

Deliverable III: Numeric and Binary targets Forecasting Models



INDEX

Input variables	3
Model construction only with numeric explanatory variables	4
Multivariant Data Analysis	4
Model Construction	4
Initial Modelling	5
Transforming variables	11
Adding factors as explanatory variables	13
Interactions between numeric variables and factors	25
Model Additiu	25
Model interaccions	28
Binary Regression	30
Explanatory numeric variables	30
Initial modelling	30
Transforming variables	34
Adding factors	35
Adding new factors	37
Add to the best model: INTERACTIONS	42
Model final	46
Interpretació del model final	47
Validació del model	48
Anàlisi dels residus	48
Predicció	51
Work	51
Test	51

Deliverable3

Montserrat Martinez i Aleix Costa

02 de Juny de 2019

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')

Model construction only with numeric explanatory variables

Multivariant Data Analysis

Ara el que farem serà analitzar quines són les variables numèriques més relacionades amb el nostre target *duration*, per tal de decidir quines d'aquestes utilitzarem en la construcció dels diferents models fins trobar l'òptim.

#En vars_model també tenim la variable "duration" perquè és necessari per poder veure les més relacionades amb aquesta

```
vars_model<-names(df)[c(1,11:14,16:20)]; vars_model
```

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

```
condes(df[,vars_model],which(vars_model == "duration"))
```

```
## $quanti
##          correlation      p.value
## previous    0.02859224 4.435374e-02
## nr.employed -0.03619203 1.091224e-02
## campaign    -0.04179341 3.284450e-03
## pdays       -0.06147234 1.516945e-05
```

A partir d'executar la comanda “condes” podem veure que les variables més relacionades són *previous*, *nr.employed*, *campaign* i *pdays*, tot i que la correlació que presenten és molt baixa i poc significativa. Tot i així les podem considerar com a candidates a formar part de la construcció del nostre model.

Model Construction

A partir de tot l'anàlisi realitzat fins ara, començarem la construcció dels models, partint d'un model més complex de totes les variables numèriques. Realitzarem diferents anàlisis per a cada model fins a trobar el model més adient o òptim a la nostra situació o joc de dades.

Initial modelling

~~#Las variables socioeconomicas estan relacionadas entre ellas, pero no tienen nada que ver con el target~~

```
#vars_exp<-names(df)[c(1,12:14,16:20)]; vars_exp
```

```
vars_conaux #numèriques = vars_exp
```

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

```
#vars_con_aux2 #numeriques (sense age) que es la que utilitzem!
condes(df,11)
```

```
## $quanti
##               correlation      p      ue
## previous      0.02859224 4.4353 0.02
## errors_indiv  -0.03476735 1.447588e-02
## nr.employed   -0.03619203 1.091224e-02
## CLUSTER       -0.04004368 4.853468e-03
## campaign      -0.04179341 3.284450e-03
## pdays         -0.06147234 1.516945e-05
## missings_indiv -0.07328498 2.474678e-07
##
## $quali
##               R2      p.value
## factor_duration 0.8271873066 0.000000e+00
## y               0.1863696068 9.891372e-224
## factor_Pdays   0.0051824450 4.017238e-07
## poutcome        0.0041874670 3.132625e-05
## f.CLUSTER       0.0061553592 3.146859e-05
## month           0.0073478185 3.327154e-05
## factor_cons.price.idx 0.0039803615 5.696640e-04
## factor_Previous  0.0019228074 2.038492e-03
## day_of_week      0.0029955473 5.075577e-03
## factor_cons.conf.idx 0.0026002247 1.194404e-02
## contact          0.0011105265 1.909343e-02
## default          0.0009897216 2.693284e-02
## factor_campaign  0.0013152237 3.866909e-02
##
## $category
##               Estimate      p.value
## factor_duration-(504,2.12e+03] 547.162252 0.000000e+00
## Y_yes                          169.675531 9.891372e-224
```

```
## factor_duration-(329,504]      138.462468  3.985182e-48
## factor_Pdays-[0,15]           49.355073  4.017238e-07
## CLUSTER-4                      82.017790  5.318613e-06
## Poutcome_success               62.641078  7.933875e-06
## factor_cons.price.idx-(93.4,93.9] 27.117765  2.010384e-04
## Month_jul                      12.946601  2.986551e-04
## factor_Previous-(1,5]          34.966136  2.038492e-03
## Contact_cellular               8.850090  1.909343e-02
## Default_no                     9.913335  2.693284e-02
## Month_dec                      104.090396  2.868142e-02
## Day_of_week_tue                14.917687  4.872420e-02
## Education_illiterate           178.585152  4.932974e-02
## CLUSTER-7                      -37.598946  4.049876e-02
## Education_university.degree    -38.308971  3.857651e-02
## factor_cons.conf.idx-(-36.4,-29.8] -13.574401  3.768483e-02
## CLUSTER-5                      -20.182210  3.375761e-02
## factor_cons.conf.idx-(-42,-40.3] -17.926886  2.695593e-02
## Default_unknown                -9.913335  2.693284e-02
## Contact_telephone              -8.850090  1.909343e-02
## Month_jun                      -37.404273  1.736971e-02
## factor_campaign-(3,14]         -16.741883  1.148865e-02
## Job_technician                 -25.341033  1.106827e-02
## Day_of_week_mon                -19.239047  7.577039e-03
## Month_aug                      -39.248662  5.073298e-03
## factor_cons.price.idx-(93,93.4] -19.809889  2.312144e-03
## factor_Previous-[0,1]          -34.966136  2.038492e-03
## factor_Pdays-(15,17]          -49.355073  4.017238e-07
## factor_duration-(182,236]      -56.414720  8.764699e-09
## factor_duration-(139,182]      -103.067426  8.297196e-27
## factor_duration-(104,139]      -141.910732  3.245807e-49
## factor_duration-(68,104]       -177.221056  2.195363e-78
## factor_duration-[1,68]         -222.636796  8.250905e-127
## Y_no                           -169.675531  9.891372e-224
```

```
m1<-lm(duration~previous+euribor3m+campaign+pdays+nr.employed,data=df)
```

```
#summary(m1)
```

```
Anova(m1)
```

```
## Anova Table (Type II tests)
```

```
##
```

```
## Response: duration
```

##	Sum Sq	Df	F value	Pr(>F)
## previous	69540	1	1.0663	0.3018273
## euribor3m	393980	1	6.0413	0.0140094 *
## campaign	441217	1	6.7656	0.0093209 **
## pdays	726966	1	11.1473	0.0008478 ***

```
## nr.employed      478090      1  7.3310 0.0068008 **
## Residuals      322161286 4940
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Inferential criteria o Bayesian info criteria
# Remove non significant variables
```

#Les variables que sobran son las que tienen un pvalor por encima del 5%
#Aqui se ponen las que tengan un p valor menor que 5

Veiem que aquest model i segurament tots els que realitzarem amb el target numèric tenen una explicabilitat molt baixa (menys del 0.005 del % de les dades), i per tant serà difícil obtenir dades rellevants. Tot i així procedirem a fer un procés metodològic de “Modeling” del target numèric.

Ara el que farem és fer un segon model i només posaré les variables que tenen un p-valor per sota d'un 5%, llavors em queda el mateix model que m1 però sense les variables previous.

```
m2<-lm(duration~euribor3m+campaign+pdays+nr.employed,data=df)
summary(m2)
```

```
##
## Call:
## lm(formula = duration ~ euribor3m + campaign + pdays + nr.employed,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -305.24 -158.09  -83.76   65.34 1858.59
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2419.0524    788.7005   3.067  0.00217 **
## euribor3m     15.9367     6.5155   2.446  0.01448 *
## campaign     -4.7524     1.8455  -2.575  0.01005 *
## pdays        -6.2056     1.9320  -3.212  0.00133 **
## nr.employed   -0.4075     0.1584  -2.573  0.01012 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 255.4 on 4941 degrees of freedom
## Multiple R-squared:  0.006577,    Adjusted R-squared:  0.005773
## F-statistic: 8.178 on 4 and 4941 DF,  p-value: 1.434e-06
```

Anova(m2)

```
## Anova Table (Type II tests)
##
## Response: duration
##           Sum Sq   Df F value    Pr(>F)
## euribor3m   390168    1  5.9827 0.014481 *
## campaign    432446    1  6.6310 0.010051 *
## pdays      672831    1 10.3170 0.001327 **
## nr.employed  431626    1  6.6184 0.010122 *
## Residuals   322230826 4941
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#I ara el que farem serà un vif (variance inflation factor) per veure les variables explicatives del model que estan correlacionades

vif(m2)

```
##   euribor3m   campaign    pdays nr.employed
##   9.620996    1.016309    1.168931    10.105172
```

Ara en el nostre tercer model el que farem és que quan executem el vif veiem que tenim les variables nr.employed i euribor3m amb un vif > 3, llavors això no és vàlid, perquè inflarà la variança de la nostra mostra. Llavors primer el que fem és eliminar nr.employed y després en el model número 4 eliminarem euribor3m també per veure quin és el que té una millor explicabilitat.

```
m3<-lm(duration~campaign+pdays+euribor3m,data=df)
summary(m3)
```

```
##
## Call:
## lm(formula = duration ~ campaign + pdays + euribor3m, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -319.98 -159.03  -83.08   67.50 1854.92
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   391.323     28.316   13.820 < 2e-16 ***
## campaign       -4.967       1.845   -2.692  0.00712 **
## pdays         -7.505       1.866   -4.023 5.84e-05 ***
## euribor3m       0.162       2.204    0.074  0.94141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 255.5 on 4942 degrees of freedom
## Multiple R-squared:  0.005247,    Adjusted R-squared:  0.004643
## F-statistic: 8.688 on 3 and 4942 DF,  p-value: 9.541e-06

m4<-lm(duration~campaign+pdays+nr.employed,data=df)
summary(m4)

##
## Call:
## lm(formula = duration ~ campaign + pdays + nr.employed, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -316.92 -158.62  -83.03   66.73 1857.76
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  604.23452   267.59919    2.258 0.023990 *
## campaign     -4.78623     1.84642   -2.592 0.009565 **
## pdays        -6.93604     1.90973   -3.632 0.000284 ***
## nr.employed  -0.04289     0.05359   -0.800 0.423582
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 255.5 on 4942 degrees of freedom
## Multiple R-squared:  0.005374,    Adjusted R-squared:  0.004771
## F-statistic: 8.901 on 3 and 4942 DF,  p-value: 7.024e-06

m5<-step(m1, k=log(nrow(df)))

## Start:  AIC=54873.63
## duration ~ previous + euribor3m + campaign + pdays + nr.employed
##
##              Df Sum of Sq      RSS   AIC
## - previous    1      69540 322230826 54866
## - euribor3m    1      393980 322555266 54871
## - campaign     1      441217 322602503 54872
## - nr.employed  1      478090 322639376 54872
## <none>                    322161286 54874
## - pdays        1      726966 322888252 54876
##
## Step:  AIC=54866.19
## duration ~ euribor3m + campaign + pdays + nr.employed
##
##              Df Sum of Sq      RSS   AIC
```



```

## - euribor3m      1      390168 322620995 54864
## - nr.employed    1      431626 322662452 54864
## - campaign       1      432446 322663273 54864
## <none>                                322230826 54866
## - pdays          1      672831 322903657 54868
##
## Step: AIC=54863.67
## duration ~ campaign + pdays + nr.employed
##
##              Df Sum of Sq      RSS   AIC
## - nr.employed  1       41810 322662805 54856
## - campaign     1       438650 323059645 54862
## <none>                                322620995 54864
## - pdays        1       861130 323482124 54868
##
## Step: AIC=54855.81
## duration ~ campaign + pdays
##
##              Df Sum of Sq      RSS   AIC
## - campaign     1       475707 323138512 54855
## <none>                                322662805 54856
## - pdays        1      1134867 323797672 54865
##
## Step: AIC=54854.59
## duration ~ pdays
##
##              Df Sum of Sq      RSS   AIC
## <none>                                323138512 54855
## - pdays       1      1225723 324364235 54865

```

#vif(m5) # Dóna error perquè tenim menys de dos variables!

