Contents

---

title: “Entrega-3”
author: “Ivan Cala Mesa - Pau Bosch Ribalta”
date: “June 4, 2023”
output:
pdf_document:
toc: yes
toc_depth: 4
html_document:
toc: yes
toc_depth: ‘4’
df_print: paged
geometry: left=1.9cm,right=1.9cm,top=1.25cm,bottom=1.52cm
fontsize: 18pt
classoption: a4paper
editor_options:
chunk_output_type: console

```
setwd("/home/pau/Escriptori/part pau")
load("./bank-additional-clean.RData")
```

Listing out variables

```
# List of variable names for the target variable
target_vars <- names(df)[c(11, 20)]

# List of variable names for the discrete variables
discrete_vars <- names(Filter(is.factor, df))
discrete_vars = discrete_vars[discrete_vars != 'y']

# List of variable names for the continuous variables
continuous_vars <- names(Filter(is.numeric, df))
```

Model de regresió lineal (target: duration)

Variables numèriques

Per començar a contruïr el nostre model, inclourem totes les variables numèriques que tenim en el nostre data set.

```
lm1 <- lm(duration ~ campaign + emp.var.rate + cons.price.idx +
            cons.conf.idx + euribor3m + nr.employed + age_num, data=df)
summary(lm1)
```

```
##
## Call:
## lm(formula = duration ~ campaign + emp.var.rate + cons.price.idx +
##     cons.conf.idx + euribor3m + nr.employed + age_num, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -805.73 -187.89  -82.33   97.75 1758.66
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -9.705e+04  5.921e+03 -16.391  < 2e-16 ***
## campaign        6.858e+00  3.135e+00   2.187  0.02877 *
## emp.var.rate   -1.743e+01  3.180e+01  -0.548  0.58366
## cons.price.idx  6.670e+01  3.164e+01   2.108  0.03505 *
## cons.conf.idx   1.095e+01  3.550e+00   3.084  0.00206 **
## euribor3m      -5.913e+02  2.611e+01 -22.652  < 2e-16 ***
## nr.employed     1.818e+01  7.080e-01  25.675  < 2e-16 ***
## age_num        -2.932e-01  4.368e-01  -0.671  0.50208
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 305.7 on 4992 degrees of freedom
## Multiple R-squared:  0.3452, Adjusted R-squared:  0.3442
## F-statistic: 375.9 on 7 and 4992 DF,  p-value: < 2.2e-16
```

```
vif(lm1)
```

```
##       campaign   emp.var.rate cons.price.idx  cons.conf.idx      euribor3m
##       1.041612      80.899677      11.053022      13.320954      74.694858
##    nr.employed        age_num
##      50.456723       1.011967
```

A partir d'aquest primer model, treurem les variables emp.var.rate i age_num ja que el model ens diu que no tenen relació amb la target. També treurem cons.price.idx ja que esta correlacionada amb cons.conf.idx i té menys importància dins del model.

```
lm2 <- lm(duration ~ campaign + cons.conf.idx + euribor3m +
            nr.employed, data=df)
summary(lm2)
```

```
##
## Call:
## lm(formula = duration ~ campaign + cons.conf.idx + euribor3m +
##     nr.employed, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -843.46 -187.86  -83.17   97.06 1752.91
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -8.689e+04  2.278e+03 -38.150  < 2e-16 ***
## campaign       6.941e+00  3.110e+00   2.232  0.02565 *
## cons.conf.idx  7.557e+00  2.792e+00   2.706  0.00683 **
## euribor3m     -5.545e+02  1.984e+01 -27.953  < 2e-16 ***
## nr.employed    1.736e+01  4.695e-01  36.987  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 305.8 on 4995 degrees of freedom
## Multiple R-squared:  0.3444, Adjusted R-squared:  0.3439
## F-statistic: 655.9 on 4 and 4995 DF,  p-value: < 2.2e-16
```

```
vif(lm2)
```

```
##      campaign cons.conf.idx     euribor3m   nr.employed
##      1.024222      8.234320     43.098511     22.170482
```

Amb aquest segon model podem veure que tenim molta colinealitat. Per arreglar aquest problema treiem l'indicador euribor3m (la variable amb més colinealitat).

```
lm3 <- lm(duration ~ campaign + cons.conf.idx + nr.employed, data=df)
summary(lm3)
```

```
##
## Call:
## lm(formula = duration ~ campaign + cons.conf.idx + nr.employed,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -698.24 -213.26  -96.62  139.88 1710.14
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.728e+04  8.602e+02 -31.714  < 2e-16 ***
## campaign       1.276e+01  3.336e+00   3.824 0.000133 ***
## cons.conf.idx -5.962e+01  1.529e+00 -38.997  < 2e-16 ***
## nr.employed    4.897e+00  1.576e-01  31.076  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 328.8 on 4996 degrees of freedom
## Multiple R-squared:  0.2418, Adjusted R-squared:  0.2414
## F-statistic: 531.1 on 3 and 4996 DF,  p-value: < 2.2e-16
```

```
vif(lm3)
```

```
##      campaign cons.conf.idx   nr.employed
##      1.019637      2.134960      2.160390
```

Amb aquest model hem solucionat el problema de la colinealitat i hem obtingut una explicativitat del 24%. Tot i que el model 3 sigui pitjor que el 2, l'agafem perquè té una colinealitat menor que 3. Si agafessim el model 2, aquest no seria vàlid. Així doncs el nostre model inicial amb les variables numèriques és el model lm3.

Variables categòriques

Una vegada tenim el model amb les variables numèriques addients, procedirem a afegir-li les variables categòriques. Per veure quines variables categòriques hem d'afegir al model començarem per fer un condes i veure'm les categories més relacionades amb la target.

```
condes(df[, c(continuous_vars[1], discrete_vars)], num.var = 1)
```

```
##
## Link between the variable and the categorical variable (1-way anova)
## ================================================
##                          R2       p.value
## month       0.2687566917  0.000000e+00
## contact     0.1775236255 2.093855e-214
## day_of_week 0.0126405296  5.186401e-13
## poutcome    0.0108893480  1.316225e-12
## default     0.0023416784  6.195630e-04
## age         0.0027889411  2.967867e-03
## marital     0.0020923915  5.335796e-03
## housing     0.0008873747  3.517575e-02
## job         0.0026763395  3.729645e-02
##
## Link between variable abd the categories of the categorical variables
## ================================================================
##                                 Estimate       p.value
## contact=cellular              165.16324 2.093855e-214
## month=jul                     333.85272 3.965288e-118
## month=aug                     280.72756  5.068452e-57
## month=jun                     341.05236  1.938499e-53
## month=nov                     208.13506  2.463898e-25
## poutcome=failure              106.55349  7.581864e-13
## day_of_week=wed                50.66271  1.397880e-06
## default=no                     21.12415  6.195630e-04
## marital=single                 20.58647  2.151955e-03
## day_of_week=thu                26.20140  3.632670e-03
## housing=yes                    11.24417  3.517575e-02
## day_of_week=fri                19.96409  3.575222e-02
## age=Jove-Adult                 57.19128  3.944233e-02
## housing=no                    -11.24417  3.517575e-02
## education=university.degree   -16.52057  1.044162e-02
## month=oct                    -245.16912  2.484858e-03
## marital=married               -19.32388  2.432148e-03
## day_of_week=mon               -36.34532  2.098132e-03
## job=unemployed                -47.45417  1.484963e-03
```

```
## age=Gran                    -146.97090  9.341034e-04
## default=NA                   -21.12415  6.195630e-04
## day_of_week=tue              -60.48287  3.926094e-09
## month=mar                   -265.09390  3.802976e-09
## poutcome=nonexistent         -90.99622  3.314190e-12
## month=may                   -178.18275 5.706054e-198
## contact=telephone           -165.16324 2.093855e-214
```

Incourem al model les 5 variables més representatives amb excepció de default, que conté un alt nombre de NA's:

- Month
- Contact
- Day_of_week
- Poutcome
- Age

```
lm4 <- lm(duration ~ campaign + cons.conf.idx + nr.employed + month
          + contact + day_of_week + poutcome + age, data=df)
summary(lm4)
```

```
##
## Call:
## lm(formula = duration ~ campaign + cons.conf.idx + nr.employed +
##     month + contact + day_of_week + poutcome + age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -663.62 -186.92  -79.99   96.84 1762.47
##
## Coefficients: (1 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -10704.282  43708.626  -0.245  0.80654
## campaign               7.204      3.123   2.307  0.02110 *
## cons.conf.idx        -45.266     73.573  -0.615  0.53841
## nr.employed            1.703      7.851   0.217  0.82828
## monthapr             324.912    218.076   1.490  0.13631
## monthmay             556.528    284.353   1.957  0.05038 .
## monthjun             816.261    407.809   2.002  0.04538 *
## monthjul             688.056    475.285   1.448  0.14777
## monthaug             927.723     57.809  16.048  < 2e-16 ***
## monthoct             361.404     75.000   4.819 1.49e-06 ***
## monthnov             649.212    173.850   3.734  0.00019 ***
## monthdec                  NA         NA      NA       NA
```

```
## contacttelephone       -90.773     29.319  -3.096  0.00197 **
## day_of_weektue          10.977     12.749   0.861  0.38926
## day_of_weekwed          42.721     13.851   3.084  0.00205 **
## day_of_weekthu          36.801     13.594   2.707  0.00681 **
## day_of_weekfri          39.557     13.990   2.828  0.00471 **
## poutcomenonexistent    -41.989     23.922  -1.755  0.07928 .
## poutcomesuccess        -97.534     39.637  -2.461  0.01390 *
## ageJove-Adult           56.449     23.945   2.357  0.01844 *
## ageAdult                56.958     24.853   2.292  0.02196 *
## ageGran               -101.233     54.363  -1.862  0.06264 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 302.3 on 4979 degrees of freedom
## Multiple R-squared:  0.3612, Adjusted R-squared:  0.3586
## F-statistic: 140.8 on 20 and 4979 DF,  p-value: < 2.2e-16
```

```
vif(lm4, type = "predictor") # Molta colinealitat
```

```
## GVIFs computed for predictors
```

```
##                         GVIF Df GVIF^(1/(2*Df)) Interacts With
## campaign       1.190082e+00  1        1.090909           --
## cons.conf.idx -2.906791e+15  1             NaN           --
## nr.employed   -3.148208e+15  1             NaN           --
## month         -1.722988e+16  8             NaN           --
## contact        1.090011e+01  1        3.301531           --
## day_of_week    1.097444e+00  4        1.011691           --
## poutcome       1.275171e+00  2        1.062654           --
## age            1.135605e+00  3        1.021420           --
##                                                                    Other Predictors
## campaign           cons.conf.idx, nr.employed, month, contact, day_of_week, poutcome, age
## cons.conf.idx           campaign, nr.employed, month, contact, day_of_week, poutcome, age
## nr.employed           campaign, cons.conf.idx, month, contact, day_of_week, poutcome, age
## month           campaign, cons.conf.idx, nr.employed, contact, day_of_week, poutcome, age
## contact           campaign, cons.conf.idx, nr.employed, month, day_of_week, poutcome, age
## day_of_week           campaign, cons.conf.idx, nr.employed, month, contact, poutcome, age
## poutcome           campaign, cons.conf.idx, nr.employed, month, contact, day_of_week, age
## age           campaign, cons.conf.idx, nr.employed, month, contact, day_of_week, poutcome
```

Amb aquest primer model que hi intervenen les variables categòriques, hem passat a tenir una explicativitat del 36%, més de 10 punts més que sols amb numèriques. Degut a l'alta colinealitat, el vif dona error. Per solucionar-ho ens veiem obligats a treure una variable numèrica. En aquest cas treurem la

variable nr.employed ja que ens dona millors resultats que si treiem la variable cons.conf.idx.

```
lm5 <- lm(duration ~ campaign + cons.conf.idx + month + contact
          + day_of_week + poutcome + age, data=df)
summary(lm5)
```

```
##
## Call:
## lm(formula = duration ~ campaign + cons.conf.idx + month + contact +
##     day_of_week + poutcome + age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -663.62 -186.92  -79.99   96.84 1762.47
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1222.639    175.269  -6.976 3.44e-12 ***
## campaign                7.204      3.123   2.307  0.02110 *
## cons.conf.idx         -29.298      3.304  -8.867  < 2e-16 ***
## monthapr              278.603     32.623   8.540  < 2e-16 ***
## monthmay              495.846     35.988  13.778  < 2e-16 ***
## monthjun              904.988     37.902  23.877  < 2e-16 ***
## monthjul              791.155     39.278  20.142  < 2e-16 ***
## monthaug              925.427     56.682  16.327  < 2e-16 ***
## monthoct              372.772     56.408   6.609 4.29e-11 ***
## monthnov              686.130     42.909  15.990  < 2e-16 ***
## monthdec               65.990    304.223   0.217  0.82828
## contacttelephone      -90.773     29.319  -3.096  0.00197 **
## day_of_weektue         10.977     12.749   0.861  0.38926
## day_of_weekwed         42.721     13.851   3.084  0.00205 **
## day_of_weekthu         36.801     13.594   2.707  0.00681 **
## day_of_weekfri         39.557     13.990   2.828  0.00471 **
## poutcomenonexistent   -41.989     23.922  -1.755  0.07928 .
## poutcomesuccess       -97.534     39.637  -2.461  0.01390 *
## ageJove-Adult          56.449     23.945   2.357  0.01844 *
## ageAdult               56.958     24.853   2.292  0.02196 *
## ageGran              -101.233     54.363  -1.862  0.06264 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 302.3 on 4979 degrees of freedom
## Multiple R-squared:  0.3612, Adjusted R-squared:  0.3586
## F-statistic: 140.8 on 20 and 4979 DF,  p-value: < 2.2e-16
```

```
vif(lm5)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## campaign        1.056717  1        1.027967
## cons.conf.idx  11.795409  1        3.434444
## month          13.461628  8        1.176437
## contact        10.900106  1        3.301531
## day_of_week     1.097444  4        1.011691
## poutcome        1.275171  2        1.062654
## age             1.135605  3        1.021420
```

Com que seguim tenint colinealitat, treure'm la variable contact, ja que de les dues variables amb colinealitat (contact i cons.conf.idx) és la que té menys importància (p valor més elevat).

```
lm6 <- lm(duration ~ campaign + cons.conf.idx+ month
          + day_of_week + poutcome + age, data=df)
summary(lm6)
```

```
##
## Call:
## lm(formula = duration ~ campaign + cons.conf.idx + month + day_of_week +
##     poutcome + age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -677.20 -186.81  -79.87   98.24 1761.82
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1667.362    100.518 -16.588  < 2e-16 ***
## campaign                6.643      3.120   2.129  0.03332 *
## cons.conf.idx         -38.189      1.635 -23.358  < 2e-16 ***
## monthapr              302.907     31.692   9.558  < 2e-16 ***
## monthmay              530.875     34.193  15.526  < 2e-16 ***
## monthjun              892.220     37.709  23.660  < 2e-16 ***
## monthjul              854.760     33.506  25.511  < 2e-16 ***
## monthaug             1050.202     39.892  26.326  < 2e-16 ***
## monthoct              371.530     56.455   6.581 5.15e-11 ***
## monthnov              748.352     37.945  19.722  < 2e-16 ***
## monthdec               15.892    304.054   0.052  0.95832
## day_of_weektue         10.638     12.760   0.834  0.40449
## day_of_weekwed         42.622     13.863   3.074  0.00212 **
## day_of_weekthu         37.807     13.602   2.779  0.00547 **
```

```
## day_of_weekfri          39.535     14.002   2.823  0.00477 **
## poutcomenonexistent    -46.060     23.906  -1.927  0.05408 .
## poutcomesuccess       -103.536     39.624  -2.613  0.00900 **
## ageJove-Adult           56.995     23.965   2.378  0.01743 *
## ageAdult                57.094     24.874   2.295  0.02176 *
## ageGran                -97.874     54.399  -1.799  0.07205 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 302.6 on 4980 degrees of freedom
## Multiple R-squared:   0.36,  Adjusted R-squared:  0.3575
## F-statistic: 147.4 on 19 and 4980 DF,  p-value: < 2.2e-16
```

```
vif(lm6)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## campaign       1.053151  1        1.026232
## cons.conf.idx 2.883342  1        1.698041
## month          3.029829  8        1.071738
## day_of_week    1.096127  4        1.011539
## poutcome       1.270512  2        1.061683
## age            1.134940  3        1.021321
```

Amb aquest model no tenim colinealitat i tenim una explicativitat del 36%. Tot i així realitzarem un step a partir del model inicial amb totes les variables categòriques per validar el nostre model.

```
comp <- step(lm4, trace = F)
summary(comp)
```

```
##
## Call:
## lm(formula = duration ~ campaign + cons.conf.idx + month + contact +
##     day_of_week + poutcome + age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -663.62 -186.92  -79.99   96.84 1762.47
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1222.639    175.269  -6.976 3.44e-12 ***
## campaign                7.204      3.123   2.307  0.02110 *
## cons.conf.idx         -29.298      3.304  -8.867  < 2e-16 ***
```

```
## monthapr              278.603     32.623   8.540  < 2e-16 ***
## monthmay              495.846     35.988  13.778  < 2e-16 ***
## monthjun              904.988     37.902  23.877  < 2e-16 ***
## monthjul              791.155     39.278  20.142  < 2e-16 ***
## monthaug              925.427     56.682  16.327  < 2e-16 ***
## monthoct              372.772     56.408   6.609 4.29e-11 ***
## monthnov              686.130     42.909  15.990  < 2e-16 ***
## monthdec               65.990    304.223   0.217  0.82828
## contacttelephone      -90.773     29.319  -3.096  0.00197 **
## day_of_weektue         10.977     12.749   0.861  0.38926
## day_of_weekwed         42.721     13.851   3.084  0.00205 **
## day_of_weekthu         36.801     13.594   2.707  0.00681 **
## day_of_weekfri         39.557     13.990   2.828  0.00471 **
## poutcomenonexistent   -41.989     23.922  -1.755  0.07928 .
## poutcomesuccess       -97.534     39.637  -2.461  0.01390 *
## ageJove-Adult          56.449     23.945   2.357  0.01844 *
## ageAdult               56.958     24.853   2.292  0.02196 *
## ageGran              -101.233     54.363  -1.862  0.06264 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 302.3 on 4979 degrees of freedom
## Multiple R-squared:  0.3612, Adjusted R-squared:  0.3586
## F-statistic: 140.8 on 20 and 4979 DF,  p-value: < 2.2e-16
```

```
vif(comp)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## campaign        1.056717  1        1.027967
## cons.conf.idx  11.795409  1        3.434444
## month          13.461628  8        1.176437
## contact        10.900106  1        3.301531
## day_of_week     1.097444  4        1.011691
## poutcome        1.275171  2        1.062654
## age             1.135605  3        1.021420
```

Amb el step podem veure que hem fet la mateixa transformació però, a més, hem fet una iteració més per treure la colinealitat i eliminar la variable contact.

Interaccions

Per a relitzar les interaccions entre variables, hem agafat totes les variables seleccionades en els apartats anteriors i les hem elevat al quadrat. D'aquesta manera totes les variables interactuaran entre si i amb un step podrem veure

quines d'aquestes interaccions són les més útils pel nostre model. A partir d'aquest moment no tindrem en compte la colinealitat ja que al fer les interaccions, com és obvi, hi haurà colinealitat alta entre molts factors del model al tenir la mateixa variable intervenint en diversos camps.

```
lm7 <- lm(duration ~ (campaign + cons.conf.idx + month
                      + day_of_week + poutcome + age)^2, data=df)
lm8 <- step(lm7, trace = F)
summary(lm8)
```

```
##
## Call:
## lm(formula = duration ~ campaign + cons.conf.idx + month + day_of_week +
##     poutcome + age + campaign:cons.conf.idx + cons.conf.idx:day_of_week +
##     cons.conf.idx:age + month:day_of_week, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -694.72 -185.70  -78.39   99.19 1753.35
##
## Coefficients: (4 not defined because of singularities)
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -671.8735   271.0847  -2.478 0.013228 *
## campaign                       -75.5175    29.4752  -2.562 0.010434 *
## cons.conf.idx                  -16.2191     5.8865  -2.755 0.005885 **
## monthapr                       379.1892    73.8897   5.132 2.98e-07 ***
## monthmay                       427.2491    71.7833   5.952 2.83e-09 ***
## monthjun                       730.3971    81.4786   8.964  < 2e-16 ***
## monthjul                       791.9446    73.2126  10.817  < 2e-16 ***
## monthaug                       865.4275    84.7744  10.209  < 2e-16 ***
## monthoct                       589.9510   132.4819   4.453 8.65e-06 ***
## monthnov                       740.4928    93.4938   7.920 2.91e-15 ***
## monthdec                       -25.8841   307.5039  -0.084 0.932921
## day_of_weektue                -363.6050   238.7221  -1.523 0.127790
## day_of_weekwed                -410.0736   249.7296  -1.642 0.100639
## day_of_weekthu                -965.7648   250.0517  -3.862 0.000114 ***
## day_of_weekfri                 -87.8400   235.8882  -0.372 0.709626
## poutcomenonexistent            -46.4907    23.9119  -1.944 0.051922 .
## poutcomesuccess               -104.6240    39.9476  -2.619 0.008845 **
## ageJove-Adult                 -474.3745   215.3042  -2.203 0.027621 *
## ageAdult                      -409.3842   224.8332  -1.821 0.068692 .
## ageGran                      -2832.2970  1745.4401  -1.623 0.104721
## campaign:cons.conf.idx          -2.0870     0.7419  -2.813 0.004925 **
## cons.conf.idx:day_of_weektue    -7.2422     4.5068  -1.607 0.108127
## cons.conf.idx:day_of_weekwed    -7.5317     4.6243  -1.629 0.103430
```

```
## cons.conf.idx:day_of_weekthu   -18.0121     4.6580  -3.867 0.000112 ***
## cons.conf.idx:day_of_weekfri    -1.7654     4.4111  -0.400 0.689015
## cons.conf.idx:ageJove-Adult    -12.4813     5.0322  -2.480 0.013161 *
## cons.conf.idx:ageAdult         -10.8165     5.2991  -2.041 0.041284 *
## cons.conf.idx:ageGran          -58.5094    36.3768  -1.608 0.107805
## monthapr:day_of_weektue       -176.7698    98.6899  -1.791 0.073328 .
## monthmay:day_of_weektue        102.3285    98.7360   1.036 0.300073
## monthjun:day_of_weektue        240.8458   117.1207   2.056 0.039797 *
## monthjul:day_of_weektue         56.1731    98.5281   0.570 0.568620
## monthaug:day_of_weektue        252.5789   116.0267   2.177 0.029535 *
## monthoct:day_of_weektue       -298.4208   169.7293  -1.758 0.078773 .
## monthnov:day_of_weektue         77.2612   125.9322   0.614 0.539565
## monthdec:day_of_weektue             NA         NA      NA       NA
## monthapr:day_of_weekwed        -61.5404   108.6259  -0.567 0.571056
## monthmay:day_of_weekwed        192.4683   110.7102   1.738 0.082187 .
## monthjun:day_of_weekwed        207.4730   121.6739   1.705 0.088228 .
## monthjul:day_of_weekwed         61.1790   110.2488   0.555 0.578976
## monthaug:day_of_weekwed        244.2173   126.8782   1.925 0.054310 .
## monthoct:day_of_weekwed       -187.5553   187.4798  -1.000 0.317164
## monthnov:day_of_weekwed         90.6598   128.2933   0.707 0.479811
## monthdec:day_of_weekwed             NA         NA      NA       NA
## monthapr:day_of_weekthu         51.5978   104.1431   0.495 0.620304
## monthmay:day_of_weekthu        312.1013   109.4682   2.851 0.004375 **
## monthjun:day_of_weekthu        388.2938   120.3355   3.227 0.001260 **
## monthjul:day_of_weekthu        267.9857   109.7992   2.441 0.014694 *
## monthaug:day_of_weekthu        501.1970   127.2834   3.938 8.34e-05 ***
## monthoct:day_of_weekthu       -110.7404   196.5259  -0.563 0.573127
## monthnov:day_of_weekthu        156.1081   126.4902   1.234 0.217205
## monthdec:day_of_weekthu             NA         NA      NA       NA
## monthapr:day_of_weekfri        -24.8768   103.5358  -0.240 0.810129
## monthmay:day_of_weekfri         70.1215   100.7695   0.696 0.486549
## monthjun:day_of_weekfri        119.0602   114.3255   1.041 0.297734
## monthjul:day_of_weekfri         48.5704   102.8987   0.472 0.636933
## monthaug:day_of_weekfri         81.1501   123.9403   0.655 0.512658
## monthoct:day_of_weekfri       -286.5132   178.7476  -1.603 0.109022
## monthnov:day_of_weekfri        -73.9207   122.1154  -0.605 0.544984
## monthdec:day_of_weekfri             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 300.9 on 4944 degrees of freedom
## Multiple R-squared:  0.3717, Adjusted R-squared:  0.3647
## F-statistic: 53.17 on 55 and 4944 DF,  p-value: < 2.2e-16
```

Repliquem el model que ens ha donat el step amb un model propi per veure exactament les variables que interactuen:

```
lm9 <- lm(formula = duration ~ campaign + cons.conf.idx + month
          + day_of_week + poutcome + age + campaign * cons.conf.idx
          + cons.conf.idx * day_of_week +
    cons.conf.idx * age + month * day_of_week, data = df)
anova(lm6, lm9)
```

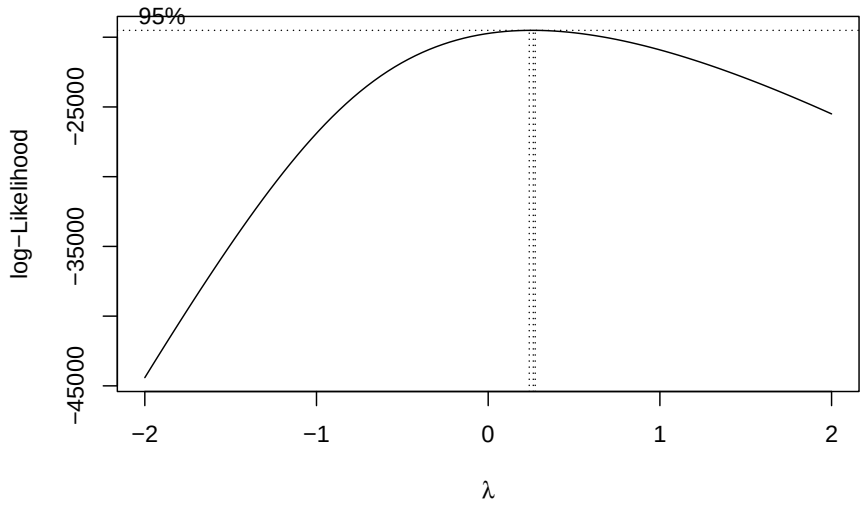
```
## Analysis of Variance Table
##
## Model 1: duration ~ campaign + cons.conf.idx + month + day_of_week + poutcome +
##     age
## Model 2: duration ~ campaign + cons.conf.idx + month + day_of_week + poutcome +
##     age + campaign * cons.conf.idx + cons.conf.idx * day_of_week +
##     cons.conf.idx * age + month * day_of_week
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1   4980 455906158
## 2   4944 447570095 36   8336063 2.5579 9.853e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fent la comparació entre el model sense i amb interaccions (utilityzant la funció anova()) podem veure que el model amb interaccions (lm9), és molt millor que el model sense (lm6). Amb la incorporació de les interaccions hem elevat fins a un 37% el percentatge d'explicativitat.

Transformacions

En aquest punt realitzarem les transformacions addients en les variables per ajustar el model a les dades que tenim. Començarem evaluant la variable original (duration), per a fer-ho utilitzarem la funció boxcox() per trobar el valor de la lambda.

```
boxcox(lm9, data=df)
```



Veiem que el valor de lambda és major que 0 però proper a aquest. Per tant, haurem de realitzar una transformació d'arrel quadrada sobre duration.

```
lm10 <- lm(sqrt(duration) ~ campaign + cons.conf.idx + month
           + day_of_week + poutcome + age + campaign * cons.conf.idx
           + cons.conf.idx * day_of_week +
            cons.conf.idx * age + month * day_of_week, data = df)
summary(lm10)
```

```
##
## Call:
## lm(formula = sqrt(duration) ~ campaign + cons.conf.idx + month +
##     day_of_week + poutcome + age + campaign * cons.conf.idx +
##     cons.conf.idx * day_of_week + cons.conf.idx * age + month *
##     day_of_week, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.160  -4.442  -1.044   3.318  29.554
##
## Coefficients: (4 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -11.03352    5.89887  -1.870 0.061481 .
## campaign                       -1.67917    0.64139  -2.618 0.008871 **
## cons.conf.idx                  -0.48380    0.12809  -3.777 0.000161 ***
```

```
## monthapr                          8.17270    1.60786   5.083 3.85e-07 ***
## monthmay                         10.59484    1.56202   6.783 1.32e-11 ***
## monthjun                         17.67570    1.77299   9.969  < 2e-16 ***
## monthjul                         18.56905    1.59312  11.656  < 2e-16 ***
## monthaug                         21.49478    1.84471  11.652  < 2e-16 ***
## monthoct                         13.06170    2.88284   4.531 6.01e-06 ***
## monthnov                         17.56963    2.03445   8.636  < 2e-16 ***
## monthdec                         -1.60212    6.69136  -0.239 0.810781
## day_of_weektue                   -7.14755    5.19465  -1.376 0.168901
## day_of_weekwed                   -7.50872    5.43417  -1.382 0.167108
## day_of_weekthu                  -21.49377    5.44118  -3.950 7.92e-05 ***
## day_of_weekfri                   -1.55254    5.13298  -0.302 0.762311
## poutcomenonexistent              -1.03818    0.52033  -1.995 0.046072 *
## poutcomesuccess                  -2.17045    0.86927  -2.497 0.012562 *
## ageJove-Adult                   -10.31877    4.68507  -2.202 0.027677 *
## ageAdult                         -8.94038    4.89242  -1.827 0.067701 .
## ageGran                         -51.10162   37.98117  -1.345 0.178543
## campaign:cons.conf.idx           -0.04503    0.01614  -2.789 0.005299 **
## cons.conf.idx:day_of_weektue     -0.13956    0.09807  -1.423 0.154789
## cons.conf.idx:day_of_weekwed     -0.13850    0.10062  -1.376 0.168771
## cons.conf.idx:day_of_weekthu     -0.39080    0.10136  -3.856 0.000117 ***
## cons.conf.idx:day_of_weekfri     -0.04135    0.09599  -0.431 0.666673
## cons.conf.idx:ageJove-Adult      -0.26813    0.10950  -2.449 0.014375 *
## cons.conf.idx:ageAdult           -0.23353    0.11531  -2.025 0.042901 *
## cons.conf.idx:ageGran            -1.05169    0.79157  -1.329 0.184035
## monthapr:day_of_weektue          -3.41734    2.14751  -1.591 0.111606
## monthmay:day_of_weektue           2.07267    2.14852   0.965 0.334744
## monthjun:day_of_weektue           4.44618    2.54857   1.745 0.081121 .
## monthjul:day_of_weektue           1.21563    2.14399   0.567 0.570743
## monthaug:day_of_weektue           4.77463    2.52477   1.891 0.058667 .
## monthoct:day_of_weektue          -6.44779    3.69335  -1.746 0.080910 .
## monthnov:day_of_weektue           0.84904    2.74031   0.310 0.756700
## monthdec:day_of_weektue                NA         NA      NA       NA
## monthapr:day_of_weekwed          -1.14216    2.36372  -0.483 0.628972
## monthmay:day_of_weekwed           3.78817    2.40908   1.572 0.115909
## monthjun:day_of_weekwed           3.74214    2.64765   1.413 0.157606
## monthjul:day_of_weekwed           1.04019    2.39904   0.434 0.664608
## monthaug:day_of_weekwed           4.41616    2.76090   1.600 0.109765
## monthoct:day_of_weekwed          -3.44174    4.07960  -0.844 0.398908
## monthnov:day_of_weekwed           1.51802    2.79169   0.544 0.586629
## monthdec:day_of_weekwed                NA         NA      NA       NA
## monthapr:day_of_weekthu           2.89658    2.26618   1.278 0.201246
## monthmay:day_of_weekthu           7.07295    2.38205   2.969 0.002999 **
## monthjun:day_of_weekthu           8.12065    2.61853   3.101 0.001938 **
## monthjul:day_of_weekthu           5.82162    2.38925   2.437 0.014862 *
## monthaug:day_of_weekthu          10.63019    2.76972   3.838 0.000126 ***
```

```
## monthoct:day_of_weekthu       -0.97342    4.27645  -0.228 0.819948
## monthnov:day_of_weekthu        3.92246    2.75246   1.425 0.154198
## monthdec:day_of_weekthu             NA         NA      NA       NA
## monthapr:day_of_weekfri       -0.34052    2.25296  -0.151 0.879869
## monthmay:day_of_weekfri        0.96361    2.19277   0.439 0.660354
## monthjun:day_of_weekfri        1.53078    2.48775   0.615 0.538367
## monthjul:day_of_weekfri        0.13530    2.23910   0.060 0.951820
## monthaug:day_of_weekfri        0.95159    2.69697   0.353 0.724225
## monthoct:day_of_weekfri       -5.86000    3.88959  -1.507 0.131981
## monthnov:day_of_weekfri       -1.98306    2.65726  -0.746 0.455535
## monthdec:day_of_weekfri             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.547 on 4944 degrees of freedom
## Multiple R-squared:  0.4048, Adjusted R-squared:  0.3982
## F-statistic: 61.15 on 55 and 4944 DF,  p-value: < 2.2e-16
```

Amb aquesta transformació hem aconseguit elevar l'explicativitat fins al 40%, però el més destacable és que hem rebaixat el residual standard error fins a un valor de 6,5, quan en l'anterior model tenia un valor de més de 300.

Seguidament realitzarem un boxTidwell sobre les variables numèriques per veure si és necessària alguna transformació. Sols realitzarem el procés per la variable campaign, ja que la variable cons.conf.idx té valors negatius i no es pot realitzar l'anàlisis.

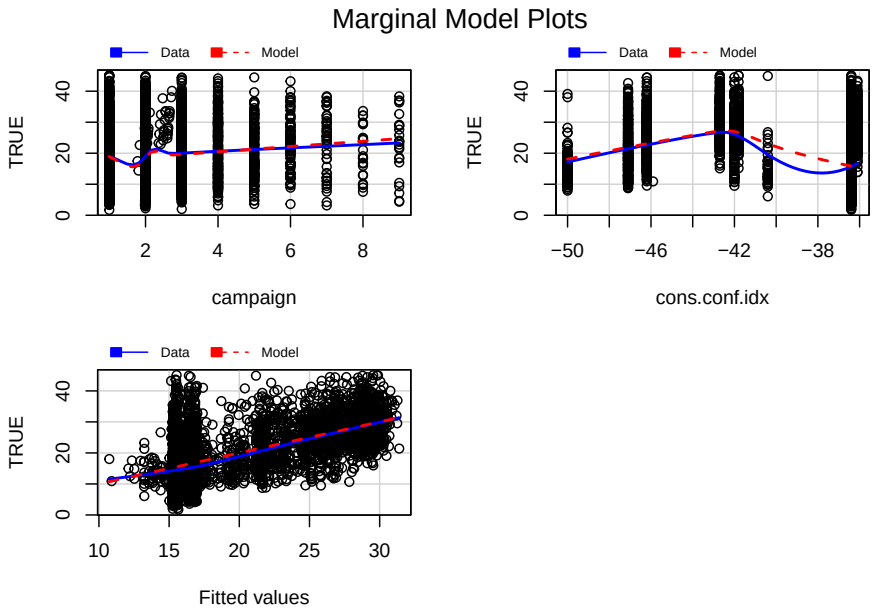
```
boxTidwell(sqrt(duration) ~ campaign, data = df)
```

```
##  MLE of lambda Score Statistic (z) Pr(>|z|)
##        0.66864             -0.7127   0.4761
##
## iterations =  2
```

La funció ens diu que no val la pena realizar una transformació, ja que el resultat del p-valor és superior a 0,05. Per a validar aquesta hipòtesis realitzarem el marginalModelPlot i podrem veure si el nostre model s'adapta a les dades o no:

```
marginalModelPlots(lm10)
```

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

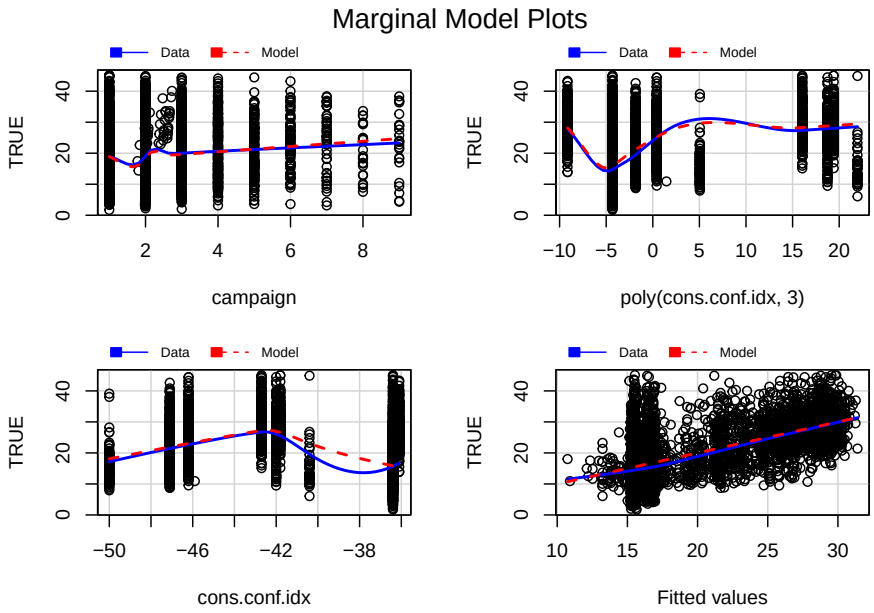


Aquest output ens mostra que, com ja hem vist anteriorment, la variable campaign ja s'ajusta prou, tot i que la variable cons.conf.idx no. Per tant, hem de transformar cons.conf.idx, per fer-ho aplicarem poly() amb un grau que ajusti el model.

```
lm11 <- lm(sqrt(duration) ~ campaign + poly(cons.conf.idx, 3)
           + day_of_week + month + poutcome + age +
             campaign * cons.conf.idx + cons.conf.idx * day_of_week +
            cons.conf.idx * age + month * day_of_week, data = df)

marginalModelPlots(lm11)
```

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

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

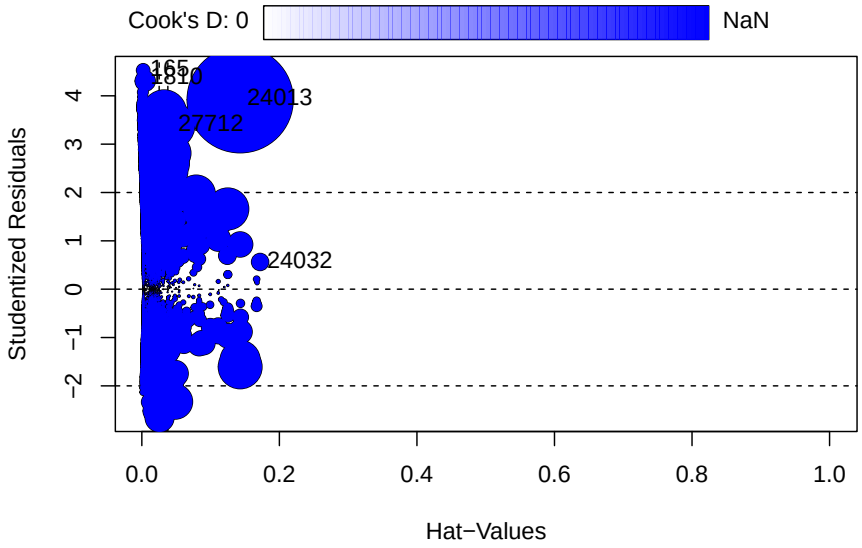


Com podem veure, amb aquesta transformació, la variable s'ajusat prou bé al model.

Individus influents

Seguidament farem una observació dels individus més influents en el model per detectar anomalies i individus fora de rang.

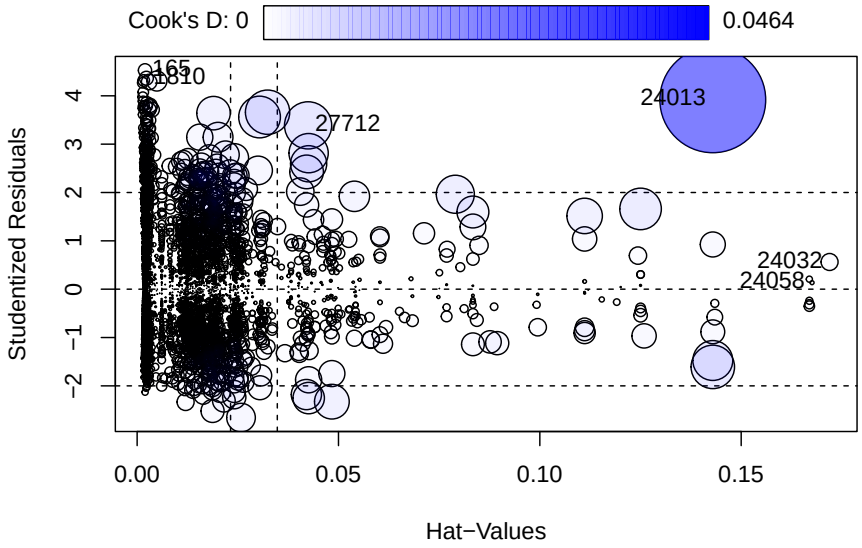
```
influencePlot(lm11)
```




```
##          StudRes         Hat        CookD
## 24032 0.5626649 0.171951716 0.0011741470
## 24013 3.9175554 0.142951939 0.0455794064
## 27712 3.3889079 0.042456976 0.0090740727
## 27690       NaN 1.000000000          NaN
## 1810  4.3643350 0.002298740 0.0007808272
## 165   4.5273485 0.001951042 0.0007126975
```

En la taula anterior podem veure com hi ha un individu el qual està destorbant el model ja que no es pot obtenir la seva Cook's distance. Per tant, el treurem del model amb un nou data frame. Una vegada tret l'individu problemàtic tornarem a realitzar el model i la influencePlot().

```
r <- which(row.names(df) == 27690)
df_new <- df[-c(r),]
lm12 <- lm(sqrt(duration) ~ campaign + poly(cons.conf.idx, 3)
           + day_of_week + month + poutcome + campaign * cons.conf.idx
           + cons.conf.idx * day_of_week +
            cons.conf.idx * age + month * day_of_week, data = df_new)
influencePlot(lm12)
```




```
##          StudRes         Hat        CookD
## 24032 0.5626649 0.171951716 1.195495e-03
## 24013 3.9175554 0.142951939 4.640812e-02
## 24058 0.1349383 0.167600252 6.667098e-05
## 27712 3.3889079 0.042456976 9.239056e-03
## 1810  4.3643350 0.002298740 7.950241e-04
## 165   4.5273485 0.001951042 7.256556e-04
```

Com podem veure, ara els individus si que són representatius i la plot té sentit. Podem veure com hi ha un individu aïllat amb una forta influència (cercle bastant gran) i una cook's distance important (blau bastant intens). La resta d'individus són menys influents i estan més agrupats. Per detectar i eliminar individus amb una alta cook's distance i un residuu alt realitzarem el següent filtrate:

```
threshold_cook <- 4/(nrow(df_new)-length(coef(lm12)))
llcoo <- which( cooks.distance(lm12) > threshold_cook)
llresid <- which(abs(rstudent(lm12))>3)

df_new <- df_new[-c(llcoo, llresid),]

lm13 <- lm(sqrt(duration) ~ campaign + poly(cons.conf.idx, 3)
           + day_of_week + month + poutcome + age
           + campaign * cons.conf.idx + cons.conf.idx * day_of_week +
```

```
            cons.conf.idx * age + month * day_of_week, data = df_new)
summary(lm13)
```

```
##
## Call:
## lm(formula = sqrt(duration) ~ campaign + poly(cons.conf.idx,
##     3) + day_of_week + month + poutcome + age + campaign * cons.conf.idx +
##     cons.conf.idx * day_of_week + cons.conf.idx * age + month *
##     day_of_week, data = df_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5421  -3.9603  -0.7935   3.1509  20.0582
##
## Coefficients: (3 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   39.170541  18.650865   2.100  0.03576 *
## campaign                      -1.514674   0.585560  -2.587  0.00972 **
## poly(cons.conf.idx, 3)1       85.960372 227.676701   0.378  0.70578
## poly(cons.conf.idx, 3)2       84.537065 269.127910   0.314  0.75345
## poly(cons.conf.idx, 3)3      460.054759 422.622853   1.089  0.27640
## day_of_weektue               -12.597103   4.702450  -2.679  0.00741 **
## day_of_weekwed                -6.710877   4.875092  -1.377  0.16871
## day_of_weekthu               -22.888307   4.911617  -4.660 3.25e-06 ***
## day_of_weekfri                -2.859184   4.565501  -0.626  0.53118
## monthapr                     -26.965850  22.430903  -1.202  0.22936
## monthmay                     -25.033608  21.839261  -1.146  0.25174
## monthjun                       0.787338   1.856188   0.424  0.67146
## monthjul                      -3.229451   4.580348  -0.705  0.48080
## monthaug                     -19.537424  27.581761  -0.708  0.47877
## monthoct                            NA         NA      NA       NA
## monthnov                            NA         NA      NA       NA
## poutcomenonexistent           -0.834889   0.486824  -1.715  0.08642 .
## poutcomesuccess               -2.287330   0.804185  -2.844  0.00447 **
## ageJove-Adult                 -3.334241   4.399893  -0.758  0.44861
## ageAdult                      -0.934495   4.575136  -0.204  0.83816
## ageGran                      -16.969017  39.366187  -0.431  0.66645
## cons.conf.idx                       NA         NA      NA       NA
## campaign:cons.conf.idx        -0.040315   0.014796  -2.725  0.00646 **
## day_of_weektue:cons.conf.idx  -0.274432   0.088567  -3.099  0.00196 **
## day_of_weekwed:cons.conf.idx  -0.159876   0.089072  -1.795  0.07273 .
## day_of_weekthu:cons.conf.idx  -0.426513   0.089981  -4.740 2.20e-06 ***
## day_of_weekfri:cons.conf.idx  -0.105810   0.085124  -1.243  0.21393
## ageJove-Adult:cons.conf.idx   -0.112943   0.102727  -1.099  0.27163
## ageAdult:cons.conf.idx        -0.050192   0.107670  -0.466  0.64112
```

```
## ageGran:cons.conf.idx          -0.347376   0.819838  -0.424  0.67180
## day_of_weektue:monthapr        -4.242844   2.027676  -2.092  0.03645 *
## day_of_weekwed:monthapr        -3.697205   2.305447  -1.604  0.10885
## day_of_weekthu:monthapr         4.105845   2.261332   1.816  0.06948 .
## day_of_weekfri:monthapr        -0.164332   2.174262  -0.076  0.93976
## day_of_weektue:monthmay         2.528148   1.969013   1.284  0.19922
## day_of_weekwed:monthmay         2.036793   2.275338   0.895  0.37075
## day_of_weekthu:monthmay         7.330692   2.306599   3.178  0.00149 **
## day_of_weekfri:monthmay        -0.290971   1.980091  -0.147  0.88318
## day_of_weektue:monthjun         4.955059   2.354238   2.105  0.03537 *
## day_of_weekwed:monthjun         1.648442   2.490135   0.662  0.50801
## day_of_weekthu:monthjun         7.153700   2.511510   2.848  0.00441 **
## day_of_weekfri:monthjun        -0.351396   2.283509  -0.154  0.87771
## day_of_weektue:monthjul         0.736795   1.949570   0.378  0.70550
## day_of_weekwed:monthjul        -0.686848   2.263195  -0.303  0.76153
## day_of_weekthu:monthjul         5.628059   2.312149   2.434  0.01497 *
## day_of_weekfri:monthjul        -1.662650   2.026001  -0.821  0.41188
## day_of_weektue:monthaug         5.370751   2.295972   2.339  0.01937 *
## day_of_weekwed:monthaug         2.301871   2.573573   0.894  0.37114
## day_of_weekthu:monthaug        10.269399   2.628100   3.908 9.46e-05 ***
## day_of_weekfri:monthaug        -0.597809   2.489168  -0.240  0.81021
## day_of_weektue:monthoct        -5.963337   6.267914  -0.951  0.34145
## day_of_weekwed:monthoct        -4.148633   6.515976  -0.637  0.52436
## day_of_weekthu:monthoct        -0.005733   6.605576  -0.001  0.99931
## day_of_weekfri:monthoct        -5.467524   6.627710  -0.825  0.40944
## day_of_weektue:monthnov        -0.462346   2.726489  -0.170  0.86535
## day_of_weekwed:monthnov         0.385305   2.746348   0.140  0.88843
## day_of_weekthu:monthnov         3.868373   2.759380   1.402  0.16101
## day_of_weekfri:monthnov        -3.527038   2.557622  -1.379  0.16795
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.688 on 4664 degrees of freedom
## Multiple R-squared:  0.4775, Adjusted R-squared:  0.4715
## F-statistic: 78.94 on 54 and 4664 DF,  p-value: < 2.2e-16
```

Com podem veure, aquest filtratge d'individus anòmals ens ha augmentat el percentatge d'explicativitat fins a un 47%. Tot i així, podem veure que hi ha variables que no aporten res al model, com és el cas de age. Per trant, traurem aquesta variable per obtenir el model final (no és un gran canvi però pensem que és millor treure-la si no té explicativitat rellevant):

```
lm.final <- lm(sqrt(duration) ~ campaign + poly(cons.conf.idx, 3)
               + day_of_week + poutcome + campaign * cons.conf.idx
               + cons.conf.idx * day_of_week +  cons.conf.idx * age
```

```
               + month * day_of_week, data = df_new)
summary(lm.final)
```

```
##
## Call:
## lm(formula = sqrt(duration) ~ campaign + poly(cons.conf.idx,
##     3) + day_of_week + poutcome + campaign * cons.conf.idx +
##     cons.conf.idx * day_of_week + cons.conf.idx * age + month *
##     day_of_week, data = df_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5421  -3.9603  -0.7935   3.1509  20.0582
##
## Coefficients: (3 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   39.170541  18.650865   2.100  0.03576 *
## campaign                      -1.514674   0.585560  -2.587  0.00972 **
## poly(cons.conf.idx, 3)1       85.960372 227.676701   0.378  0.70578
## poly(cons.conf.idx, 3)2       84.537065 269.127910   0.314  0.75345
## poly(cons.conf.idx, 3)3      460.054759 422.622853   1.089  0.27640
## day_of_weektue               -12.597103   4.702450  -2.679  0.00741 **
## day_of_weekwed                -6.710877   4.875092  -1.377  0.16871
## day_of_weekthu               -22.888307   4.911617  -4.660 3.25e-06 ***
## day_of_weekfri                -2.859184   4.565501  -0.626  0.53118
## poutcomenonexistent           -0.834889   0.486824  -1.715  0.08642 .
## poutcomesuccess               -2.287330   0.804185  -2.844  0.00447 **
## cons.conf.idx                        NA         NA      NA       NA
## ageJove-Adult                 -3.334241   4.399893  -0.758  0.44861
## ageAdult                      -0.934495   4.575136  -0.204  0.83816
## ageGran                      -16.969017  39.366187  -0.431  0.66645
## monthapr                     -26.965850  22.430903  -1.202  0.22936
## monthmay                     -25.033608  21.839261  -1.146  0.25174
## monthjun                       0.787338   1.856188   0.424  0.67146
## monthjul                      -3.229451   4.580348  -0.705  0.48080
## monthaug                     -19.537424  27.581761  -0.708  0.47877
## monthoct                             NA         NA      NA       NA
## monthnov                             NA         NA      NA       NA
## campaign:cons.conf.idx        -0.040315   0.014796  -2.725  0.00646 **
## day_of_weektue:cons.conf.idx  -0.274432   0.088567  -3.099  0.00196 **
## day_of_weekwed:cons.conf.idx  -0.159876   0.089072  -1.795  0.07273 .
## day_of_weekthu:cons.conf.idx  -0.426513   0.089981  -4.740 2.20e-06 ***
## day_of_weekfri:cons.conf.idx  -0.105810   0.085124  -1.243  0.21393
## cons.conf.idx:ageJove-Adult   -0.112943   0.102727  -1.099  0.27163
## cons.conf.idx:ageAdult        -0.050192   0.107670  -0.466  0.64112
```

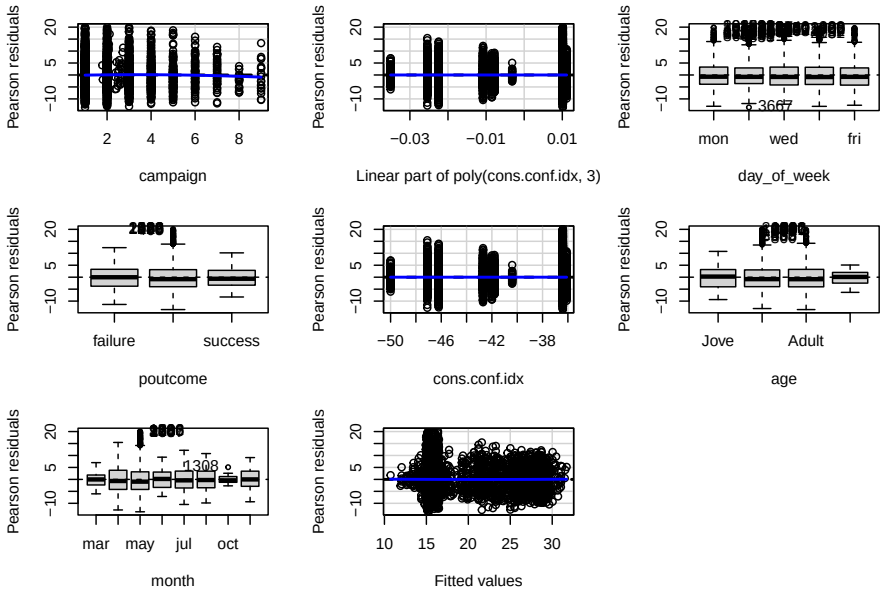
```
## cons.conf.idx:ageGran          -0.347376   0.819838  -0.424  0.67180
## day_of_weektue:monthapr        -4.242844   2.027676  -2.092  0.03645 *
## day_of_weekwed:monthapr        -3.697205   2.305447  -1.604  0.10885
## day_of_weekthu:monthapr         4.105845   2.261332   1.816  0.06948 .
## day_of_weekfri:monthapr        -0.164332   2.174262  -0.076  0.93976
## day_of_weektue:monthmay         2.528148   1.969013   1.284  0.19922
## day_of_weekwed:monthmay         2.036793   2.275338   0.895  0.37075
## day_of_weekthu:monthmay         7.330692   2.306599   3.178  0.00149 **
## day_of_weekfri:monthmay        -0.290971   1.980091  -0.147  0.88318
## day_of_weektue:monthjun         4.955059   2.354238   2.105  0.03537 *
## day_of_weekwed:monthjun         1.648442   2.490135   0.662  0.50801
## day_of_weekthu:monthjun         7.153700   2.511510   2.848  0.00441 **
## day_of_weekfri:monthjun        -0.351396   2.283509  -0.154  0.87771
## day_of_weektue:monthjul         0.736795   1.949570   0.378  0.70550
## day_of_weekwed:monthjul        -0.686848   2.263195  -0.303  0.76153
## day_of_weekthu:monthjul         5.628059   2.312149   2.434  0.01497 *
## day_of_weekfri:monthjul        -1.662650   2.026001  -0.821  0.41188
## day_of_weektue:monthaug         5.370751   2.295972   2.339  0.01937 *
## day_of_weekwed:monthaug         2.301871   2.573573   0.894  0.37114
## day_of_weekthu:monthaug        10.269399   2.628100   3.908 9.46e-05 ***
## day_of_weekfri:monthaug        -0.597809   2.489168  -0.240  0.81021
## day_of_weektue:monthoct        -5.963337   6.267914  -0.951  0.34145
## day_of_weekwed:monthoct        -4.148633   6.515976  -0.637  0.52436
## day_of_weekthu:monthoct        -0.005733   6.605576  -0.001  0.99931
## day_of_weekfri:monthoct        -5.467524   6.627710  -0.825  0.40944
## day_of_weektue:monthnov        -0.462346   2.726489  -0.170  0.86535
## day_of_weekwed:monthnov         0.385305   2.746348   0.140  0.88843
## day_of_weekthu:monthnov         3.868373   2.759380   1.402  0.16101
## day_of_weekfri:monthnov        -3.527038   2.557622  -1.379  0.16795
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.688 on 4664 degrees of freedom
## Multiple R-squared:  0.4775, Adjusted R-squared:  0.4715
## F-statistic: 78.94 on 54 and 4664 DF,  p-value: < 2.2e-16
```

Anàlisis del model final

Per acabar aquest model, farem l'anàlisis final per treure les conclusions de la validesa i qualitat del model.

Primer de tot començarem observant la distribució dels residus de pearson per cada variable del model:

```
par(mfrow = c(1, 1))
residualPlots(lm.final)
```

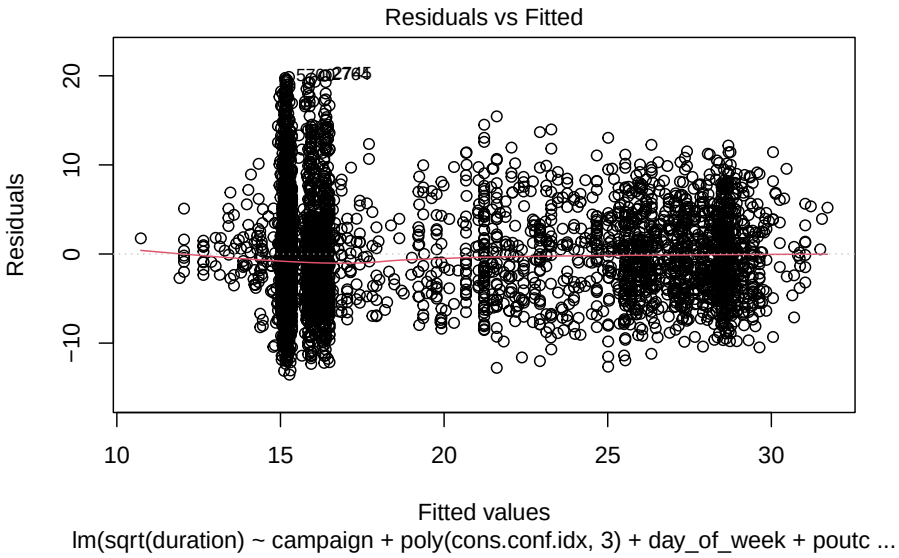


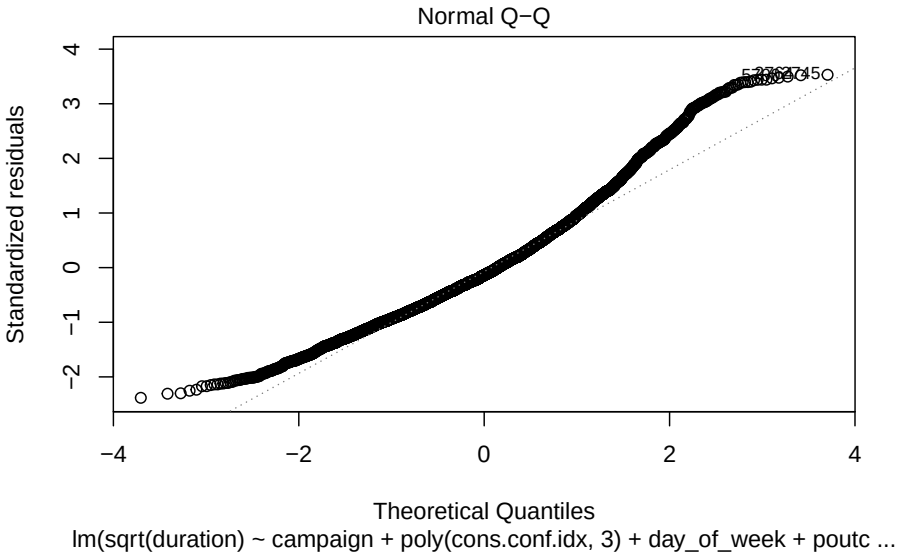

```
##                          Test stat Pr(>|Test stat|)
## campaign                   -1.2950           0.1954
## poly(cons.conf.idx, 3)
## day_of_week
## poutcome
## cons.conf.idx               0.9634           0.3354
## age
## month
## Tukey test                  0.7228           0.4698
```

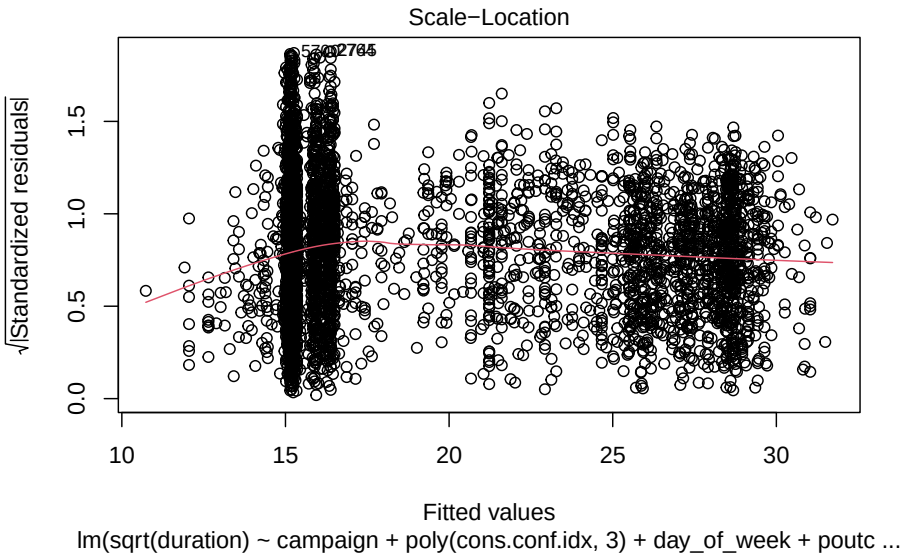
Podem veure que els residus estan, generalment, alineats en el 0, fet que ens aporta una bona seguretat en les dades del model. Tot i així hi ha alguns individus aïllasts que estan fora del rang idílic. Tot i així no són preocupants ja que no hi ha residus amb un valor més gran que

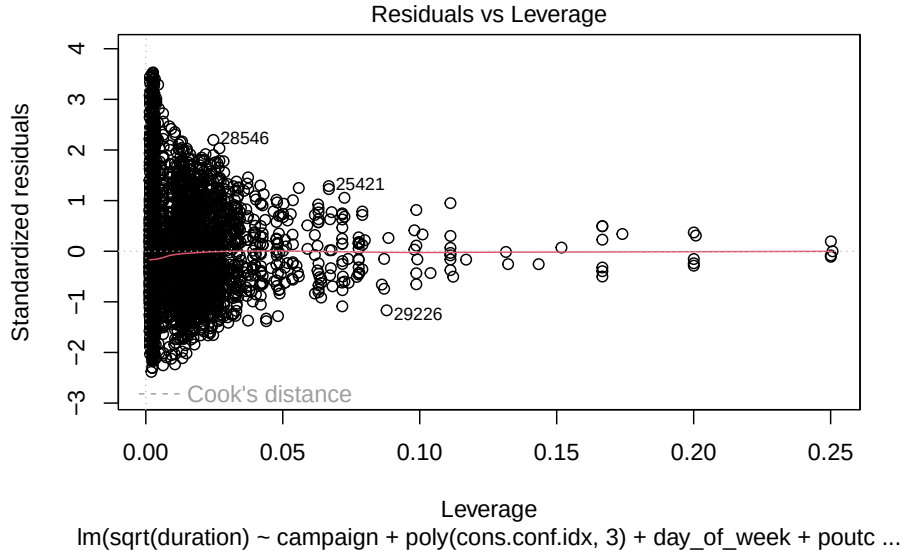
```
par(mfrow = c(1, 1))
plot(lm.final)
```

```
## Warning: not plotting observations with leverage one:
##   2047
```


Residuals vs Fitted
Residuals
Fitted values
lm(sqrt(duration) ~ campaign + poly(cons.conf.idx, 3) + day_of_week + poutc ...


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(sqrt(duration) ~ campaign + poly(cons.conf.idx, 3) + day_of_week + poutc ...





Per a cada plot farem el seu anàlisis partinent:

1. Els residus no mostren patrons significatius, estan repartits per tot l'espai i la seva variabilitat és regular i centrada al 0. Podem veure que compleix la homoscedisticitat i la independència.

2. Podem veure que els residus no segueixen una distribució normal, ja que existeixen dues cues significatives a l'inici i al final de la plot.

3. En la línia de la primera plot, podem veure que segueix sent homoscedàtica però aquí sí que podem veure dos patrons més marcats en els valors fitats 15 i 28 aproximadament. A més, la línia smooth no és regular.

4. En aquesta última plot podem veure que els residus estan ben distribuïts ja que la majoria es troben amb un leverage inferior a 0,05 i amb residus propers al 0. Tot i així podem veure alguns individus fora de rang amb leverages més elevats, tot i així la variabilitat es manté constant al 0.

Model de regresió logística (target: y)

Con el objetivo de poder encuadrar la variable binaria 'Y' en una regresión logística, será necesario asumir que las observaciones son independientes entre sí. Esto significa que los valores de la variable objetivo "Y" de una observación no están relacionados o influenciados por los valores de otras observaciones.

```
# División del conjunto de datos en muestra de trabajo y muestra de prueba
set.seed(123)  # Para reproducibilidad
sample_size <- floor(0.8 * nrow(df))  # Tamaño de muestra de trabajo (80%)
train_index <- sample(seq_len(nrow(df)),
                      size = sample_size)# Índices de muestra de trabajo

# Creación de la muestra de trabajo y muestra de prueba
train_data <- df[train_index, ]
test_data <- df[-train_index, ]
```

```
continuous_vars
```

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

```
cor(df$duration, df[,continuous_vars[-1]])
```

```
##        campaign emp.var.rate cons.price.idx cons.conf.idx  euribor3m
## [1,] 0.09357119  -0.08878741     -0.1888923     -0.291847 -0.1201745
##      nr.employed     age_num    na_count
## [1,]  0.05776931 -0.05293015 -0.02361006
```

When undertaking the analysis of various models in R, one may often confront a binomial regression problem. A binomial regression refers to the statistical

process of modeling a binary outcome as a function of one or more predictors. One potential solution is to employ a linear model, which may be a favorable option under certain circumstances, particularly when elucidating the connection between predictor variables and the likelihood of a positive binary result.

The appeal of linear models for binomial regression problems resides in three major attributes:

1. Coefficients Interpretation: Linear models facilitate the interpretation of the relationship between predictor variables and the expected outcome. They deliver coefficients that signify the direction and magnitude of the impact of various predictors on the probability of success.

2. Simplicity and Familiarity: Linear regression has a long-standing history and broad application in statistics. This universal use and basic understanding make it an accessible tool for binomial regression, particularly when the focus is on discerning the linear connection between predictors and the log-odds of success.

3. Continuous Predictor Variables: For binomial regression problems primarily featuring continuous predictors, a linear model can accurately portray their linear relationships with the log-odds of success.

However, it is essential to recognize that a linear model may not always be the optimal choice for binomial regression problems. Specifically designed for binary outcomes, alternative models may sometimes prove more suitable.

Indeed, binomial regression scenarios often necessitate models explicitly constructed for binary results. While linear models offer insights into the connection between predictors and the expected outcome, they may fail to adequately represent nonlinear relationships or consider the inherent probabilities associated with binary outcomes.

Alternative models tailored for binomial regression, such as logistic regression, probit regression, and complementary log-log regression, account for these underlying probabilities of success. They incorporate suitable link functions to model the relationship between predictors and the likelihood of success directly.

When analyzing our specific problem with 'y' as the target variable in R, fitting a linear model could help to illustrate the linear relationship between the predictor variables and the expected value of 'y'. Nonetheless, it remains crucial to reflect upon the assumptions and limitations of linear regression and investigate alternative models like logistic regression, which might more accurately model the probability of 'y' being a success.

```
# Modelo inicial con dos variables numéricas y factores significativos
model <- glm(y ~ age_num + duration + job + housing + loan
             + month + euribor3m, data = train_data, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model)
```

```
##
## Call:
## glm(formula = y ~ age_num + duration + job + housing + loan +
##     month + euribor3m, family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7724  -0.1779  -0.1011   0.0000   3.0396
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       2.577e+01  2.720e+03   0.009  0.99244
## age_num          -1.400e-02  1.201e-02  -1.165  0.24400
## duration          6.359e-03  3.543e-04  17.946  < 2e-16 ***
## jobblue-collar   -6.944e-02  3.028e-01  -0.229  0.81859
## jobmanagement    -2.609e-01  4.468e-01  -0.584  0.55930
## jobself-employed -7.843e-02  4.694e-01  -0.167  0.86730
## jobservices      -3.077e-01  3.838e-01  -0.802  0.42265
## jobtechnician     1.832e-01  3.616e-01   0.507  0.61248
## jobunemployed     1.754e-01  3.972e-01   0.442  0.65884
## housingyes        2.689e-01  2.134e-01   1.260  0.20760
## loanyes           2.397e-01  2.950e-01   0.812  0.41659
## monthapr         -1.551e+00  3.057e+03  -0.001  0.99960
## monthmay         -1.544e+01  2.720e+03  -0.006  0.99547
## monthjun          7.300e+00  3.379e+03   0.002  0.99828
## monthjul          7.375e+00  3.081e+03   0.002  0.99809
## monthaug          7.769e+00  3.233e+03   0.002  0.99808
## monthoct          1.039e+01  5.539e+03   0.002  0.99850
## monthnov          6.588e+00  3.322e+03   0.002  0.99842
## monthdec          4.825e+00  2.936e+04   0.000  0.99987
## euribor3m        -3.146e+00  1.170e+00  -2.690  0.00715 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5540.69  on 3999  degrees of freedom
## Residual deviance:  683.42  on 3980  degrees of freedom
## AIC: 723.42
##
## Number of Fisher Scoring iterations: 20
```

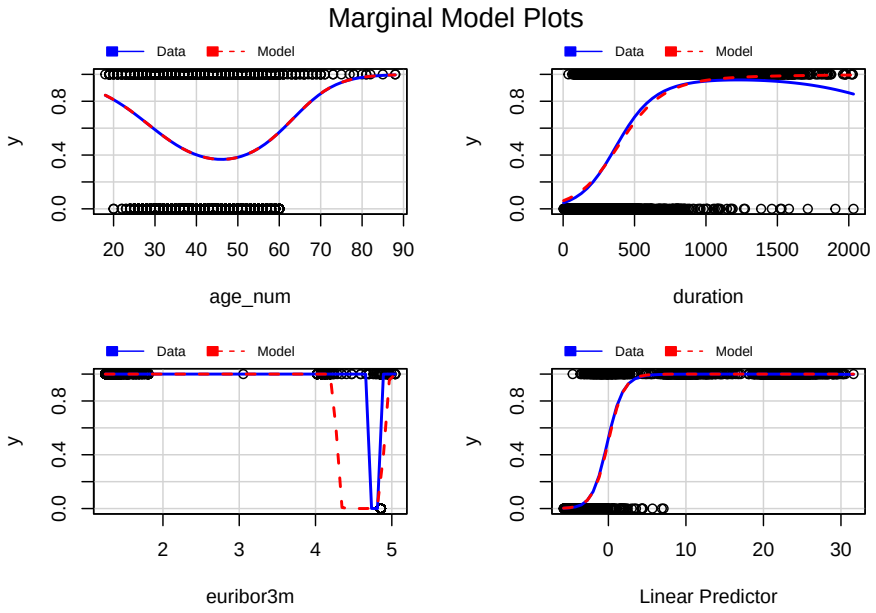
```
vif(model)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## age_num    1.082569  1        1.040466
## duration   1.015231  1        1.007587
## job        1.096155  6        1.007680
## housing    1.033585  1        1.016654
## loan       1.021006  1        1.010448
## month      1.000010  8        1.000001
## euribor3m  1.000766  1        1.000383
```

```
marginalModelPlots(model)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

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




```
Anova(model)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##           LR Chisq Df Pr(>Chisq)
## age_num       1.37  1     0.2423
## duration    611.67  1     <2e-16 ***
## job           2.23  6     0.8974
## housing       1.59  1     0.2070
## loan          0.64  1     0.4221
## month      1194.42  8     <2e-16 ***
## euribor3m   930.41  1     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model1 <- glm(y ~ duration + housing + month + euribor3m
            + marital + day_of_week, data = train_data, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model1)
```

```
##
## Call:
## glm(formula = y ~ duration + housing + month + euribor3m + marital +
##     day_of_week, family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6287  -0.1719  -0.0803   0.0000   3.1949
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.473e+01  2.674e+03   0.009 0.992623
## duration        6.369e-03  3.610e-04  17.642  < 2e-16 ***
## housingyes      2.882e-01  2.145e-01   1.344 0.178975
## monthapr       -1.674e+00  3.006e+03  -0.001 0.999556
## monthmay       -1.541e+01  2.674e+03  -0.006 0.995403
```

```
## monthjun        7.392e+00  3.333e+03   0.002 0.998230
## monthjul        7.563e+00  3.036e+03   0.002 0.998013
## monthaug        7.996e+00  3.191e+03   0.003 0.998000
## monthoct        1.074e+01  5.409e+03   0.002 0.998415
## monthnov        6.663e+00  3.278e+03   0.002 0.998378
## monthdec        5.298e+00  2.935e+04   0.000 0.999856
## euribor3m      -3.230e+00  1.191e+00  -2.712 0.006686 **
## maritalmarried  3.549e-01  3.825e-01   0.928 0.353454
## maritalsingle   8.095e-01  4.135e-01   1.958 0.050283 .
## day_of_weektue -2.484e-01  3.506e-01  -0.709 0.478619
## day_of_weekwed  6.115e-01  3.390e-01   1.804 0.071274 .
## day_of_weekthu  4.636e-01  3.621e-01   1.280 0.200520
## day_of_weekfri  1.139e+00  3.304e-01   3.449 0.000563 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5540.69  on 3999  degrees of freedom
## Residual deviance:  659.16  on 3982  degrees of freedom
## AIC: 695.16
##
## Number of Fisher Scoring iterations: 20
```

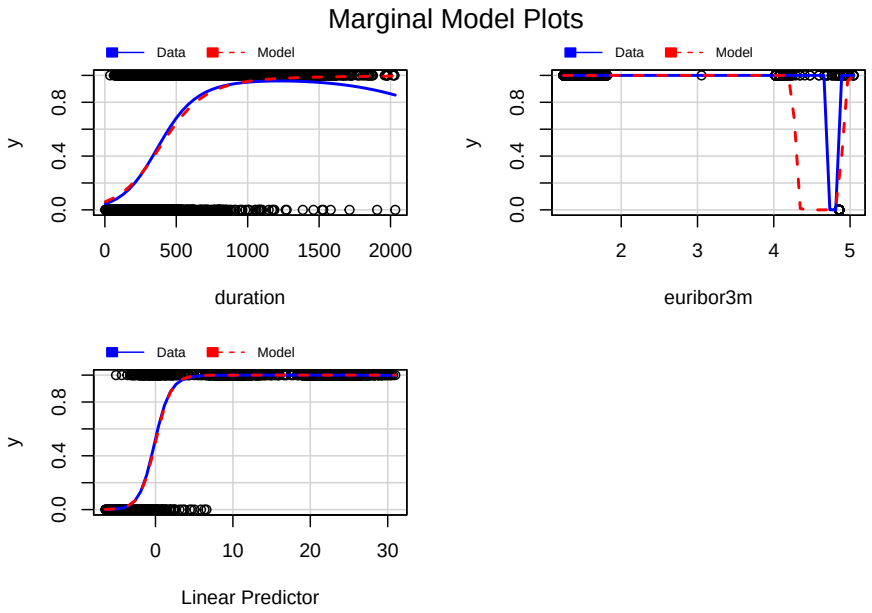
```
vif(model1)
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## duration    1.035065  1        1.017382
## housing     1.010823  1        1.005397
## month       1.000011  8        1.000001
## euribor3m   1.001608  1        1.000804
## marital     1.033009  2        1.008152
## day_of_week 1.044631  4        1.005473
```

```
marginalModelPlots(model1)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

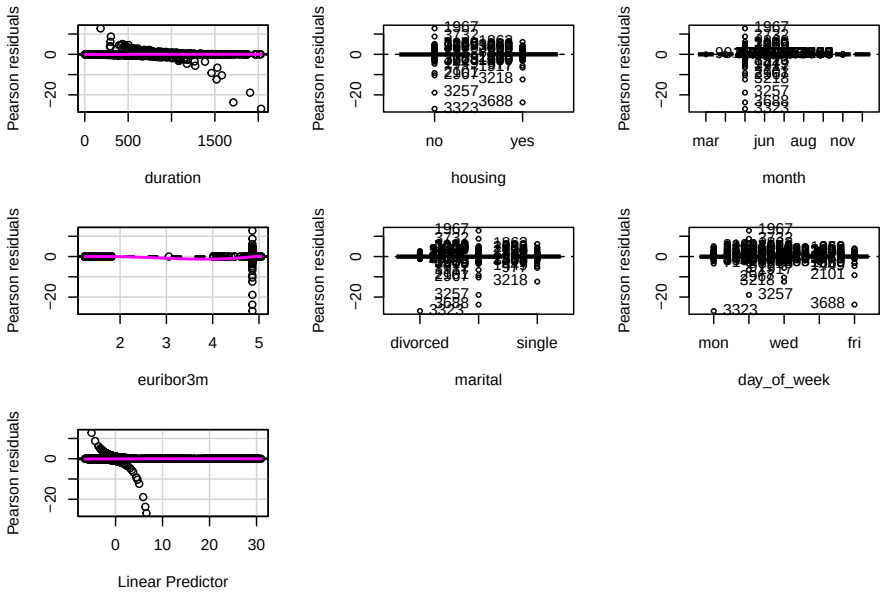
```
## Warning in mmps(...): Interactions and/or factors skipped
```




```
residualPlots(model1)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```




```
##               Test stat Pr(>|Test stat|)
## duration         99.257           <2e-16 ***
## housing
## month
## euribor3m   -202915.394                1
## marital
## day_of_week
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(model1)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##             LR Chisq Df Pr(>Chisq)
```

```
## duration      598.00  1  < 2.2e-16 ***
## housing         1.81  1  0.1783426
## month        1220.88  8  < 2.2e-16 ***
## euribor3m     933.41  1  < 2.2e-16 ***
## marital         5.08  2  0.0790392 .
## day_of_week    21.89  4  0.0002104 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(model1)
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## duration    1.035065  1        1.017382
## housing     1.010823  1        1.005397
## month       1.000011  8        1.000001
## euribor3m   1.001608  1        1.000804
## marital     1.033009  2        1.008152
## day_of_week 1.044631  4        1.005473
```

This model has non-significant variables and does not meet the model requirements, like interactions. As such, a new model has to be devised.

```
# Incorporar interacciones entre factores y una covariable
model2 <- glm(y ~ duration * housing + euribor3m + marital * day_of_week
              + month, data = train_data, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model2)
```

```
##
## Call:
## glm(formula = y ~ duration * housing + euribor3m + marital *
##     day_of_week + month, family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6501  -0.1707  -0.0612   0.0000   3.1962
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    2.298e+01  2.530e+03   0.009  0.99275
## duration                       6.162e-03  4.976e-04  12.385  < 2e-16 ***
```

```
## housingyes                     -8.476e-02  4.652e-01  -0.182  0.85543
## euribor3m                      -3.496e+00  1.235e+00  -2.831  0.00464 **
## maritalmarried                  3.326e+00  1.599e+00   2.081  0.03745 *
## maritalsingle                   3.748e+00  1.653e+00   2.267  0.02337 *
## day_of_weektue                  2.210e+00  1.988e+00   1.111  0.26638
## day_of_weekwed                  3.851e+00  1.698e+00   2.268  0.02333 *
## day_of_weekthu                  3.060e+00  1.753e+00   1.746  0.08082 .
## day_of_weekfri                  4.616e+00  1.694e+00   2.725  0.00643 **
## monthapr                       -1.550e+00  2.838e+03  -0.001  0.99956
## monthmay                       -1.507e+01  2.530e+03  -0.006  0.99525
## monthjun                        8.512e+00  3.137e+03   0.003  0.99783
## monthjul                        8.252e+00  2.888e+03   0.003  0.99772
## monthaug                        8.877e+00  3.017e+03   0.003  0.99765
## monthoct                        1.136e+01  5.267e+03   0.002  0.99828
## monthnov                        6.995e+00  3.151e+03   0.002  0.99823
## monthdec                        5.212e+00  2.934e+04   0.000  0.99986
## duration:housingyes             5.899e-04  7.192e-04   0.820  0.41207
## maritalmarried:day_of_weektue -2.694e+00  2.034e+00  -1.324  0.18552
## maritalsingle:day_of_weektue  -2.169e+00  2.107e+00  -1.030  0.30319
## maritalmarried:day_of_weekwed -3.396e+00  1.748e+00  -1.943  0.05201 .
## maritalsingle:day_of_weekwed  -3.438e+00  1.825e+00  -1.884  0.05959 .
## maritalmarried:day_of_weekthu -2.584e+00  1.804e+00  -1.432  0.15202
## maritalsingle:day_of_weekthu  -3.006e+00  1.906e+00  -1.577  0.11482
## maritalmarried:day_of_weekfri -3.701e+00  1.738e+00  -2.129  0.03321 *
## maritalsingle:day_of_weekfri  -3.618e+00  1.805e+00  -2.004  0.04504 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5540.69  on 3999  degrees of freedom
## Residual deviance:  650.72  on 3973  degrees of freedom
## AIC: 704.72
##
## Number of Fisher Scoring iterations: 20
```

```
Anova(model2)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##                      LR Chisq Df Pr(>Chisq)
## duration              601.82  1  < 2.2e-16 ***
## housing                 1.37  1  0.2412280
## euribor3m             935.10  1  < 2.2e-16 ***
## marital                 5.16  2  0.0757338 .
## day_of_week            21.84  4  0.0002154 ***
## month                1212.69  8  < 2.2e-16 ***
## duration:housing        0.67  1  0.4115943
## marital:day_of_week     7.59  8  0.4749742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

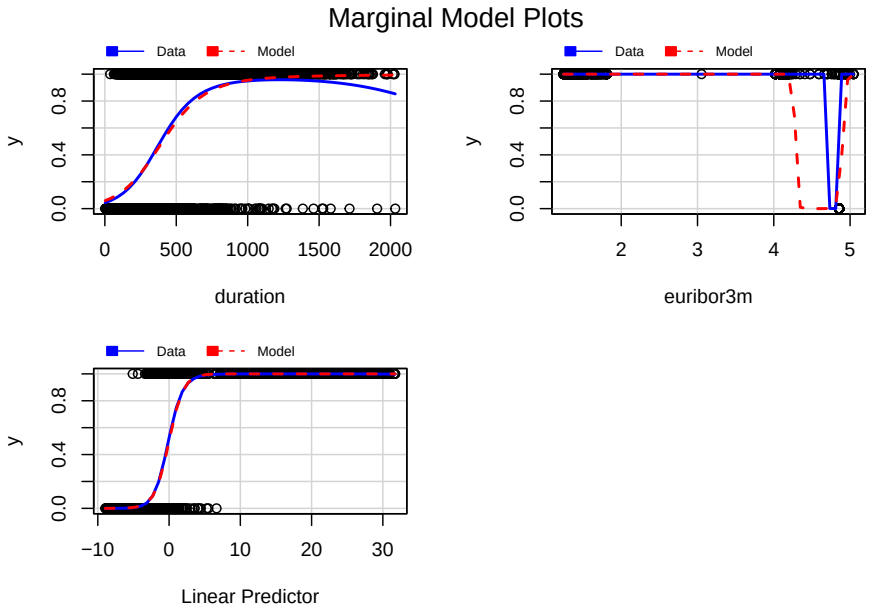
```
vif(model2)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif

##                              GVIF Df GVIF^(1/(2*Df))
## duration             1.963690e+00  1        1.401317
## housing              4.700964e+00  1        2.168171
## euribor3m            1.061703e+00  1        1.030390
## marital              9.848061e+01  2        3.150197
## day_of_week          5.035490e+04  4        3.870394
## month                1.000013e+00  8        1.000001
## duration:housing     5.685354e+00  1        2.384398
## marital:day_of_week  3.676400e+05  8        2.227609
```

```
marginalModelPlots(model2)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

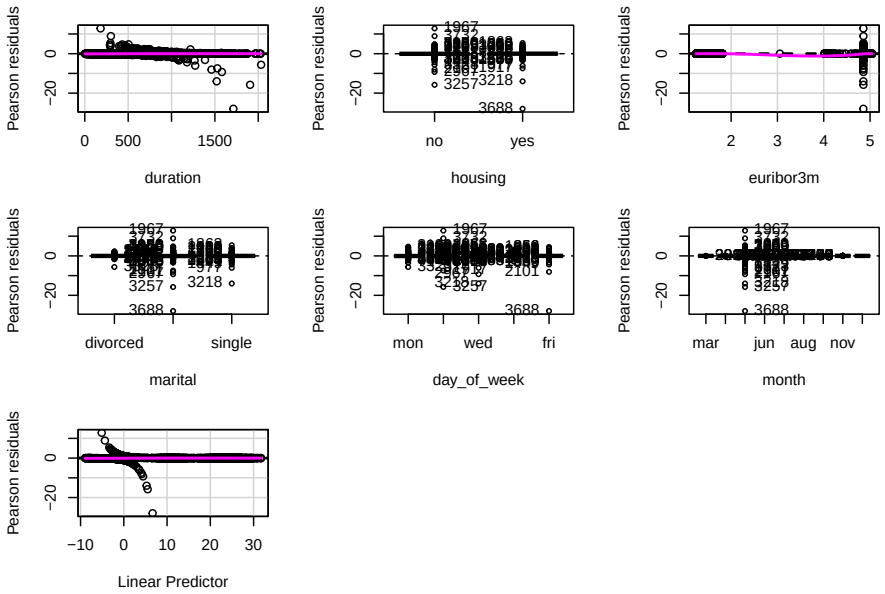
```
## Warning in mmps(...): Interactions and/or factors skipped
```




```
residualPlots(model2)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```




```
##                Test stat Pr(>|Test stat|)
## duration          93.935           <2e-16 ***
## housing
## euribor3m    -202491.307                1
## marital
## day_of_week
## month
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Incorporación de interacción entre factores
model3 <- glm(y ~ duration + housing*euribor3m + marital*day_of_week
              + month, data = train_data, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
vif(model3)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## duration             1.073273e+00  1        1.035989
```

```
## housing             2.978317e+03  1       54.573959
## euribor3m           2.244832e+00  1        1.498276
## marital             9.291725e+01  2        3.104732
## day_of_week         4.351404e+04  4        3.800394
## month               1.000017e+00  8        1.000001
## housing:euribor3m   2.979348e+03  1       54.583400
## marital:day_of_week 3.189494e+05  8        2.207916
```

```
anova(model3)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##                      Df Deviance Resid. Df Resid. Dev
## NULL                                  3999     5540.7
## duration              1  1985.19      3998     3555.5
## housing               1    30.95      3997     3524.5
## euribor3m             1  1606.93      3996     1917.6
## marital               2    13.37      3994     1904.2
## day_of_week           4    24.20      3990     1880.0
## month                 8  1220.88      3982      659.2
## housing:euribor3m     1     0.00      3981      659.2
## marital:day_of_week   8     7.76      3973      651.4
```

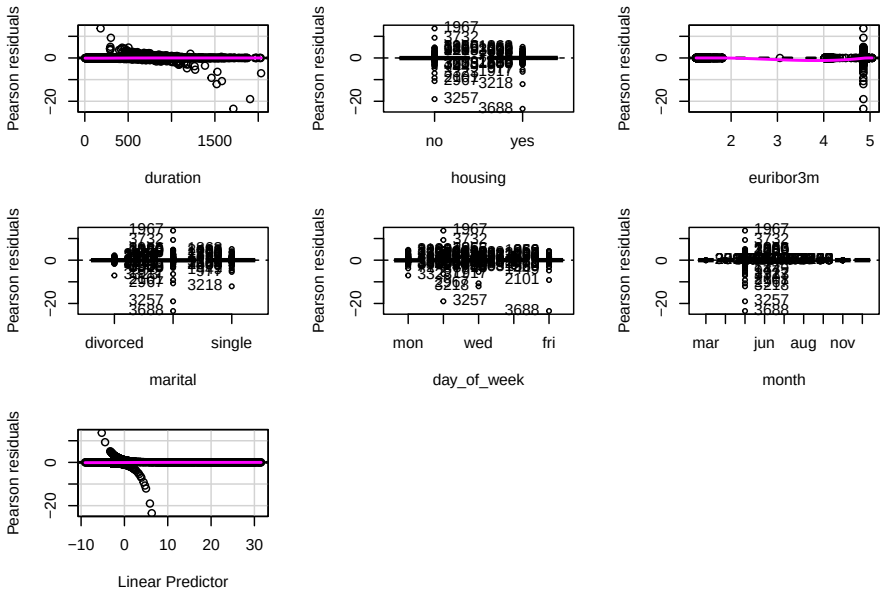
```
Anova(model3)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table (Type II tests)
##
## Response: y
##                      LR Chisq Df Pr(>Chisq)
## duration              601.82  1  < 2.2e-16 ***
## housing                 1.37  1  0.2412280
## euribor3m             934.75  1  < 2.2e-16 ***
## marital                 5.08  2  0.0790409 .
## day_of_week            21.89  4  0.0002104 ***
## month                1212.50  8  < 2.2e-16 ***
## housing:euribor3m       0.00  1  0.9774273
## marital:day_of_week     7.76  8  0.4569025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residualPlots(model3)
```

```
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```




```
##             Test stat Pr(>|Test stat|)
## duration       94.607           <2e-16 ***
## housing
## euribor3m   -8287.429                1
## marital
## day_of_week
## month
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model3)
```

```
##
## Call:
## glm(formula = y ~ duration + housing * euribor3m + marital *
##     day_of_week + month, family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5536  -0.1714  -0.0610   0.0000   3.2353
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    2.250e+01  2.529e+03   0.009  0.99290
```

```
## duration                          6.451e-03  3.679e-04  17.533  < 2e-16 ***
## housingyes                        5.855e-01  1.169e+01   0.050  0.96006
## euribor3m                        -3.445e+00  1.797e+00  -1.917  0.05519 .
## maritalmarried                    3.303e+00  1.536e+00   2.151  0.03151 *
## maritalsingle                     3.700e+00  1.593e+00   2.323  0.02018 *
## day_of_weektue                    2.226e+00  1.925e+00   1.156  0.24759
## day_of_weekwed                    3.848e+00  1.638e+00   2.349  0.01880 *
## day_of_weekthu                    3.076e+00  1.688e+00   1.822  0.06839 .
## day_of_weekfri                    4.595e+00  1.634e+00   2.812  0.00493 **
## monthapr                         -1.328e+00  2.835e+03   0.000  0.99963
## monthmay                         -1.497e+01  2.529e+03  -0.006  0.99528
## monthjun                          8.602e+00  3.133e+03   0.003  0.99781
## monthjul                          8.323e+00  2.888e+03   0.003  0.99770
## monthaug                          8.952e+00  3.018e+03   0.003  0.99763
## monthoct                          1.147e+01  5.257e+03   0.002  0.99826
## monthnov                          7.120e+00  3.149e+03   0.002  0.99820
## monthdec                          5.135e+00  2.934e+04   0.000  0.99986
## housingyes:euribor3m             -6.847e-02  2.408e+00  -0.028  0.97732
## maritalmarried:day_of_weektue    -2.729e+00  1.972e+00  -1.383  0.16653
## maritalsingle:day_of_weektue     -2.156e+00  2.048e+00  -1.053  0.29236
## maritalmarried:day_of_weekwed    -3.407e+00  1.689e+00  -2.017  0.04365 *
## maritalsingle:day_of_weekwed     -3.427e+00  1.770e+00  -1.936  0.05289 .
## maritalmarried:day_of_weekthu    -2.606e+00  1.741e+00  -1.497  0.13433
## maritalsingle:day_of_weekthu     -2.981e+00  1.847e+00  -1.614  0.10653
## maritalmarried:day_of_weekfri    -3.690e+00  1.680e+00  -2.197  0.02804 *
## maritalsingle:day_of_weekfri     -3.570e+00  1.749e+00  -2.041  0.04128 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5540.7  on 3999  degrees of freedom
## Residual deviance:  651.4  on 3973  degrees of freedom
## AIC: 705.4
##
## Number of Fisher Scoring iterations: 20
```

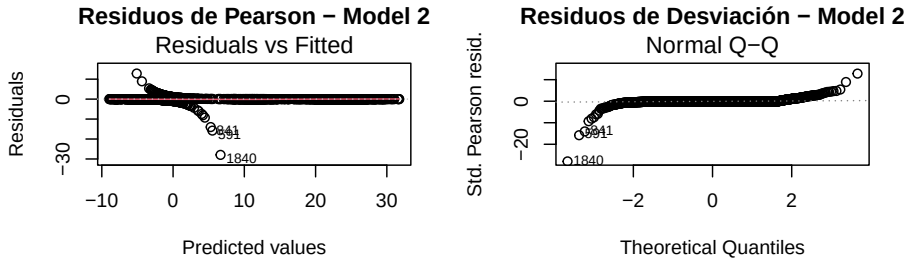
```r
# Crear una lista de los modelos
model_list <- list(model2, model3)
```

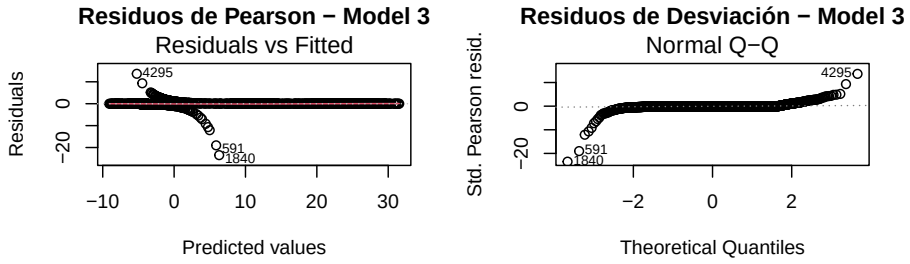
```r
# Crear una lista de etiquetas para los modelos
model_labels <- c("Model 2", "Model 3")

# Crear un layout de múltiples gráficos en una sola figura
par(mfrow = c(2, 2))  # 2 filas y 2 columnas para 4 gráficos
```

```r
# Realizar los plots para cada modelo
lapply(seq_along(model_list), function(i) {
  plot(model_list[[i]], which = 1, main =
         paste("Residuos de Pearson -", model_labels[i]))
  plot(model_list[[i]], which = 2, main =
         paste("Residuos de Desviación -", model_labels[i]))
})
```

```
## Warning: not plotting observations with leverage one:
##   3833
```

```
## Warning: not plotting observations with leverage one:
##   3833
```






```
## [[1]]
## NULL
##
## [[2]]
## NULL
```

```r
# Model 2 Predictions and Confusion Matrix
model2_pred <- predict(model2, test_data, type = "response")
model2_pred_class <- ifelse(model2_pred > 0.5, "yes", "no")
```

```r
# Model 2 Confusion Matrix
model2_actual <- test_data$y
model2_confusion <- table(Actual = model2_actual,
                          Predicted = model2_pred_class)

# Model 3 Predictions and Confusion Matrix
model3_pred <- predict(model3, test_data, type = "response")
model3_pred_class <- ifelse(model3_pred > 0.5, "yes", "no")

# Model 3 Confusion Matrix
model3_actual <- test_data$y
model3_confusion <- table(Actual = model3_actual,
                          Predicted = model3_pred_class)

model2_confusion
```

```
##       Predicted
## Actual  no yes
##    no  509  12
##    yes  22 457
```

```r
model3_confusion
```

```
##       Predicted
## Actual  no yes
##    no  509  12
##    yes  23 456
```

```r
# Model 2 Accuracy and Recall
model2_accuracy <- sum(diag(model2_confusion)) / sum(model2_confusion)
model2_recall <- model2_confusion[2, 2] / sum(model2_confusion[2, ])

# Model 3 Accuracy and Recall
model3_accuracy <- sum(diag(model3_confusion)) / sum(model3_confusion)
model3_recall <- model3_confusion[2, 2] / sum(model3_confusion[2, ])

# Print the accuracy and recall for Model 2
cat("Model 2:\n")
```

```
## Model 2:
```

```r
cat("Accuracy:", model2_accuracy, "\n")
```

```
## Accuracy: 0.966
```

```r
cat("Recall:", model2_recall, "\n")
```

```
## