family=binomial, data=dfw)
Anova(gmint2); anova(gmint1, gmint2, test="Chisq") #H0 rej -> models no equivalents
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: y
##                         LR Chisq Df Pr(>Chisq)
## poutcome                  30.767  2  2.084e-07 ***
## month                      12.929  9   0.165863
## f.cons.conf.idx            47.082  3  3.339e-10 ***
## f.euribor3m                 93.292  3 < 2.2e-16 ***
## cons.price.idx:f.euribor3m    9.542  3   0.022893 *
## f.euribor3m:f.cons.conf.idx 18.694  5   0.002191 **
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ f.cons.conf.idx * f.euribor3m + poutcome + month
## Model 2: y ~ cons.price.idx:f.euribor3m + poutcome + month + f.cons.conf.idx *
##             f.euribor3m
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3715     2128.3
## 2      3712     2118.8  3    9.5416  0.02289 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
gmint3<-step(gmint2, k=log(nrow(dfw)))
```

```
## Start: AIC=2340.91
## y ~ cons.price.idx:f.euribor3m + poutcome + month + f.cons.conf.idx *
##             f.euribor3m
##
##                         Df Deviance     AIC
## - month                      9   2131.7 2279.8
## - f.euribor3m:f.cons.conf.idx 5   2137.5 2318.5
## - cons.price.idx:f.euribor3m   3   2128.3 2325.8
## <none>                      2118.8 2340.9
## - poutcome                     2   2149.6 2355.2
##
## Step: AIC=2279.8
## y ~ poutcome + f.cons.conf.idx + f.euribor3m + cons.price.idx:f.euribor3m +
##             f.euribor3m:f.cons.conf.idx
##
##                         Df Deviance     AIC
```

```

## - f.cons.conf.idx:f.euribor3m 5 2166.0 2273.0
## - f.euribor3m:cons.price.idx 3 2154.0 2277.4
## <none> 2131.7 2279.8
## - poutcome 2 2160.9 2292.5
##
## Step: AIC=2272.98
## y ~ poutcome + f.cons.conf.idx + f.euribor3m + f.euribor3m:cons.price.idx
##
##                                     Df Deviance    AIC
## <none>                         2166.0 2273.0
## - poutcome                      2 2198.2 2288.7
## - f.cons.conf.idx                3 2279.5 2361.8
## - f.euribor3m:cons.price.idx    4 2288.1 2362.1
anova(gmint2, gmint3, test="Chisq")

## Analysis of Deviance Table
##
## Model 1: y ~ cons.price.idx:f.euribor3m + poutcome + month + f.cons.conf.idx *
##           f.euribor3m
## Model 2: y ~ poutcome + f.cons.conf.idx + f.euribor3m + f.euribor3m:cons.price.idx
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3712     2118.8
## 2      3726     2166.0 -14   -47.234 1.759e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#H0 rej -> models no equivalents

```

### Interaction between a couple of factors in our model mgint2

El model escollit mgint2 considera una interacció entre dos factors *f.cons.conf.idx:f.euribor3m*.

```
table(dfw$f.euribor3m, dfw$f.cons.conf.idx)
```

```

##
##                                     f.cons.conf.idx-[-50.8,-42.7]
## f.euribor3m-[0.635,1.33]          507
## f.euribor3m-(1.33,4.86]          281
## f.euribor3m-(4.86,4.96]          241
## f.euribor3m-(4.96,5]             351
##
##                                     f.cons.conf.idx-(-42.7,-41.8]
## f.euribor3m-[0.635,1.33]          0
## f.euribor3m-(1.33,4.86]          328
## f.euribor3m-(4.86,4.96]          347
## f.euribor3m-(4.96,5]             53
##
##                                     f.cons.conf.idx-(-41.8,-36.4]
## f.euribor3m-[0.635,1.33]          164
## f.euribor3m-(1.33,4.86]          499
## f.euribor3m-(4.86,4.96]          266
## f.euribor3m-(4.96,5]              3
##
##                                     f.cons.conf.idx-(-36.4,-26.9]
## f.euribor3m-[0.635,1.33]          260

```

```

##   f.euribor3m-(1.33,4.86]          0
##   f.euribor3m-(4.86,4.96]          0
##   f.euribor3m-(4.96,5]           439

```

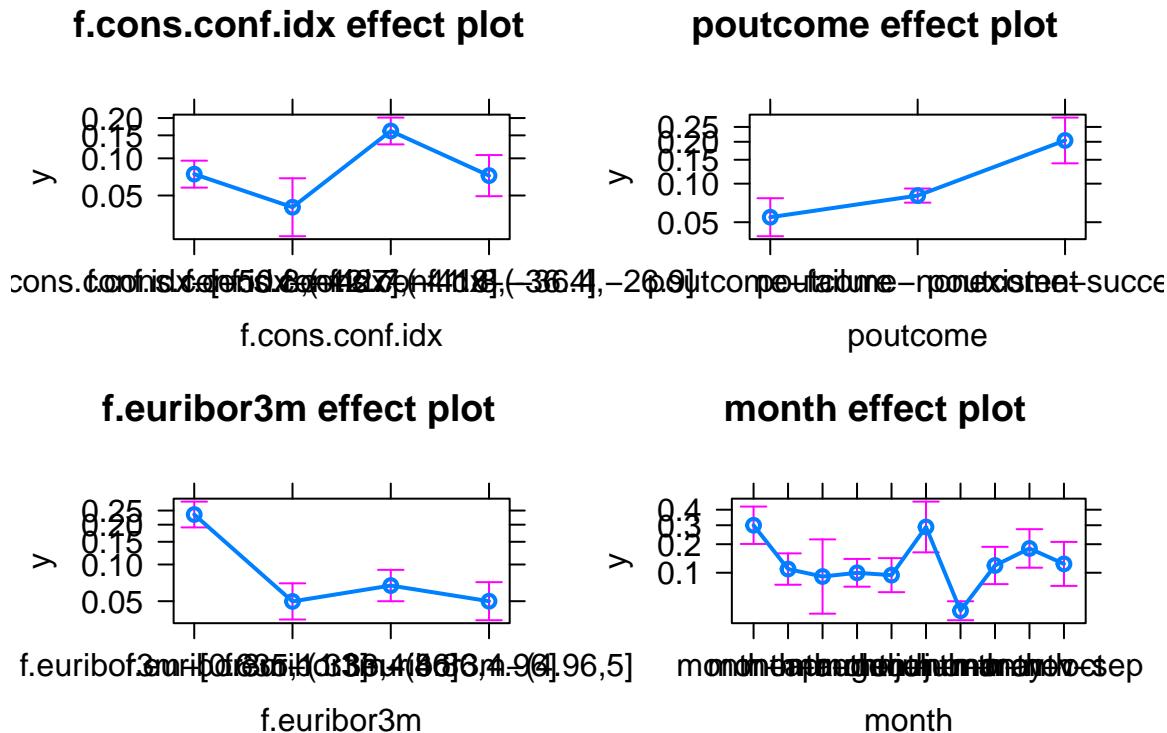
### Interaction between a factor and a covariate in our model mgint2

El model escollit mgint2 també considera una interacció entre un factor i una variable numèrica *cons.price.idx:f.euribor3m*.

```

#model sense interaccions
plot(allEffects(gmint))

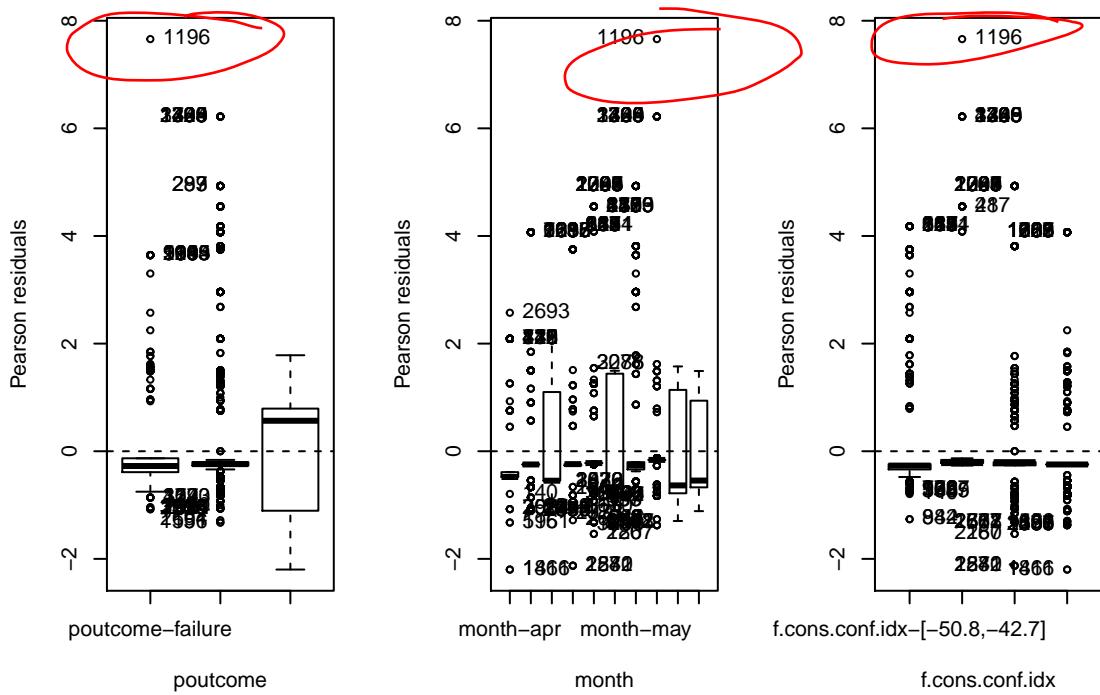
```



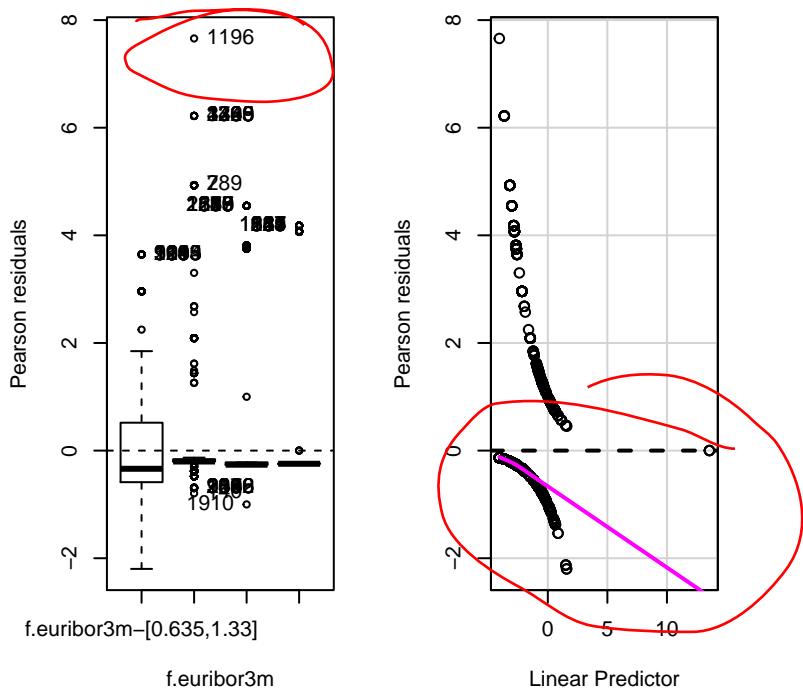
### Diagnostics del model definitiu gmint2

En els 5 primers grafics de les dues primeres figures que hi ha a continuació es poden veure els residus de les prediccions del model per a cada variable explicativa. El 6e grafic mostra un test de curvatura. En les següents figures i taula veiem les distàncies de Cook i els valors barret per a cadascuna de les observacions. Les observacions 24043 i 24031 son observacions molt influents en el model.

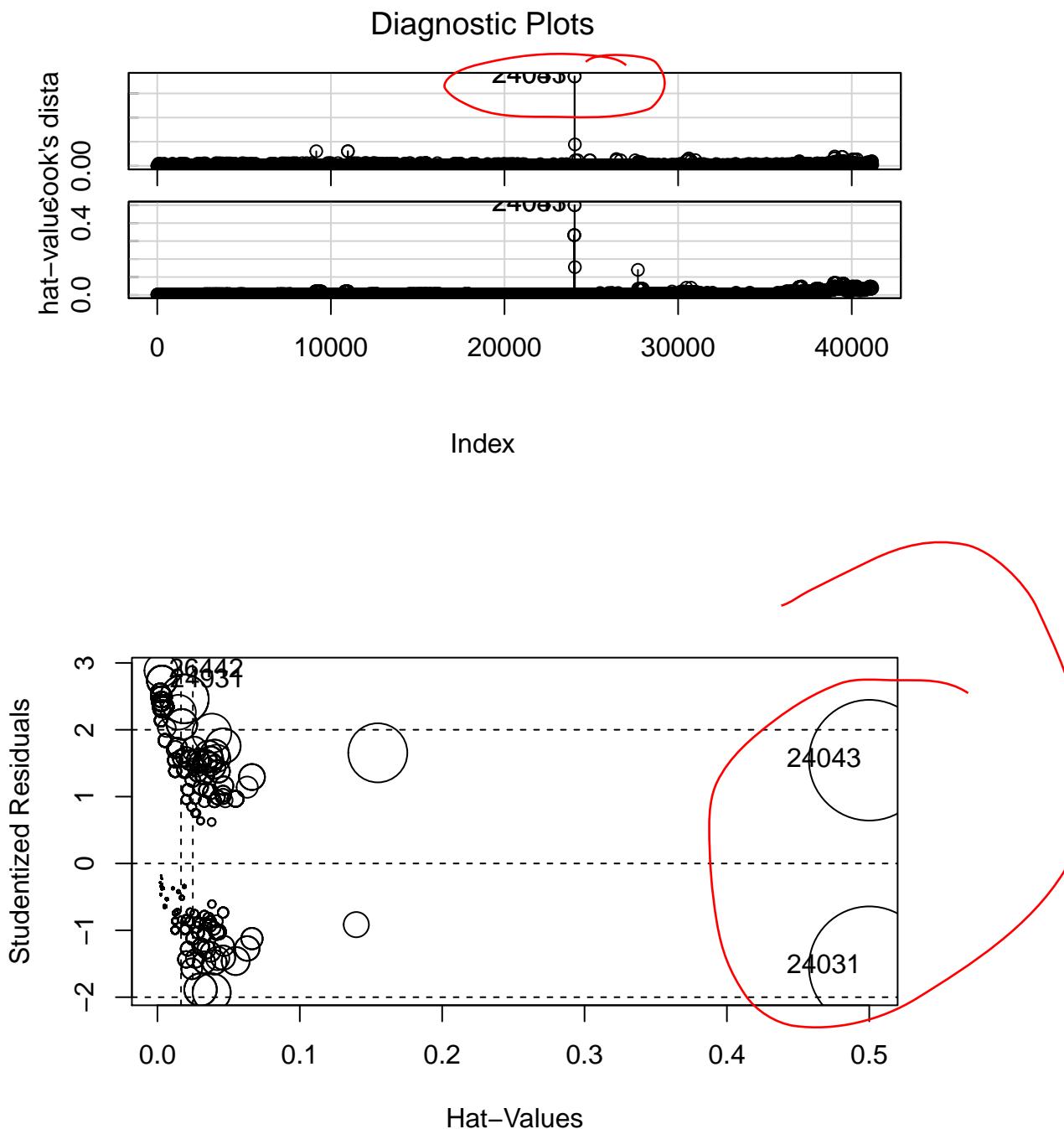
```
residualPlots(gmint2, layout=c(1, 3))
```



```
## Warning in residualPlots.default(model, ...): No possible lack-of-fit tests
```



```
influenceIndexPlot(gmint2, vars=c("Cook", "hat")); influencePlot(gmint2)
```



```
## 24043 1.544764 0.500000000 0.074074074
## 24031 -1.544764 0.500000000 0.074074074
```

## Predictions i matriu de confusió de l'acceptació del producte financer $y$

Primer utilitzarem el model obtingut per a predir la resposta  $y$  de les dades de treball  $dfw$ , que son les mateixes amb les quals s'ha construït el model. Després farem el mateix però amb les dades de test  $dft$ , les quals no s'han usat per a la construcció del model. Si obtenim resultats similars en ambdós tests voldrà dir que el nostre model està correctament parametritzat i no ens hem adaptat a les dades de treball a l'hora de construir-lo. Veiem com per ambdós jocs de dades obtenim els mateixos resultats, amb una precisió del 90% d'ençert a l'hora de determinar l'acceptació o no del producte financer. Les prediccions correctes d'acceptació del producte financer o *sensitivity* són del 60-70%, mentre que les prediccions correctes de rebutj del producte financer són del 90%.

```
#work data
predw<-predict(gmint2, type="response")
predictionw<-prediction(predw, dfw$y)
predw.y <- factor(ifelse(as.numeric(predw)<0.5, 0, 1), labels=c("predw.y-no", "predw.y-yes"))
tablew<-addmargins(table(predw.y, dfw$y)); tablew

##
## predw.y      y-no y-yes Sum
##   predw.y-no 3265 334 3599
##   predw.y-yes  56  84 140
##   Sum          3321 418 3739

#test data
predt<-predict(gmint2, type="response", newdata=dft)

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
predictiont<-prediction(predt, dft$y)
predt.y <- factor(ifelse(as.numeric(predt)<0.5, 0, 1), labels=c("predt.y-no", "predt.y-yes"))
tablet<-addmargins(table(predt.y, dft$y)); tablet

##
## predt.y      y-no y-yes Sum
##   predt.y-no 1095 108 1203
##   predt.y-yes  13  31 44
##   Sum          1108 139 1247

#confusion matrix values
predicions_correctes_w<-sum(diag(tablew[1:2, 1:2]))/sum(tablew[1:2, 1:2])*100; predicions_correctes_w

## [1] 89.5694
predicions_correctes_t<-sum(diag(tablet[1:2, 1:2]))/sum(tablet[1:2, 1:2])*100; predicions_correctes_t

## [1] 90.29671
predicions_incorrectes_w<-(100-predicions_correctes_w); predicions_incorrectes_w

## [1] 10.4306
predicions_incorrectes_t<-(100-predicions_correctes_t); predicions_incorrectes_t

## [1] 9.703288
```

```

sensibility_w<-tablew[2,2]/sum(tablew[2, 1:2])*100; sensibility_w
## [1] 60
sensibility_t<-tablet[2,2]/sum(tablet[2, 1:2])*100; sensibility_t
## [1] 70.45455
specificity_w<-tablew[1,1]/sum(tablew[1, 1:2])*100; specificity_w
## [1] 90.71964
specificity_t<-tablet[1,1]/sum(tablet[1, 1:2])*100; specificity_t
## [1] 91.02244

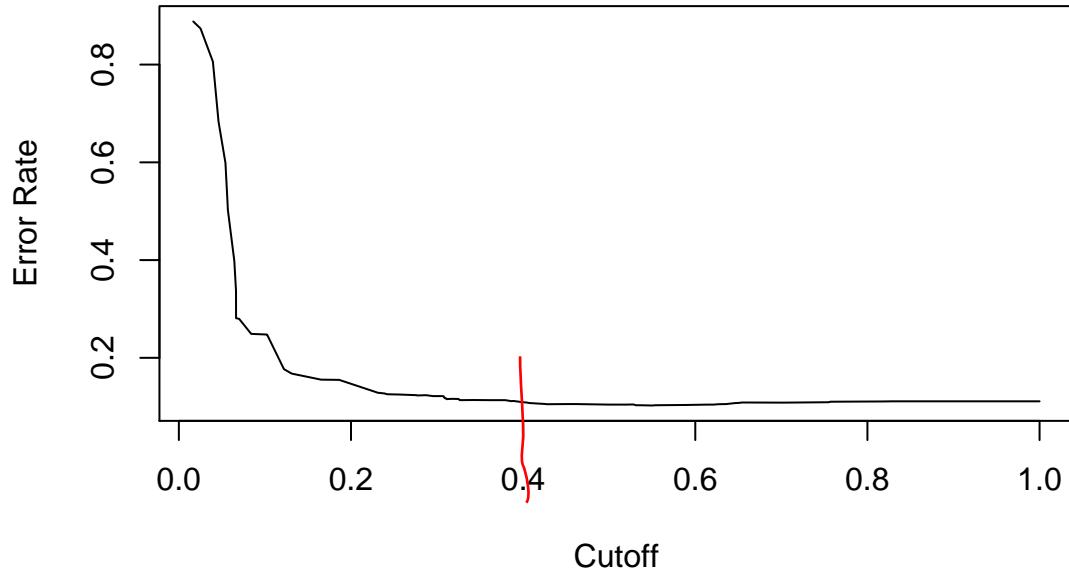
```

El llindar seleccionat anteriorment per a decidir si una probabilitat es una acceptació o un rebuig del producte ha estat l'standard de 0.5; a continuació també es buscarà si es poden millorar els resultats considerant un altre valor threshold, el qual valorarem a partir de les curves ROC. En el primer grafic veiem com el llindar actual ens garantitza el maxim % d'encert global del model; en el segon grafic però, veiem també com la sensibilitat del model té encara marge de millora, a canvi d'incrementar els falsos positius, es a dir, predir una acceptació del producte financer i que finalment aquest sigui rebutjat.

```

dadesroc<-prediction(predict(gmint2, type="response"), dfw$y)
plot(performance(dadesroc, "err"));

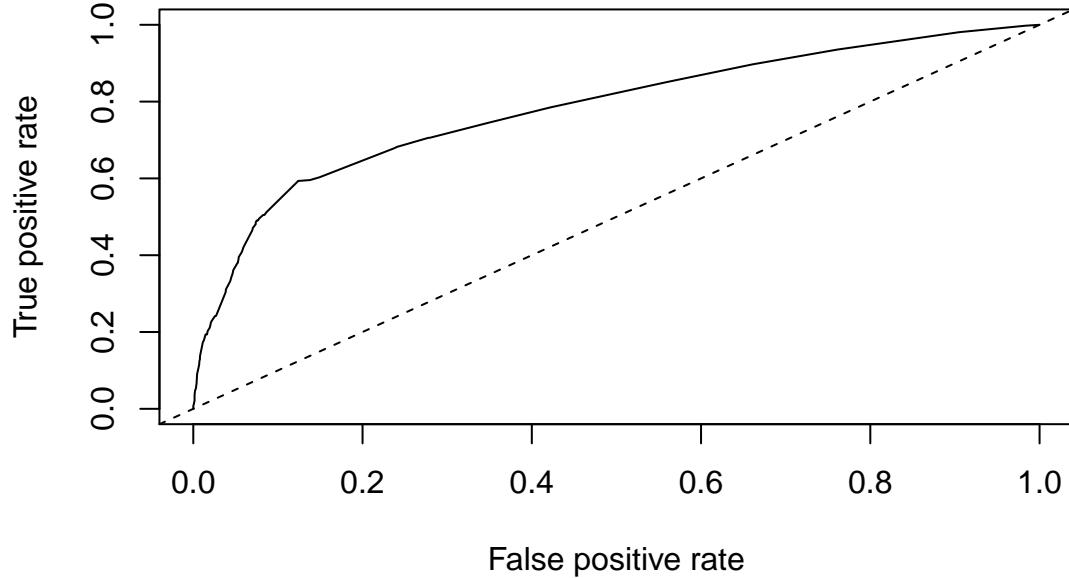
```



```

plot(performance(dadesroc, "tpr", "fpr"))
abline(0,1,lty=2)

```



Si canviem el valor llindar *cut off* a 0.7, el que vol dir que serem “mes pesimistes” a l’hora de dir que el client acceptara el producte financer a partir de la probabilitat donada pel model, obtenim uns millors resultats de sensibilitat (per sobre el 70% en ambdues mostres), mentres que la resta de valors segueixen similars. Tot i això, aquest valor es podria ajustar segons l’aplicació física real del model predictiu. Si ens interessa estar molt segurs que si predim un outcome positiu així sigui, hauríem de pujar el llindar, fet que comportaria que ens estiguessim predint algun valor positiu com a negatiu. Si el que destijem es el contrari, i simplement volem descartar nomes casos que l’outcome seria negatiu amb molta seguretat, hauríem de baixar el llindar.

```
#work data
predw<-predict(gmint2, type="response")
predictionw<-prediction(predw, dfw$y)
predw.y <- factor(ifelse(as.numeric(predw)<0.7, 0, 1), labels=c("predw.y-no", "predw.y-yes"))
tablew<-addmargins(table(predw.y, dfw$y)); tablew

##
## predw.y      y-no y-yes Sum
##   predw.y-no 3314  398 3712
##   predw.y-yes    7   20  27
##   Sum         3321  418 3739

#test data
predt<-predict(gmint2, type="response", newdata=dft)

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
predictiont<-prediction(predt, dft$y)
predt.y <- factor(ifelse(as.numeric(predt)<0.7, 0, 1), labels=c("predt.y-no", "predt.y-yes"))
tablet<-addmargins(table(predt.y, dft$y)); tablet

##
```

```

## predt.y      y-no y-yes Sum
##   predt.y-no 1105   131 1236
##   predt.y-yes    3     8  11
##   Sum          1108   139 1247
#confussion matrix values
predicions_correctes_w<-sum(diag(tablew[1:2, 1:2]))/sum(tablew[1:2, 1:2])*100; predicions_correctes_w

## [1] 89.16823
predicions_correctes_t<-sum(diag(tablet[1:2, 1:2]))/sum(tablet[1:2, 1:2])*100; predicions_correctes_t

## [1] 89.25421
predicions_incorrectes_w<-(100-predicions_correctes_w); predicions_incorrectes_w

## [1] 10.83177
predicions_incorrectes_t<-(100-predicions_correctes_t); predicions_incorrectes_t

## [1] 10.74579
sensibility_w<-tablew[2,2]/sum(tablew[2, 1:2])*100; sensibility_w

## [1] 74.07407
sensibility_t<-tablet[2,2]/sum(tablet[2, 1:2])*100; sensibility_t

## [1] 72.72727
specificity_w<-tablew[1,1]/sum(tablew[1, 1:2])*100; specificity_w

## [1] 89.27802
specificity_t<-tablet[1,1]/sum(tablet[1, 1:2])*100; specificity_t

## [1] 89.40129

```