```

m6<-lm(duration~campaign+pdays,data=df)
summary(m6)

```

```

##
## Call:
## lm(formula = duration ~ campaign + pdays, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -319.93 -158.86  -82.90   67.12 1855.14
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   391.279     28.307   13.82  < 2e-16 ***

```

```
## campaign      -4.953      1.835    -2.70  0.00697 **
## pdays        -7.467      1.791    -4.17  3.1e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 255.5 on 4943 degrees of freedom
## Multiple R-squared:  0.005245,    Adjusted R-squared:  0.004843
## F-statistic: 13.03 on 2 and 4943 DF,  p-value: 2.264e-06
```

```
vif(m6)
```

```
## campaign      pdays
## 1.003368 1.003368
```

Amb aquesta sortida el que podem comprobar és que les variables que són més significatives són **campaign** i **pdays**, però si fem el step veiem que la millor és **pdays**, però un model amb només una variable és molt poc i no explicaria el suficient, llavors agafem **campaign** i **pdays**.

Quan executem el vif en el nostre model definitiu veiem que les dos variables que tenim tenen un $vif < 3$, llavors això vol dir que el nostre model és correcte i que anem en bona direcció.

Transforming variables

Ara el que farem serà una transformació de les nostres variables per veure si podem explicar més en el nostre model.

```
m7 <- lm(log(duration)~previous+campaign+nr.employed+pdays,data=df)
Anova(m7)
```

```
## Anova Table (Type II tests)
##
## Response: log(duration)
##          Sum Sq   Df F value    Pr(>F)
## previous      0.1    1   0.0688   0.7931
## campaign     93.7    1 108.1953 < 2e-16 ***
## nr.employed   0.1    1   0.1424   0.7060
## pdays        17.0    1  19.5908 9.8e-06 ***
## Residuals   4277.7 4941
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m8<-lm (log(duration)~campaign+pdays,data=df)
summary(m8)
```

```
##
## Call:
## lm(formula = log(duration) ~ campaign + pdays, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.2586 -0.5401 -0.0011  0.6236  2.7295
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.88173     0.10307   57.066 < 2e-16 ***
## campaign     -0.06979     0.00668  -10.447 < 2e-16 ***
## pdays       -0.03458     0.00652   -5.303 1.19e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9303 on 4943 degrees of freedom
## Multiple R-squared:  0.02834,    Adjusted R-squared:  0.02795
## F-statistic: 72.09 on 2 and 4943 DF,  p-value: < 2.2e-16

#Polinomic regression
m9 <- lm(log(duration)~poly(campaign,2)+poly(pdays,2), data=df)
summary(m9)

##
## Call:
## lm(formula = log(duration) ~ poly(campaign, 2) + poly(pdays,
##      2), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.2184 -0.5456  0.0019  0.6134  2.8100
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.17462     0.01319 392.362 < 2e-16 ***
## poly(campaign, 2)1 -9.69878     0.92913  -10.439 < 2e-16 ***
## poly(campaign, 2)2 -4.30252     0.92758   -4.638 3.6e-06 ***
## poly(pdays, 2)1   -4.99650     0.92914   -5.378 7.9e-08 ***
## poly(pdays, 2)2   -2.94158     0.92757   -3.171 0.00153 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9275 on 4941 degrees of freedom
## Multiple R-squared:  0.03452,    Adjusted R-squared:  0.03374
## F-statistic: 44.16 on 4 and 4941 DF,  p-value: < 2.2e-16
```

```
Anova(m9)
```

```
## Anova Table (Type II tests)
##
## Response: log(duration)
##              Sum Sq   Df F value    Pr(>F)
## poly(campaign, 2) 112.3    2  65.273 < 2.2e-16 ***
## poly(pdays, 2)   33.5    2  19.477 3.755e-09 ***
## Residuals        4250.6 4941
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# marginalModelPlots(m9)
```

Com podem observar les nostres variables més significatives del nostre model són campaign y pdays, llavors com a conclusió el nostre model serà m8 que té una explicabilitat d'un 2,8%. Però quan fem la transformació logarítmica veiem que té una mica més d'explicabilitat el nostre model, perquè el Multiple R-squared és major, veiem que tenim una explicabilitat d'un 3,4%.



La diferència és sumament petita i fent les diferents execucions que venen a continuació hem vist que no hi ha cap tipus de diferència, com era correcte agafar un d'aquests dos models vam optar per agafar el m8 en comptes del m9.

CONCLUSIÓ: El Multiple R-squared (variabilitat de les dades) és molt petit i això vol dir que el nostre target és complicat d'interpretar, és a dir, no podem explicar el nostre target (duration, en aquest cas) amb les variables que tenim.

Adding factors as explanatory variables

Ara el que farem és afegir variables factors com a variables explicatives, llavors hem de trobar les que poden ser més significatives i ara a continuació farem aquest estudi.

```
vars_dis2<-names(df)[c(2:10,15,25,26:35)];vars_dis2

## [1] "job" "marital"
## [3] "education" "default"
## [5] "housing" "loan"
## [7] "contact" "month"
## [9] "day_of_week" "poutcome"
## [11] "season" "factor_age"
## [13] "factor_duration" "factor_campaign"
## [15] "factor_Pdays" "factor_Previous"
## [17] "factor_emp.var.rate" "factor_cons.price.idx"
```

```

## [19] "factor_cons.conf.idx" "factor_euribor3m"
## [21] "factor_nr.employed"

# Agafem el nostre millor model que tenim fins ara
m10<-step(m8,k=log(nrow(df)))

## Start: AIC=-692.34
## log(duration) ~ campaign + pdays
##
##           Df Sum of Sq    RSS      AIC
## <none>                4277.8 -692.34
## - pdays      1      24.342 4302.1 -672.78
## - campaign    1      94.458 4372.3 -592.82

# maux4<-step(m9,k=log(nrow(df))) Con el modelo que usa poly!

condes(df[,c("duration",vars_dis2)],1,proba = 0.01)

## $quali
##
##           R2      p.value
## factor_duration      0.827187307 0.000000e+00
## factor_Pdays        0.005182445 4.017238e-07
## poutcome             0.004187467 3.132625e-05
## month                0.007347818 3.327154e-05
## factor_cons.price.idx 0.003980361 5.696640e-04
## factor_Previous       0.001922807 2.038492e-03
## day_of_week           0.002995547 5.075577e-03
##
## $category
##
##           Estimate      p.value
## factor_duration-(504,2.12e+03] 547.16225 0.000000e+00
## factor_duration-(329,504]      138.46247 3.985182e-48
## factor_Pdays-[0,15]           49.35507 4.017238e-07
## Poutcome_success              62.64108 7.933875e-06
## factor_cons.price.idx-(93.4,93.9] 27.11777 2.010384e-04
## Month_jul                     12.94660 2.986551e-04
## factor_Previous-(1,5]          34.96614 2.038492e-03
## Day_of_week_mon               -19.23905 7.577039e-03
## Month_aug                     -39.24866 5.073298e-03
## factor_cons.price.idx-(93,93.4] -19.80989 2.312144e-03
## factor_Previous-[0,1]          -34.96614 2.038492e-03
## factor_Pdays-(15,17]          -49.35507 4.017238e-07
## factor_duration-(182,236]      -56.41472 8.764699e-09
## factor_duration-(139,182]      -103.06743 8.297196e-27
## factor_duration-(104,139]      -141.91073 3.245807e-49

```

```
## factor_duration-(68,104]          -177.22106  2.195363e-78
## factor_duration-[1,68]            -222.63680  8.250905e-127
```

Després de l'execució anterior el que hem vist són les variables més correlacionades amb el nostre model que són aquelles que tenen un p-valor $\ll 0.01$. Aquestes variables són: factor_Pdays+ poutcome+month+factor_cons.price.idx+ factor_Previous+day_of_week

Llavors ara estudiarem el cas, és a dir, al nostre model li afegim aquests factors.

```
#Avoid numeric and factors simultaneously for the same concept
m11<-lm(log(duration)~campaign+pdays+poutcome+month+factor_cons.price.idx+
factor_Previous+day_of_week,data = df)
summary(m11) #Take a look to NA estimates
```

```
##
## Call:
## lm(formula = log(duration) ~ campaign + pdays + poutcome + month +
##     factor_cons.price.idx + factor_Previous + day_of_week, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.1845 -0.5552 -0.0061  0.6031  2.6685
##
## Coefficients:
##                                     Estimate
## (Intercept)                        5.406988
## campaign                          -0.069743
## pdays                             0.002901
## poutcomePoutcome_nonexistent      0.009651
## poutcomePoutcome_success          0.378327
## monthMonth_aug                    -0.212340
## monthMonth_dec                     0.141391
## monthMonth_jul                    -0.187828
## monthMonth_jun                    -0.351201
## monthMonth_mar                    -0.185593
## monthMonth_may                    -0.345035
## monthMonth_nov                    -0.269914
## monthMonth_oct                    -0.228642
## monthMonth_sep                    -0.352472
## factor_cons.price.idxfactor_cons.price.idx-(93,93.4] -0.110456
## factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9] 0.088951
## factor_cons.price.idxfactor_cons.price.idx-(93.9,94]  0.219283
## factor_cons.price.idxfactor_cons.price.idx-(94,94.8]  0.002831
## factor_Previousfactor_Previous-(1,5] 0.188940
## day_of_weekDay_of_week_mon        0.060226
```

## day_of_weekDay_of_week_thu	0.085789	
## day_of_weekDay_of_week_tue	0.211005	
## day_of_weekDay_of_week_wed	0.150490	
##	Std. Error	t value
## (Intercept)	0.306736	17.627
## campaign	0.006710	-10.393
## pdays	0.018666	0.155
## poutcomePoutcome_nonexistent	0.049726	0.194
## poutcomePoutcome_success	0.207580	1.823
## monthMonth_aug	0.066472	-3.194
## monthMonth_dec	0.214603	0.659
## monthMonth_jul	0.114380	-1.642
## monthMonth_jun	0.105853	-3.318
## monthMonth_mar	0.130310	-1.424
## monthMonth_may	0.092767	-3.719
## monthMonth_nov	0.069135	-3.904
## monthMonth_oct	0.130712	-1.749
## monthMonth_sep	0.140611	-2.507
## factor_cons.price.idxfactor_cons.price.idx-(93,93.4]	0.070455	-1.568
## factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9]	0.096588	0.921
## factor_cons.price.idxfactor_cons.price.idx-(93.9,94]	0.049133	4.463
## factor_cons.price.idxfactor_cons.price.idx-(94,94.8]	0.074668	0.038
## factor_Previousfactor_Previous-(1,5]	0.098283	1.922
## day_of_weekDay_of_week_mon	0.041383	1.455
## day_of_weekDay_of_week_thu	0.041253	2.080
## day_of_weekDay_of_week_tue	0.042899	4.919
## day_of_weekDay_of_week_wed	0.041820	3.598
##	Pr(> t)	
## (Intercept)	< 2e-16	***
## campaign	< 2e-16	***
## pdays	0.876480	
## poutcomePoutcome_nonexistent	0.846126	
## poutcomePoutcome_success	0.068431	.
## monthMonth_aug	0.001410	**
## monthMonth_dec	0.510022	
## monthMonth_jul	0.100625	
## monthMonth_jun	0.000914	***
## monthMonth_mar	0.154438	
## monthMonth_may	0.000202	***
## monthMonth_nov	9.58e-05	***
## monthMonth_oct	0.080316	.
## monthMonth_sep	0.012218	*
## factor_cons.price.idxfactor_cons.price.idx-(93,93.4]	0.117007	
## factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9]	0.357133	
## factor_cons.price.idxfactor_cons.price.idx-(93.9,94]	8.26e-06	***

```
## factor_cons.price.idxfactor_cons.price.idx-(94,94.8]    0.969754
## factor_Previousfactor_Previous-(1,5]                  0.054612 .
## day_of_weekDay_of_week_mon                             0.145640
## day_of_weekDay_of_week_thu                             0.037615 *
## day_of_weekDay_of_week_tue                             9.00e-07 ***
## day_of_weekDay_of_week_wed                             0.000323 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9212 on 4923 degrees of freedom
## Multiple R-squared:  0.05104,    Adjusted R-squared:  0.0468
## F-statistic: 12.03 on 22 and 4923 DF,  p-value: < 2.2e-16
```

#Com no ha sortit cap NA, de moment no tenim cap variable problemàtica!

Anova (m11)

```
## Anova Table (Type II tests)
##
## Response: log(duration)
##
##              Sum Sq   Df  F value    Pr(>F)
## campaign           91.7    1 108.0209 < 2.2e-16 ***
## pdays              0.0    1   0.0242  0.876480
## poutcome            2.8    2   1.6624  0.189794
## month              22.6    9   2.9525  0.001679 **
## factor_cons.price.idx 20.6    4   6.0598 7.335e-05 ***
## factor_Previous       3.1    1   3.6957  0.054612 .
## day_of_week          24.8    4   7.3018 7.367e-06 ***
## Residuals          4177.9 4923
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#Para limpiar! Efectes nets

#Poutcome és problemàtica perquè es 0.1 i les demás veiem que sí que són significatives!

A partir d'executar Anova(m11) podem veure quines són les variables significatives llavors agafem el nou model, que el que li hem tret és la variables poutcome i factor_Previous(encara que aquesta última es podria agafar també com a significativa, perquè hi ha un .).

Ara quan tenim el nostre model m8 amb els factors significatius corresponents el que hem de fer és veure si les nostres variables numèriques inicials del nostre model són més explicatives com a numèriques o com a factors.


```

#Our model
m12<-
lm(log(duration)~campaign+pdays+poutcome+month+factor_cons.price.idx+day_of_w
eek,data = df)
summary(m12)

##
## Call:
## lm(formula = log(duration) ~ campaign + pdays + poutcome + month +
##     factor_cons.price.idx + day_of_week, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.2483 -0.5570 -0.0058  0.6015  2.6707
##
## Coefficients:
##                                     Estimate
## (Intercept)                        5.531569
## campaign                          -0.069960
## pdays                             -0.003735
## poutcomePoutcome_nonexistent      -0.013441
## poutcomePoutcome_success           0.350904
## monthMonth_aug                     -0.208718
## monthMonth_dec                      0.163868
## monthMonth_jul                     -0.193449
## monthMonth_jun                     -0.370057
## monthMonth_mar                     -0.185277
## monthMonth_may                     -0.343337
## monthMonth_nov                     -0.268959
## monthMonth_oct                     -0.219786
## monthMonth_sep                     -0.336518
## factor_cons.price.idxfactor_cons.price.idx-(93,93.4] -0.110291
## factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9] 0.099605
## factor_cons.price.idxfactor_cons.price.idx-(93.9,94]  0.221876
## factor_cons.price.idxfactor_cons.price.idx-(94,94.8]  0.030606
## day_of_weekDay_of_week_mon         0.060586
## day_of_weekDay_of_week_thu         0.086819
## day_of_weekDay_of_week_tue         0.212060
## day_of_weekDay_of_week_wed         0.152392
##                                     Std. Error t value
## (Intercept)                        0.299894  18.445
## campaign                          0.006711 -10.424
## pdays                             0.018349  -0.204
## poutcomePoutcome_nonexistent      0.048267  -0.278
## poutcomePoutcome_success           0.207146   1.694
## monthMonth_aug                     0.066463  -3.140

```

```

## monthMonth_dec                0.214343    0.765
## monthMonth_jul                0.114374   -1.691
## monthMonth_jun                0.105427   -3.510
## monthMonth_mar                0.130345   -1.421
## monthMonth_may                0.092788   -3.700
## monthMonth_nov                0.069152   -3.889
## monthMonth_oct                0.130666   -1.682
## monthMonth_sep                0.140404   -2.397
## factor_cons.price.idxfactor_cons.price.idx-(93,93.4] 0.070475   -1.565
## factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9] 0.096455    1.033
## factor_cons.price.idxfactor_cons.price.idx-(93.9,94]  0.049128    4.516
## factor_cons.price.idxfactor_cons.price.idx-(94,94.8]  0.073276    0.418
## day_of_weekDay_of_week_mon    0.041394    1.464
## day_of_weekDay_of_week_thu    0.041260    2.104
## day_of_weekDay_of_week_tue    0.042907    4.942
## day_of_weekDay_of_week_wed    0.041820    3.644
##                               Pr(>|t|)
## (Intercept)                   < 2e-16 ***
## campaign                       < 2e-16 ***
## pdays                        0.838711
## poutcomePoutcome_nonexistent 0.780664
## poutcomePoutcome_success     0.090330 .
## monthMonth_aug               0.001697 **
## monthMonth_dec               0.444597
## monthMonth_jul               0.090828 .
## monthMonth_jun               0.000452 ***
## monthMonth_mar               0.155254
## monthMonth_may               0.000218 ***
## monthMonth_nov               0.000102 ***
## monthMonth_oct               0.092623 .
## monthMonth_sep               0.016577 *
## factor_cons.price.idxfactor_cons.price.idx-(93,93.4] 0.117652
## factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9] 0.301815
## factor_cons.price.idxfactor_cons.price.idx-(93.9,94]  6.44e-06 ***
## factor_cons.price.idxfactor_cons.price.idx-(94,94.8]  0.676204
## day_of_weekDay_of_week_mon    0.143350
## day_of_weekDay_of_week_thu    0.035414 *
## day_of_weekDay_of_week_tue    7.98e-07 ***
## day_of_weekDay_of_week_wed    0.000271 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9215 on 4924 degrees of freedom
## Multiple R-squared:  0.05032,    Adjusted R-squared:  0.04627
## F-statistic: 12.43 on 21 and 4924 DF,  p-value: < 2.2e-16

```

```
#Estudi de campaign
```

```
# Decide wether campaign should be considered either numeric, or factor  
(never both)
```

```
maux<-
```

```
lm(log(duration)~factor_campaign+pdays+month+factor_cons.price.idx+day_of_week, data = df)
```

```
BIC(m12,maux) #Choose option with minimum BIC
```

```
##      df      BIC
```

```
## m12  23 13400.74
```

```
## maux 22 13420.62
```

```
#El BIC més petit es el recomanable
```

```
#La variable campaign numèrica m'explica més que factor_campaign perquè el  
BIC de m12 és més petit que el de maux
```

```
# Estudi de pdays
```

```
maux2<-
```

```
lm(log(duration)~campaign+factor_Pdays+poutcome+month+factor_cons.price.idx+day_of_week, data = df)
```

```
BIC(m12,maux2) #Choose option with minimum BIC, for me pdays as numeric is  
not an option
```

```
##      df      BIC
```

```
## m12  23 13400.74
```

```
## maux2 23 13395.80
```

```
#El factor_Pdays m'explica més que la variable numèrica pdays perquè el BIC  
de maux2 és més petit que el de m12
```

```
maux3<-
```

```
lm(log(duration)~factor_campaign+factor_Pdays+poutcome+month+factor_cons.price.idx+day_of_week, data = df)
```

```
BIC(m12,maux3)
```

```
##      df      BIC
```

```
## m12  23 13400.74
```

```
## maux3 24 13429.43
```

```
#Hi ha una millor explicabilitat en el maux2!
```

```
#Best solution:
```

```
m13<-
```

```
lm(log(duration)~campaign+factor_Pdays+poutcome+month+factor_cons.price.idx+day_of_week, data = df)
```

Després del nostre estudi, el que podem veure o les conclusions que podem treure és que les nostres variables numèriques del model inical, campaign i pdays, és que campaign és més explicativa sent numèrica mentre que la variable pdays és més explicativa quan s'utilitza com a factor i això es pot comprovar amb la comanda “BIC”.

És pot veure com en maux3 tenim un BIC més petit que en el nostre model m12, però si comprovem tots els models auxiliar veiem que el BIC més petit és el que ens dona el model maux2.

```
#Try to combine both criteria
Anova(m13) #Check significant variables

## Anova Table (Type II tests)
##
## Response: log(duration)
##
```

	Sum Sq	Df	F value	Pr(>F)	
campaign	91.8	1	108.2467	< 2.2e-16	***
factor_Pdays	4.2	1	4.9628	0.025943	*
poutcome	0.2	2	0.1296	0.878431	
month	22.5	9	2.9462	0.001715	**
factor_cons.price.idx	20.6	4	6.0794	7.075e-05	***
day_of_week	25.6	4	7.5441	4.692e-06	***
Residuals	4176.8	4924			

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

m14<-step(m13,k=log(nrow(df))) #I prioritize BIC criteria

## Start:  AIC=-648.84
## log(duration) ~ campaign + factor_Pdays + poutcome + month +
##      factor_cons.price.idx + day_of_week
##
##
```

	Df	Sum of Sq	RSS	AIC
- month	9	22.492	4199.3	-698.84
- poutcome	2	0.220	4177.1	-665.60
- factor_cons.price.idx	4	20.628	4197.5	-658.50
- day_of_week	4	25.597	4202.4	-652.65
- factor_Pdays	1	4.210	4181.1	-652.37
<none>			4176.8	-648.84
- campaign	1	91.822	4268.7	-549.80

```
##
## Step:  AIC=-698.84
## log(duration) ~ campaign + factor_Pdays + poutcome + factor_cons.price.idx
## +
##      day_of_week
```

```
##
##              Df Sum of Sq    RSS    AIC
## - poutcome      2      0.401 4199.7 -715.38
## - day_of_week    4     22.889 4222.2 -705.98
## - factor_Pdays   1      5.071 4204.4 -701.38
## <none>                                4199.3 -698.84
## - factor_cons.price.idx  4     43.631 4243.0 -681.74
## - campaign        1     94.896 4294.2 -596.82
##
## Step:   AIC=-715.38
## log(duration) ~ campaign + factor_Pdays + factor_cons.price.idx +
##   day_of_week
##
##              Df Sum of Sq    RSS    AIC
## - day_of_week    4     22.803 4222.5 -722.62
## <none>                                4199.7 -715.38
## - factor_cons.price.idx  4     45.083 4244.8 -696.59
## - factor_Pdays         1     39.056 4238.8 -678.10
## - campaign              1     95.751 4295.5 -612.39
##
## Step:   AIC=-722.62
## log(duration) ~ campaign + factor_Pdays + factor_cons.price.idx
##
##              Df Sum of Sq    RSS    AIC
## <none>                                4222.5 -722.62
## - factor_cons.price.idx  4     48.066 4270.6 -700.66
## - factor_Pdays         1     40.106 4262.7 -684.37
## - campaign              1    100.169 4322.7 -615.17

summary(m14)

##
## Call:
## lm(formula = log(duration) ~ campaign + factor_Pdays +
##   factor_cons.price.idx,
##   data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.1686 -0.5522 -0.0012  0.6094  2.6940
##
## Coefficients:
##              Estimate
## (Intercept)      5.746773
## campaign        -0.072224
## factor_Pdaysfactor_Pdays-(15,17] -0.491280
```

```

## factor_cons.price.idxfactor_cons.price.idx-(93,93.4]    0.004904
## factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9]  0.219195
## factor_cons.price.idxfactor_cons.price.idx-(93.9,94]    0.189446
## factor_cons.price.idxfactor_cons.price.idx-(94,94.8]    -0.014655
##                                                         Std. Error t value
## (Intercept)                                             0.072690  79.059
## campaign                                                0.006672 -10.824
## factor_Pdaysfactor_Pdays-(15,17]                     0.071729  -6.849
## factor_cons.price.idxfactor_cons.price.idx-(93,93.4]    0.038153   0.129
## factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9]  0.042427   5.166
## factor_cons.price.idxfactor_cons.price.idx-(93.9,94]    0.042045   4.506
## factor_cons.price.idxfactor_cons.price.idx-(94,94.8]    0.044780  -0.327
##                                                         Pr(>|t|)
## (Intercept)                                           < 2e-16 ***
## campaign                                              < 2e-16 ***
## factor_Pdaysfactor_Pdays-(15,17]                   8.34e-12 ***
## factor_cons.price.idxfactor_cons.price.idx-(93,93.4]    0.898
## factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9]  2.48e-07 ***
## factor_cons.price.idxfactor_cons.price.idx-(93.9,94]    6.76e-06 ***
## factor_cons.price.idxfactor_cons.price.idx-(94,94.8]    0.743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9246 on 4939 degrees of freedom
## Multiple R-squared:  0.04089,    Adjusted R-squared:  0.03973
## F-statistic: 35.1 on 6 and 4939 DF,  p-value: < 2.2e-16

#No tenim NA! -> PERFECTE!

#Anova(m13)

#m15<-
lm(log(duration)~campaign+factor_Pdays+factor_cons.price.idx+day_of_week,data
= df)
#summary(m15)
#Anova(m15)

#Ara volem saber els nivells que tenim
summary(df[,c("campaign", "factor_Pdays","factor_cons.price.idx")])

##      campaign                factor_Pdays
## Min.      : 1.000    factor_Pdays-[0,15] : 179
## 1st Qu.: 1.000    factor_Pdays-(15,17]:4767
## Median : 2.000
## Mean      : 2.389
## 3rd Qu.: 3.000

```

```
## Max.      :14.000
##                               factor_cons.price.idx
## factor_cons.price.idx-[92.2,93]  :1059
## factor_cons.price.idx-(93,93.4]  :1359
## factor_cons.price.idx-(93.4,93.9]: 889
## factor_cons.price.idx-(93.9,94]   : 921
## factor_cons.price.idx-(94,94.8]   : 718
##
```

```
#model.matrix(m14)
```

Per aconseguir la nostra matriu he agafat les variables més significatives que m'ha donat la comanda “step”, podem agafar també a partir de fer l'Anova del nostre model final que teníem fins el moment, però hem decidit agafar el model m14 per averiguar els nivells que tenim. Fent l'Anova tenim el model m15 que també posaria en el summary les variables “month” i “day_of_week”, mentre que el model m14 ens dóna les variables que tenim en el summary. (Era correcte agafar qualsevol de les dues opcions).

Després de tot l'estudi hem vist que nosaltres hem fet un model i un estudi Variable Numèrica VS. Factor Mai es pot donar una interacció entre dos variables numèriques!

```
##Interaction: order 2 no more
```

```
m15<-lm(log(duration)~(campaign+factor_Pdays+factor_cons.price.idx)^2,data = df)
```

```
#summary(m15)
```

```
#coef(m15)
```

```
m16<-step(m15,k=log(nrow(df)))
```

```
## Start: AIC=-726.41
```

```
## log(duration) ~ (campaign + factor_Pdays + factor_cons.price.idx)^2
```

```
##
```

```
##
```

	Df	Sum of Sq	RSS	AIC
## - factor_Pdays:factor_cons.price.idx	3	2.215	4163.9	-749.30
## - campaign:factor_Pdays	1	0.356	4162.0	-734.50
## <none>			4161.7	-726.41
## - campaign:factor_cons.price.idx	4	58.796	4220.5	-691.05

```
##
```

```
## Step: AIC=-749.3      24
```

```
## log(duration) ~ campaign + factor_Pdays + factor_cons.price.idx +
```

```
##   campaign:factor_Pdays + campaign:factor_cons.price.idx
```

```
##
```

```
##
```

	Df	Sum of Sq	RSS	AIC
## - campaign:factor_Pdays	1	0.454	4164.3	-757.27

```
## <none> 4163.9 -749.30
## - campaign:factor_cons.price.idx 4 58.630 4222.5 -714.17
##
## Step: AIC=-757.27
## log(duration) ~ campaign + factor_Pdays + factor_cons.price.idx +
## campaign:factor_cons.price.idx
##
## Df Sum of Sq RSS AIC
## <none> 4164.3 -757.27
## - campaign:factor_cons.price.idx 4 58.222 4222.5 -722.62
## - factor_Pdays 1 36.552 4200.9 -722.55
```

```
#Anova(m16)
```

```
anova(m16,m15) #Fisher test - Priority to BIC criteria
```

```
## Analysis of Variance Table
```

```
##
```

```
## Model 1: log(duration) ~ campaign + factor_Pdays + factor_cons.price.idx +
## campaign:factor_cons.price.idx
```

```
## Model 2: log(duration) ~ (campaign + factor_Pdays +
## factor_cons.price.idx)^2
```

```
## Res.Df RSS Df Sum of Sq F Pr(>F)
```

```
## 1 4935 4164.3
```

```
## 2 4931 4161.7 4 2.6684 0.7904 0.5312
```

```
#Prioritzo el criteri step per agafar les redundants
```

Després d'aquesta execució podem veure segons el criteri de Fisher que els dos models no són equivalents, i això ho podem saber mirant el p-valor i és molt petit!

Interactions between numeric variables and factors

Model Additiu

```
#Exemple adhoc: Y ~ X+A
```

```
m17<-lm(log(duration)~campaign+factor_Pdays,data = df)
```

```
summary(m17)
```

```
##
```

```
## Call:
```

```
## lm(formula = log(duration) ~ campaign + factor_Pdays, data = df)
```

```
##
```

```
## Residuals:
```

```
## Min 1Q Median 3Q Max
```

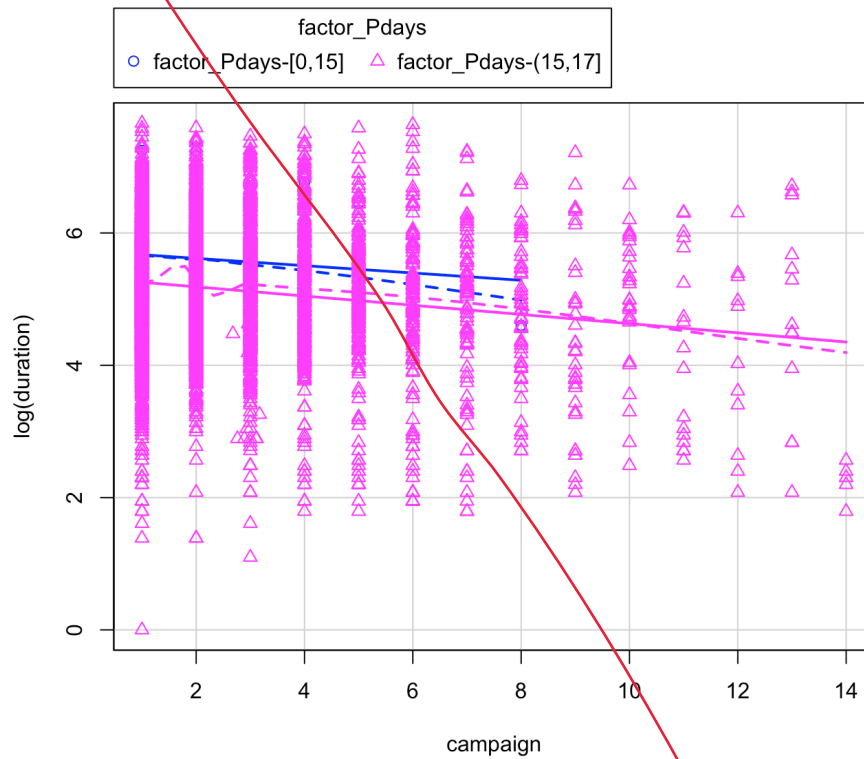
```
## -5.2555 -0.5417 0.0013 0.6222 2.7306
```



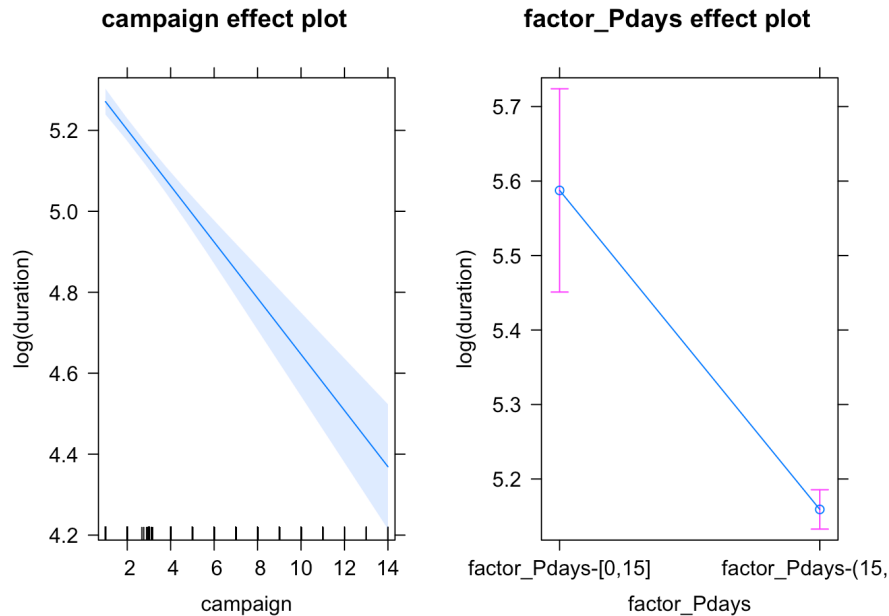
```
##
## Coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.753204   0.070467  81.644 < 2e-16 ***
## campaign         -0.069384   0.006676 -10.394 < 2e-16 ***
## factor_Pdaysfactor_Pdays-(15,17] -0.428324   0.070898  -6.041 1.64e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9295 on 4943 degrees of freedom
## Multiple R-squared:  0.02997,    Adjusted R-squared:  0.02958
## F-statistic: 76.37 on 2 and 4943 DF,  p-value: < 2.2e-16
```

#Suport visual

```
scatterplot(log(duration)~campaign|factor_Pdays,data=df)
```



```
#Interpretation of models through effects library
library(effects)
plot(allEffects(m17))
```



A l'eix de les ordenades tenim el logaritme de “duration” (eix vertical), campaign en aquest cas augmenta, és a dir, el número de campanyes implica una disminució en el logaritme de la durada = efecte negatiu Però el factor_Pdays calcula un valor de confiança segons els intervals que tenim i d'aquesta manera ens ajuda a interpretar el que tenim com a sortida

Lavors ara és hora de interpretar el nostre model: $Y \sim X + A$ i = 1 (que és equivalent al factor_Pdays[0,15]) $Y_i = Y_1 = 5.75 - 0.069X$ i = 2 (que és equivalent al factor_Pdays[15,17]) $Y_i = Y_2 = (5.75 - 0.428) - 0.069X$

Model Interaccions

```
# Y ~ X*A (que és equivalent a X+A+A:X)
m18<-lm(log(duration)~campaign*factor_Pdays,data = df) #Concepte d'interacció ara
summary(m18)

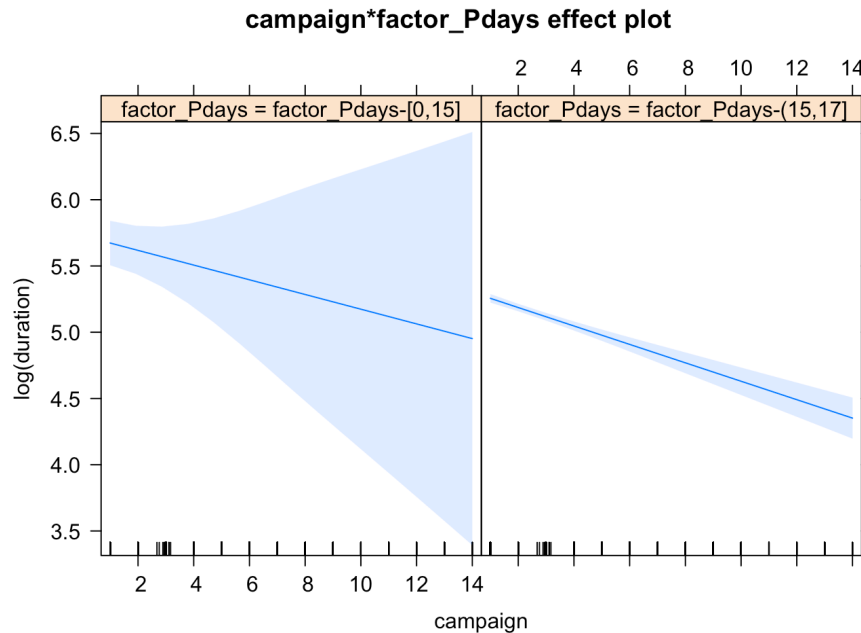
##
## Call:
## lm(formula = log(duration) ~ campaign * factor_Pdays, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.2557 -0.5418  0.0014  0.6220  2.7311
##
## Coefficients:
##              Estimate Std. Error t value
## (Intercept)      5.72867    0.13376  42.828
## campaign        -0.05549    0.06474  -0.857
## factor_Pdaysfactor_Pdays-(15,17] -0.40343    0.13541  -2.979
## campaign:factor_Pdaysfactor_Pdays-(15,17] -0.01405    0.06509  -0.216
##              Pr(>|t|)
## (Intercept)      <2e-16 ***
## campaign          0.3915
## factor_Pdaysfactor_Pdays-(15,17]  0.0029 **
## campaign:factor_Pdaysfactor_Pdays-(15,17]  0.8291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9296 on 4942 degrees of freedom
## Multiple R-squared:  0.02998,    Adjusted R-squared:  0.0294
## F-statistic: 50.92 on 3 and 4942 DF,  p-value: < 2.2e-16

# Las interacciones son rellevants?
anova(m17,m18)

## Analysis of Variance Table
##
## Model 1: log(duration) ~ campaign + factor_Pdays
## Model 2: log(duration) ~ campaign * factor_Pdays
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     4943 4270.6
## 2     4942 4270.6   1   0.040249 0.0466 0.8291

#pvalue << 0.05 -> H0 Rejected -> m18 X*A
#anova(petit, gran)
```

```
plot(allEffects(m18))
```



Hi han moltes observacions influents per això hi ha tanta zona blau clar, per l'interval de confiança que tenim!

També el que hem pogut comprobar és si les nostres interaccions són rellevants i amb la comanda “anova” fem com unaména de comparació per veure els dos models que tenim i poder treure com a conclusió que haure d’acceptar la hipòtesi nula, perquè el pvalor que surt és més gran que 0.05 (5%).

Ara és hora d’interpretar el nostre model: $Y \sim X * A$ i = 1 (que és equivalent al factor_Pdays[0,15]) $Y_i = Y_1 = 5.73 - 0.055X$ i = 2 (que és equivalent al factor_Pdays[15,17]) $Y_i = Y_2 = (5.73 - 0.403) + (-0.055 - 0.014)X$



Binary Regression

Explanatory numeric variables

Initial modelling

El que farem al començament de tot és dividir la modelització inicial (que tenim fins ara) en mostres de treball i mostres per testejar. En aquest apartat trobarem el “Eta2”, que no el podem interpretar del tot bé ja que s'utilitza més en el MCA i no l'hem pogut fer a classe, però és com un coeficient de determinació quan tenim variables involucrades que són factors. A l'hora d'escollir el nostre millor model, és bona tècnica agafar com a referència també el “Estimate” que ens dóna el pes que se li dóna a cada variable en el model, llavors veiem quines són les més explicatives. I finalment, el “z value” és una aproximació del “Estimate/Std.Error”, valors de la normal estàndard.

```
# Divide into work and test samples
```

```
set.seed(123)
```

```
sam<-sample(1:nrow(df),0.75*nrow(df)) #Random sample without replacement
```

```
dfw<-df[sam,]
```

```
dft<-df[-sam,]
```

```
# Numeric variables
```

```
vars_con
```

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

```
catdes(dfw[,c("y",vars_con)],1) #Numericas relacionadas
```

```
##
```

```
## Link between the cluster variable and the quantitative variables
```

```
## =====
```

##		Eta2	P-value
##	duration	0.17671414	9.254637e-159
##	nr.employed	0.14477732	4.417482e-128
##	pdays	0.13675760	1.481722e-120
##	euribor3m	0.10793163	4.600661e-94
##	emp.var.rate	0.09974083	1.089368e-86
##	previous	0.07808778	1.666707e-67
##	cons.price.idx	0.01621864	6.967791e-15

```

## campaign      0.00438049  5.487012e-05
##
## Description of each cluster by quantitative variables
## =====
## $Y_no
##          v.test Mean in category Overall mean sd in category
## nr.employed    23.169685      5177.7302797 5.167214e+03    64.7069872
## pdays         22.518818      15.8902551 1.559935e+01     1.1196236
## euribor3m      20.005261       3.8549862 3.641860e+00     1.6193552
## emp.var.rate   19.231198       0.2851214 9.937989e-02     1.4698800
## cons.price.idx  7.754916      93.6098528 9.358235e+01     0.5538129
## campaign       4.030243       2.4041326 2.356065e+00     1.9968564
## previous      -17.016154       0.1251153 1.763279e-01     0.4006136
## duration      -25.597969      223.6446357 2.640345e+02    203.6701199
##          Overall sd          p.value
## nr.employed    73.8222624 9.207180e-119
## pdays         2.1010235 2.715126e-112
## euribor3m      1.7326984 4.955848e-89
## emp.var.rate   1.5708408 2.028852e-82
## cons.price.idx  0.5767261 8.840227e-15
## campaign       1.9397909 5.571924e-05
## previous       0.4894910 6.233339e-65
## duration      256.6235243 1.607064e-144
##
## $Y_yes
##          v.test Mean in category Overall mean sd in category
## duration      25.597969      552.1666667 2.640345e+02    380.8900798
## previous      17.016154       0.5416667 1.763279e-01     0.8073244
## campaign      -4.030243       2.0131579 2.356065e+00     1.4234264
## cons.price.idx -7.754916      93.3861820 9.358235e+01     0.6881347
## emp.var.rate  -19.231198      -1.2256579 9.937989e-02     1.6296390
## euribor3m     -20.005261       2.1214627 3.641860e+00     1.7541244
## pdays        -22.518818      13.5241228 1.559935e+01     4.6959610
## nr.employed   -23.169685      5092.1901316 5.167214e+03    89.6674427
##          Overall sd          p.value
## duration      256.6235243 1.607064e-144
## previous      0.4894910 6.233339e-65
## campaign      1.9397909 5.571924e-05
## cons.price.idx 0.5767261 8.840227e-15
## emp.var.rate   1.5708408 2.028852e-82
## euribor3m      1.7326984 4.955848e-89
## pdays         2.1010235 2.715126e-112
## nr.employed    73.8222624 9.207180e-119

```

```

# EXEMPLE!
# Model NULL, només tenim una constant
# gm0<-glm(y~1,family=binomial,data = dfw)
# summary(gm0)

# binomial = distribucion que le damos a la variable de respuesta
# Si volem podem utilitzar duration, sino no, si es posa és com fer una mica
de trampa, no té sentit utilitzar-la com a variable explicativa, però si
volem és pot utilitzar.
gm1<-
glm(y~nr.employed+pdays+euribor3m+emp.var.rate+previous+cons.price.idx+campai
gn,family=binomial,data = dfw)
# summary(gm1)
# Anova(gm1) #Test efectes nets
vif(gm1)

##      nr.employed      pdays      euribor3m      emp.var.rate      previous
##      16.957527      1.416024      24.098435      31.623083      1.692257
## cons.price.idx      campaign
##      7.702834      1.027985

#Saca los problemas de col·linealitat!
#Més gran que 3 SON DOLENTES!

#Remove colinear variables
#Es treuran per separat i la que canviï menys el model s'agafa fins que
siguin quasi totes significatives
gm2<-
glm(y~nr.employed+pdays+euribor3m+previous+cons.price.idx+campaign,family=bin
omial,data = dfw)
# summary(gm2)
vif(gm2)

##      nr.employed      pdays      euribor3m      previous cons.price.idx
##      14.181816      1.417321      18.347138      1.684602      2.968792
##      campaign
##      1.022954

# Anova(gm2)

# gm3<-
glm(y~nr.employed+pdays+previous+cons.price.idx+campaign,family=binomial,data
= dfw)
# summary(gm3)
# vif(gm3)
# Anova(gm3)

```

```
gm4<-glm(y~pdays+previous+cons.price.idx+campaign,family=binomial,data = dfw)
summary(gm4)
```

```
##
## Call:
## glm(formula = y ~ pdays + previous + cons.price.idx + campaign,
##      family = binomial, data = dfw)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.2876  -0.4763  -0.4141  -0.3734   2.5103
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   44.22567    8.67333   5.099 3.41e-07 ***
## pdays        -0.23029    0.02344  -9.824 < 2e-16 ***
## previous       0.49007    0.10292   4.762 1.92e-06 ***
## cons.price.idx -0.45626    0.09254  -4.930 8.21e-07 ***
## campaign      -0.06844    0.03318  -2.063  0.0391 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 2765.1  on 3708  degrees of freedom
## Residual deviance: 2406.1  on 3704  degrees of freedom
## AIC: 2416.1
##
## Number of Fisher Scoring iterations: 5
```

```
vif(gm4)
```

```
##           pdays           previous cons.price.idx           campaign
##           1.366062           1.394791           1.023703           1.015790
```

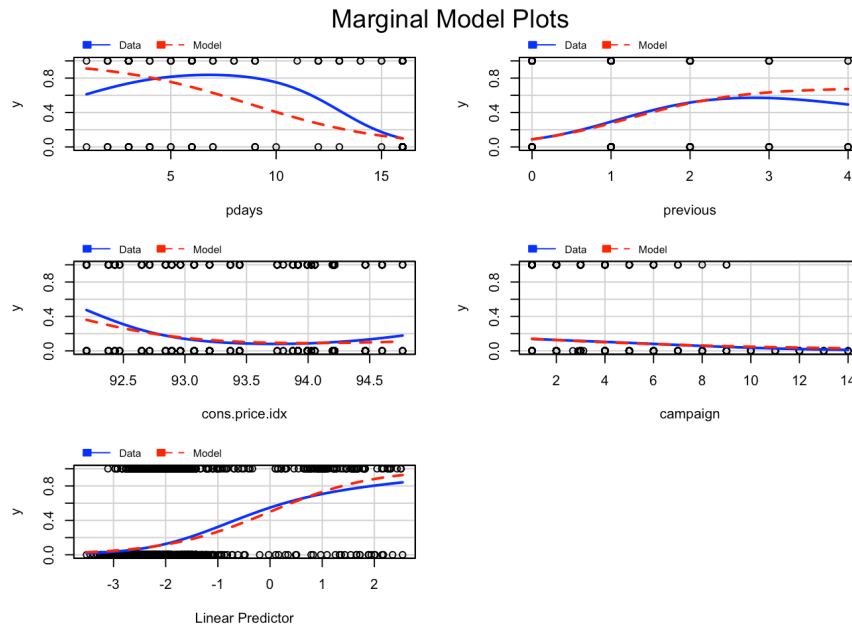
```
Anova(gm4)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: y
##              LR Chisq Df Pr(>Chisq)
## pdays          120.636  1 < 2.2e-16 ***
## previous         20.643  1 5.535e-06 ***
## cons.price.idx   24.457  1 7.600e-07 ***
## campaign         4.603  1  0.03192 *
```



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

marginalModelPlots(gm4) # Some missfit data vs model
```



Ara el que hem fet ha sigut trobar el nostre millor model lineas generalitzat i el que hem fet per aconseguir-ho ha sigut que a partir d'una mostra aleatòria hem anat elaborant els nostres models i amb la comanda “vif” hem anat treient els problemes de col·linealitat, és a dir, les variables que tenien un $vif > 3$ s'han de treure i anar probant diferents models amb les variables corresponents fins arribar a tenir un model on totes les nostres variables són significatives, però no hi ha cap estratègia òptima per dur a terme aquestes comprovacions.

Hem aconseguit disminuir la discrepància amb el nostre últim model (Residual deviance < Null deviance) i també es pot considerar correcte ja que Grau de llibertat = Num. observacions (3709) - Num. variables (5) = 3704 i una altra manera de veure que anem bé és que la Residual deviance és igual o inferior als graus de llibertat (2232.7 < 3704).

Com podem veure en les nostres transformacions, al model gm3 li hem tret la variable “euribor3m” respecte al model gm2 perquè segons el vif era una variable que afectava molt a la variança, però quan executàvem Anova hem vist que hi havien dos variables que no eren significatives, llavors hem optat per treure la variable “nr.employed” (Que en el model gm2 també sortia amb el vif elevat) que és el nostre model gm4 i ara quan executem Anova(gm4) podem veure que totes les variables implicades en el model són significatives, que és el que buscàvem.

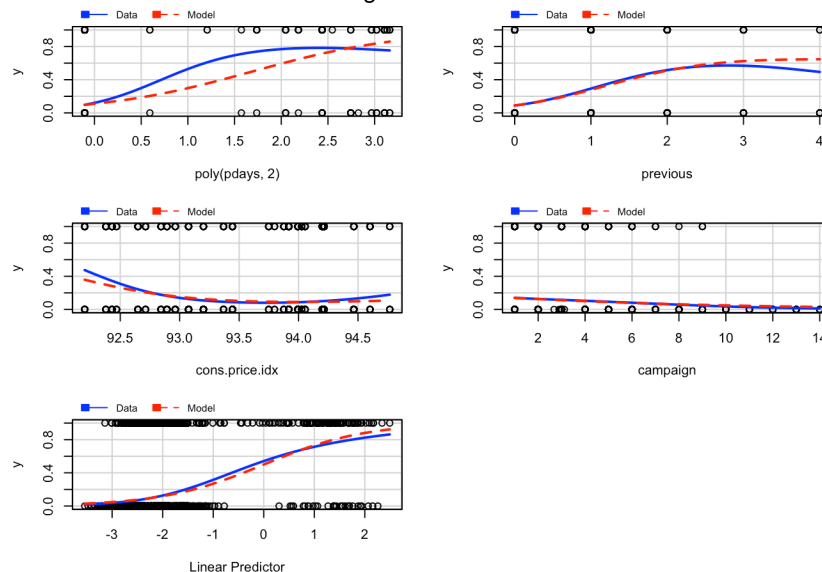
Transforming variables

El que farem a continuació és a partir del `marginalPlots` podem veure on hi ha un desajust entre les observacions i la predicció, llavors hem de trobar la manera d’arreglar-ho:

```
gm5<-glm(y~poly(pdays,
2)+previous+cons.price.idx+campaign,family=binomial,data = dfw)
# summary(gm5)
# Anova(gm5)
marginalModelPlots(gm5)
```

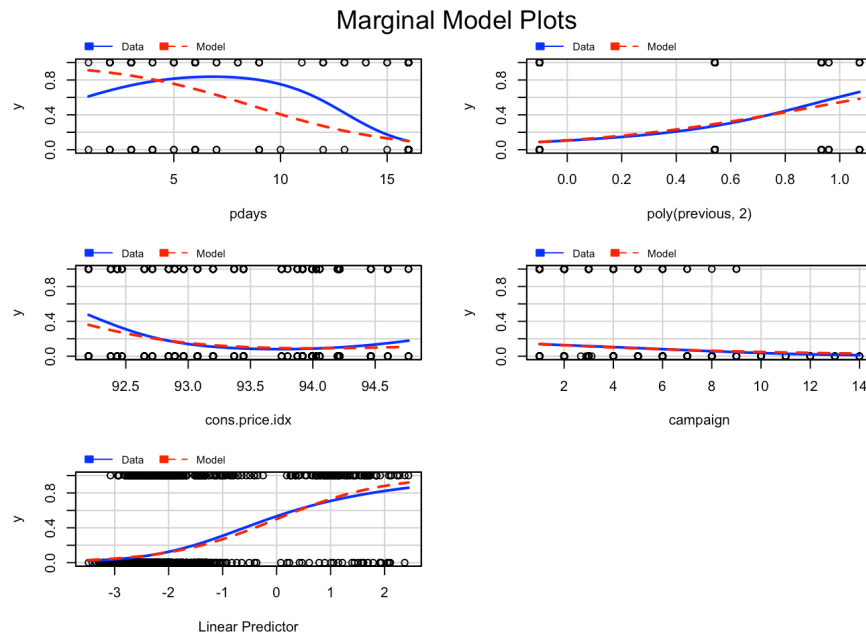
```
## Warning in mmps(...): Splines and/or polynomials replaced by a fitted
## linear combination
```

Marginal Model Plots



```
gm6<-glm(y~pdays+poly(previous,
2)+cons.price.idx+campaign,family=binomial,data = dfw)
marginalModelPlots(gm6)
```

Warning in mmps(...): Splines and/or polynomials replaced by a fitted
linear combination



Després de fer les comprovacions aplicant el quadràtic, veiem que en la variable pdays no canvia, sino que provoca un desajust més gran, després era hora de provar-ho amb previous i amb aquesta variable si que hi ha hagut una mica de millora, amb les variables que no són numèriques no fa falta fer-ho perquè mai sortirà res al marginalModelPlots. Llavors el model que ens quedarem serà el gm6 que és el que té menor desajust entre les observacions i les prediccions fetes.

Adding Factors

Seguidament el que hem de fer és agafar el nostre millor model des del punt anterior i introduïm els factors. El que s'ha de fer és anar probant totes les variables numèriques del nostre model fins ara com a factors i llavors ens quedem amb la que més t'expliqui segons ens indiqui el BIC.

```

gm10<-glm(y~pdays+poly(previous,
2)+cons.price.idx+campaign,family=binomial,data = dfw)

# First step: Choose between numeric explanatory variable or factor
# Check for all numerical variables: one by one

# Pdays: covariate or factor??
gm10a<-
glm(y~factor_Pdays+previous+cons.price.idx+campaign,family=binomial,data =
dfw)
BIC(gm10,gm10a)

##          df          BIC
## gm10      6 2453.155
## gm10a     5 2421.241

```

Explica més com a factor que com a numèrica! (BIC gm10a < BIC gm10)
L'ordre pot modificar els resultats pero no es pot fer res

Previous?

```

gm10b<-
glm(y~factor_Pdays+factor_Previous+cons.price.idx+campaign,family=binomial,da
ta = dfw)
BIC(gm10,gm10b)

##          df          BIC
## gm10      6 2453.155
## gm10b     5 2418.271

```

Explica més com a factor que com a numèrica! (BIC gm10b < BIC gm10)

Cons.price.idx?

```

gm10c<-
glm(y~factor_Pdays+factor_Previous+factor_cons.price.idx+campaign,family=bino
mial,data = dfw)
BIC(gm10,gm10c)

##          df          BIC
## gm10      6 2453.155
## gm10c     8 2394.856

```

Explica més com a factor que com a numèrica! (BIC gm10c < BIC gm10)

Campaign?

```

gm10d<-
glm(y~factor_Pdays+factor_Previous+factor_cons.price.idx+factor_campaign,fami
ly=binomial,data = dfw)

```

```

BIC(gm10,gm10d)

##          df          BIC
## gm10      6 2453.155
## gm10d     9 2406.311

# Explica més com a factor que com a numèrica! (BIC gm10d < BIC gm10)

## MILLOR MODEL FINS ARA:
gm11<-
glm(y~factor_Pdays+factor_Previous+factor_cons.price.idx+factor_campaign,fami
ly=binomial,data = dfw)

```

Podem veure o arribar a la conclusió després dels resultats que totes les variables del nostre model ideal fins ara que és el gm10 expliquen més com a factors que com a variables numèriques.

Adding new factors

Ara a continuació el que farem serà després del nostre model elaborat fins ara (gm11), li afegirem les variables factors que surtin que són més explicatives al nostre model.

```

# Add to your best model all new factors that are significantly related "y"
according to catdes(). I assume gm10 as the best model at this point
vars_dis2

```

```

## [1] "job" "marital"
## [3] "education" "default"
## [5] "housing" "loan"
## [7] "contact" "month"
## [9] "day_of_week" "poutcome"
## [11] "season" "factor_age"
## [13] "factor_duration" "factor_campaign"
## [15] "factor_Pdays" "factor_Previous"
## [17] "factor_emp.var.rate" "factor_cons.price.idx"
## [19] "factor_cons.conf.idx" "factor_euribor3m"
## [21] "factor_nr.employed"

```

```
catdes(dfw[,c("y",vars_dis2)],1)
```

```

# No hem de repetir els factors que ja tenim fins al moment comprovats i això
s'ha de fer agafant el model estudiat anteriorment

```

```

gm12<-
glm(y~factor_Pdays+factor_Previous+factor_cons.price.idx+factor_campaign+pout
come+month+job+season+default+education,family=binomial,data = dfw)

```

```
# Anova(gml2)
# summary(gml2)

#Amb el summary(gml2) he vist que tinc NA a la meua vostra en la variable
factor "season" i per això també em surt error en l'execució del vif, perquè
tenia aquesta variable que no era molt redundant, llavors:

gml2a<-
glm(y~factor_Pdays+factor_Previous+factor_cons.price.idx+factor_campaign+pout
come+month+job+default+education,family=binomial,data = dfw)
Anova(gml2a) # Mirem les que ens interessen i les que no!

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##
##          LR Chisq Df Pr(>Chisq)
## factor_Pdays      1.112  1    0.29164
## factor_Previous     4.045  1    0.04430 *
## factor_cons.price.idx 57.732  4  8.686e-12 ***
## factor_campaign     1.580  2    0.45392
## poutcome           6.035  2    0.04892 *
## month              87.675  9  4.762e-15 ***
## job                12.743 11    0.31047
## default            6.003  1    0.01428 *
## education          7.193  6    0.30338
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

vif(gml2a)

##
##          GVIF Df GVIF^(1/(2*Df))
## factor_Pdays      9.527644  1      3.086688
## factor_Previous     1.560871  1      1.249348
## factor_cons.price.idx 31.904305  4      1.541634
## factor_campaign     1.055823  2      1.013673
## poutcome           11.555512  2      1.843730
## month              36.559308  9      1.221331
## job                 3.689568 11      1.061137
## default             1.089252  1      1.043672
## education           3.182190  6      1.101270

#A partir de l'Anova veiem que hi han variables factors no significatives,
que no ens aporten res al model, llavors les treiem:
```

```

gm12b<-
glm(y~factor_Previous+factor_cons.price.idx+poutcome+month+default,family=binomial,data = dfw)
Anova(gm12b)

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##
##          LR Chisq Df Pr(>Chisq)
## factor_Previous      7.266  1  0.007027 **
## factor_cons.price.idx 65.835  4  1.716e-13 ***
## poutcome            120.651  2  < 2.2e-16 ***
## month               109.822  9  < 2.2e-16 ***
## default              8.504  1  0.003543 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

vif(gm12b)

##
##          GVIF Df GVIF^(1/(2*Df))
## factor_Previous      1.351135  1      1.162383
## factor_cons.price.idx 28.887284  4      1.522609
## poutcome            1.521054  2      1.110545
## month               28.115574  9      1.203641
## default             1.035864  1      1.017774

gm13<-step(gm12b,k=log(nrow(dfw)))

## Start:  AIC=2354.17
## y ~ factor_Previous + factor_cons.price.idx + poutcome + month +
##      default
##
##          Df Deviance    AIC
## - factor_Previous      1  2213.5 2353.2
## <none>                  2206.2 2354.2
## - default              1  2214.7 2354.5
## - factor_cons.price.idx  4  2272.1 2387.1
## - month                 9  2316.1 2390.0
## - poutcome              2  2326.9 2458.4
##
## Step:  AIC=2353.22
## y ~ factor_cons.price.idx + poutcome + month + default
##
##          Df Deviance    AIC
## <none>                  2213.5 2353.2
## - default              1  2222.1 2353.6

```

```
## - factor_cons.price.idx 4 2278.1 2384.9
## - month 9 2327.5 2393.3
## - poutcome 2 2374.7 2498.0
```

```
#vif(gm13)
```

```
# END POINT: No colinearity, all net effects for factors and numeric
variables should be significant
# colinearity: Se mira con el vig, el apartado GVIF que sean < 3
```

Després de fer el procés de modelització introduint les millores pas a pas, hem pogut observar que el nostre millor model completat amb els factors que faltaven és el model gm12b, i també ho podem comprovar executant la comanda Anova i veiem com totes les variables factors són significatives. Un model també òptim i correcte seria el gm13, ja que aquest surt després d'executar la comanda “step” al model gm12b. (Seria correcte agafar qualsevol dels dos)



```
# Check your final model at this point: all coefficients should be available
in the summary(model)
```

```
summary(gm12b)
```

```
##
## Call:
## glm(formula = y ~ factor_Previous + factor_cons.price.idx + poutcome +
##      month + default, family = binomial, data = dfw)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3646  -0.4763  -0.3483  -0.2866   2.7158
##
## Coefficients:
##                                     Estimate Std. Error
## (Intercept)                        0.20017    0.29558
## factor_Previousfactor_Previous-(1,5] 0.79436    0.29289
## factor_cons.price.idxfactor_cons.price.idx-(93,93.4] -1.65895    0.23230
## factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9] -1.13814    0.31381
## factor_cons.price.idxfactor_cons.price.idx-(93.9,94] -1.08805    0.26039
## factor_cons.price.idxfactor_cons.price.idx-(94,94.8] -0.40926    0.23599
## poutcomePoutcome_nonexistent        -0.03669    0.17995
## poutcomePoutcome_success            2.47038    0.27019
## monthMonth_aug                      -1.32216    0.25693
## monthMonth_dec                      -0.30063    0.60409
```



```

## monthMonth_jul -1.29686 0.35683
## monthMonth_jun -1.87335 0.34855
## monthMonth_mar 0.07422 0.37630
## monthMonth_may -2.24742 0.31011
## monthMonth_nov -1.31315 0.26964
## monthMonth_oct -0.47742 0.38193
## monthMonth_sep -0.73219 0.41880
## defaultDefault_unknown -0.49048 0.17571
## z value Pr(>|z|)
## (Intercept) 0.677 0.498265
## factor_Previousfactor_Previous-(1,5] 2.712 0.006684
## factor_cons.price.idxfactor_cons.price.idx-(93,93.4] -7.142 9.23e-13
## factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9] -3.627 0.000287
## factor_cons.price.idxfactor_cons.price.idx-(93.9,94] -4.179 2.93e-05
## factor_cons.price.idxfactor_cons.price.idx-(94,94.8] -1.734 0.082873
## poutcomePoutcome_nonexistent -0.204 0.838448
## poutcomePoutcome_success 9.143 < 2e-16
## monthMonth_aug -5.146 2.66e-07
## monthMonth_dec -0.498 0.618717
## monthMonth_jul -3.634 0.000279
## monthMonth_jun -5.375 7.67e-08
## monthMonth_mar 0.197 0.843652
## monthMonth_may -7.247 4.25e-13
## monthMonth_nov -4.870 1.12e-06
## monthMonth_oct -1.250 0.211293
## monthMonth_sep -1.748 0.080411
## defaultDefault_unknown -2.791 0.005248
##
## (Intercept)
## factor_Previousfactor_Previous-(1,5] **
## factor_cons.price.idxfactor_cons.price.idx-(93,93.4] ***
## factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9] ***
## factor_cons.price.idxfactor_cons.price.idx-(93.9,94] ***
## factor_cons.price.idxfactor_cons.price.idx-(94,94.8] .
## poutcomePoutcome_nonexistent
## poutcomePoutcome_success ***
## monthMonth_aug ***
## monthMonth_dec
## monthMonth_jul ***
## monthMonth_jun ***
## monthMonth_mar
## monthMonth_may ***
## monthMonth_nov ***
## monthMonth_oct
## monthMonth_sep .

```

```
## defaultDefault_unknown **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 2765.1  on 3708  degrees of freedom
## Residual deviance: 2206.2  on 3691  degrees of freedom
## AIC: 2242.2
##
## Number of Fisher Scoring iterations: 6

# Month too many levels. Try to use season
gm14<-
glm(y~factor_Previous+factor_cons.price.idx+poutcome+season+default,family=binomial,data = dfw)
#Ahora no nos aparecen NA!

#anova(gm12b,gm12) #Test for nested models not equivalent
Anova(gm12b, test="LR")

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##               LR Chisq Df Pr(>Chisq)
## factor_Previous      7.266  1  0.007027 **
## factor_cons.price.idx 65.835  4  1.716e-13 ***
## poutcome            120.651  2  < 2.2e-16 ***
## month               109.822  9  < 2.2e-16 ***
## default              8.504  1  0.003543 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Add to the best model: INTERACTIONS

Un cop utilitzades variables numèriques i factors en la construcció del model, en aquest apartat utilitzarem les interaccions per tal de veure si aquesta eina millora el nostre model. I el model que tenim fins ara és el model gm12b i si surten NA agafem el model gm14, llavors farem les interaccions sobre aquest.



En el primer cas provarem de utilitzar factor_Previous com a interacció:

```

mf1<-glm(y ~
(factor_cons.price.idx+poutcome+month+default)*(factor_Previous), family =
binomial, data = dfw)

Anova(mf1,test="LR")

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##
##          LR Chisq Df Pr(>Chisq)
## factor_cons.price.idx      58.580  4  5.765e-12 ***
## poutcome          112.230  2  < 2.2e-16 ***
## month            116.016  9  < 2.2e-16 ***
## default           7.624  1  0.005759 **
## factor_Previous       7.266  1  0.007027 **
## factor_cons.price.idx:factor_Previous  2.694  3  0.441214
## poutcome:factor_Previous  1.244  1  0.264685
## month:factor_Previous    7.044  9  0.632521
## default:factor_Previous  0.880  1  0.348089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

A partir del test d'efectes nets veiem que la interacció amb factor_Previous no aporta res rellevant al model. Continuem amb el model anterior gm12b.

A continuació intentarem una interacció amb poutcome:

```

mf2<-glm(y ~
(factor_Previous+factor_cons.price.idx+month+default)*(poutcome), family =
binomial, data = dfw)

Anova(mf2,test="LR")

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##
##          LR Chisq Df Pr(>Chisq)
## factor_Previous      2.484  1  0.114983
## factor_cons.price.idx  57.032  4  1.218e-11 ***
## month            115.339  9  < 2.2e-16 ***
## default           5.134  1  0.023460 *
## poutcome          120.651  2  < 2.2e-16 ***
## factor_Previous:poutcome  0.391  1  0.531576
## factor_cons.price.idx:poutcome  10.417  6  0.108173
## month:poutcome      41.408 18  0.001337 **

```

```
## default:poutcome                1.727  2    0.421763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

BIC(mf2, gm12b)

##          df          BIC
## mf2      45 2513.298
## gm12b    18 2354.171
```

Es pot veure que hi ha una interacció que si que és rellevant, que és la month:poutcome

```
mf3<-step(mf2, k=log(nrow(dfw)))

## Start:  AIC=2513.3
## y ~ (factor_Previous + factor_cons.price.idx + month + default) *
##      (poutcome)
##
##              Df Deviance    AIC
## - month:poutcome      18   2184.9 2406.8
## - factor_cons.price.idx:poutcome  6   2153.9 2474.4
## - default:poutcome      2   2145.2 2498.6
## - factor_Previous:poutcome      1   2143.9 2505.5
## <none>                  2143.5 2513.3
##
## Step:  AIC=2406.77
## y ~ factor_Previous + factor_cons.price.idx + month + default +
##      poutcome + factor_Previous:poutcome + factor_cons.price.idx:poutcome +
##      default:poutcome
##
##              Df Deviance    AIC
## - factor_cons.price.idx:poutcome  6   2203.2 2375.8
## - default:poutcome      2   2186.8 2392.3
## - factor_Previous:poutcome      1   2185.0 2398.6
## <none>                  2184.9 2406.8
## - month                  9   2300.2 2448.2
##
## Step:  AIC=2375.77
## y ~ factor_Previous + factor_cons.price.idx + month + default +
##      poutcome + factor_Previous:poutcome + default:poutcome
##
##              Df Deviance    AIC
## - default:poutcome      2   2205.4 2361.5
## - factor_Previous:poutcome      1   2204.1 2368.4
## <none>                  2203.2 2375.8
## - factor_cons.price.idx      4   2269.2 2408.9
```

```

## - month          9    2315.2 2413.8
##
## Step:  AIC=2361.53
## y ~ factor_Previous + factor_cons.price.idx + month + default +
##      poutcome + factor_Previous:poutcome
##
##              Df Deviance    AIC
## - factor_Previous:poutcome  1    2206.2 2354.2
## <none>                      2205.4 2361.5
## - default                  1    2213.8 2361.7
## - factor_cons.price.idx     4    2272.0 2395.3
## - month                    9    2316.1 2398.2
##
## Step:  AIC=2354.17
## y ~ factor_Previous + factor_cons.price.idx + month + default +
##      poutcome
##
##              Df Deviance    AIC
## - factor_Previous          1    2213.5 2353.2
## <none>                      2206.2 2354.2
## - default                  1    2214.7 2354.5
## - factor_cons.price.idx     4    2272.1 2387.1
## - month                    9    2316.1 2390.0
## - poutcome                 2    2326.9 2458.4
##
## Step:  AIC=2353.22
## y ~ factor_cons.price.idx + month + default + poutcome
##
##              Df Deviance    AIC
## <none>                      2213.5 2353.2
## - default                  1    2222.1 2353.6
## - factor_cons.price.idx     4    2278.1 2384.9
## - month                    9    2327.5 2393.3
## - poutcome                 2    2374.7 2498.0

Anova(mf3, test="LR")

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##              LR Chisq Df Pr(>Chisq)
## factor_cons.price.idx    64.601  4  3.122e-13 ***
## month                   114.026  9  < 2.2e-16 ***
## default                   8.582  1  0.003396 **
## poutcome                 161.220  2  < 2.2e-16 ***

```



```

## (Intercept) -9.260 < 2e-16
## factor_Previousfactor_Previous-(1,5] 3.019 0.002537
## factor_cons.price.idxfactor_cons.price.idx-(93,93.4] -5.259 1.45e-07
## factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9] -2.339 0.019337
## factor_cons.price.idxfactor_cons.price.idx-(93.9,94] -6.602 4.05e-11
## factor_cons.price.idxfactor_cons.price.idx-(94,94.8] -0.925 0.354887
## poutcomePoutcome_nonexistent -0.386 0.699287
## poutcomePoutcome_success 10.434 < 2e-16
## seasonSummer -1.398 0.162027
## seasonAut-Win 1.705 0.088134
## defaultDefault_unknown -3.474 0.000514
##
## (Intercept) ***
## factor_Previousfactor_Previous-(1,5] **
## factor_cons.price.idxfactor_cons.price.idx-(93,93.4] ***
## factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9] *
## factor_cons.price.idxfactor_cons.price.idx-(93.9,94] ***
## factor_cons.price.idxfactor_cons.price.idx-(94,94.8]
## poutcomePoutcome_nonexistent
## poutcomePoutcome_success ***
## seasonSummer
## seasonAut-Win .
## defaultDefault_unknown ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2765.1 on 3708 degrees of freedom
## Residual deviance: 2306.5 on 3698 degrees of freedom
## AIC: 2328.5
##
## Number of Fisher Scoring iterations: 6

```

Interpretació del model final

$$Y = -1.475 + 0.863 \text{factor_Previousfactor_Previous-(1,5]} -$$

$$0.855 \text{factor_cons.price.idxfactor_cons.price.idx-(93,93.4]} -$$

$$0.469 \text{factor_cons.price.idxfactor_cons.price.idx-(93.4,93.9]} -$$

$$1.607 \text{factor_cons.price.idxfactor_cons.price.idx-(93.9,94]} +$$

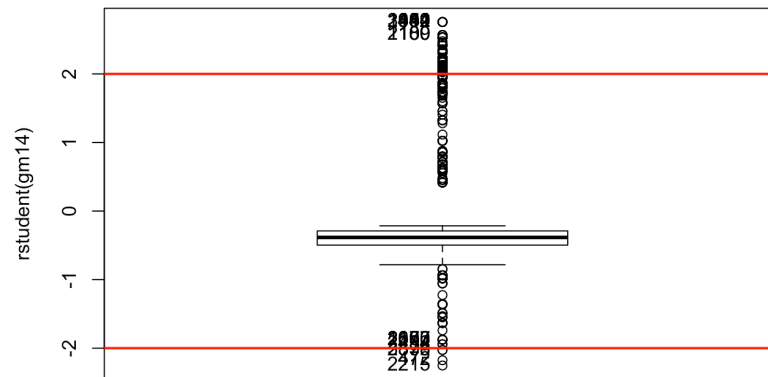
$$2.712 \text{poutcomePoutcome_success} + 0.298 \text{seasonAut-Win} - 0.598 \text{defaultDefault_unknown}$$

Anàlisi dels residus

```
Boxplot(rstudent(gm14), id.n=2)
```

```
## [1] 2215 472 2899 3378 2252 2434 1053 1373 2167 2690 144 460 612 932
## [15] 1491 2359 3432 100 1180 2109
```

```
abline(h=c(2,-2),col="red",lwd=2)
```

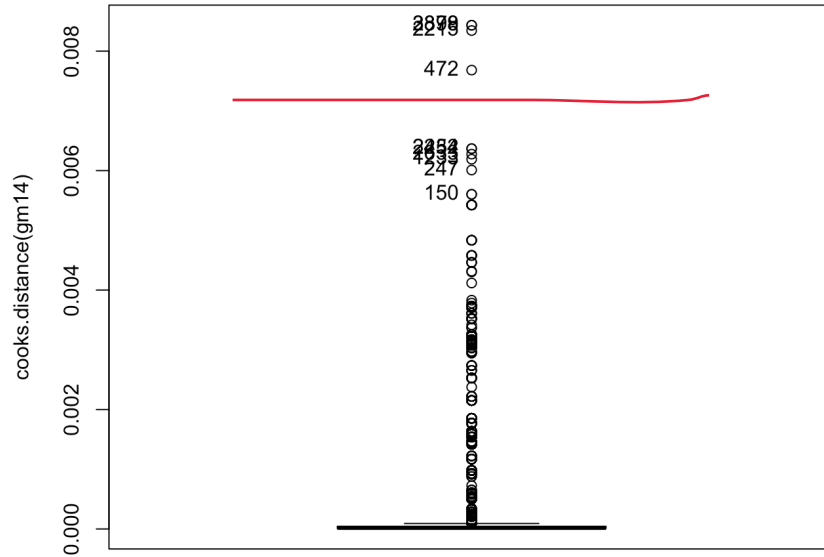


```
out2 <- which(rstudent(gm14) >= 2 | rstudent(gm14) <= -2);length(out2)
```

```
## [1] 210
```

A partir de l'anàlisi de residus veiem que no hi han quasi possibles outliers. Però ens centrarem en buscar si existeix alguna dada influent entre aquests:


```
infl<-Boxplot(cooks.distance(gm14), id.n=4)
```



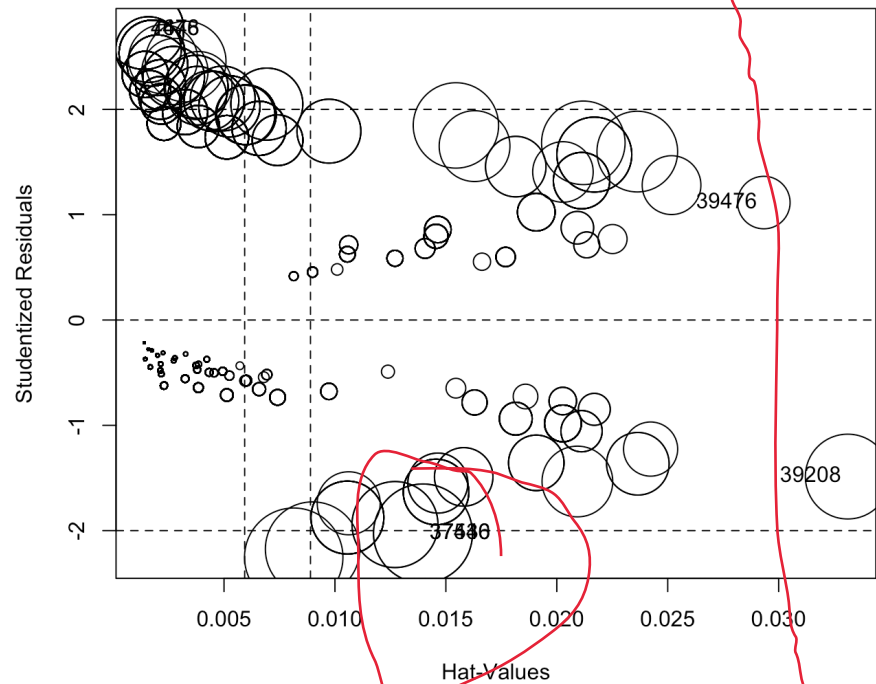
```
l1infl<-which(abs(cooks.distance(gm14))>3);length(l1infl)
```

```
## [1] 0
```

```
dfw[l1infl,]
```

```
## [1] age job marital
## [4] education default housing
## [7] loan contact month
## [10] day_of_week duration campaign
## [13] pdays previous poutcome
## [16] emp.var.rate cons.price.idx cons.conf.idx
## [19] euribor3m nr.employed y
## [22] missings_indiv errors_indiv outliers_indiv
## [25] season factor_age factor_duration
## [28] factor_campaign factor_Pdays factor_Previous
## [31] factor_emp.var.rate factor_cons.price.idx factor_cons.conf.idx
## [34] factor_euribor3m factor_nr.employed CLUSTER
## [37] f.CLUSTER
## <0 rows> (or 0-length row.names)
```

```
influencePlot(gml14,id.n=3)
```



##	Hat			
## 7446	2.757323	0.001401683	0.005424479	
## 4678	2.757323	0.001401683	0.005424479	
## 39208	-1.486823	0.033113887	0.006190260	
## 37440	-2.027270	0.013967882	0.008433304	
## 39476	1.115459	0.029331514	0.002375588	
## 37536	-2.027270	0.013967882	0.008433304	

StudRes
CookD

A partir del gràfic observat a priori es pot veure que les dades més influents són les “39208” i “39476” observant el leverage que hi ha en el plot corresponent.



Predicció

WORK

```
pre1<-predict(gml4,type="response")
pn<- as.numeric(pre1)
summary(df$y)

##   Y_no Y_yes
##  4349   597

pre.y <- factor(ifelse(pn<0.5,0,1),labels=c("pre.Success?-no","pre.Success?-yes"))

tt<-table(pre.y,dfw$y);tt


##
## pre.y           Y_no Y_yes
## pre.Success?-no  3224   353
## pre.Success?-yes    29   103

100*sum(diag(tt))/sum(tt)

## [1] 89.70073
```

TEST

```
pre<-predict(gml4,type="response",newdata=dft)
pn<- as.numeric(pre)
summary(df$y)
```



```
##   Y_no Y_yes
##  4349   597

pre.y <- factor(ifelse(pn<0.5,0,1),labels=c("pre.Success?-no","pre.Success?-yes"))

tt<-table(pre.y,dft$y);tt

##
## pre.y           Y_no Y_yes
## pre.Success?-no  1086   116
## pre.Success?-yes    10    25

100*sum(diag(tt))/sum(tt)

## [1] 89.81407
```

En aquest apartat hem realitzat les prediccions per tal de veure les taxes d'encert del nostre model. Tenim una taxa d'encert del 89.814%.

Ara tenim una altra manera de calcular la predicció: (Però aquesta em dóna error i no se perquè, perquè en el meu últim model no tinc cap NA)

```
#library("ROCR")  
#dadesroc<-prediction(predict(gml4,type="response"),df$Y)  
#par(mfrow=c(1,2))  
#plot(performance(dadesroc,"err"))  
#plot(performance(dadesroc,"tpr","fpr")) > abline(0,1,lty=2)
```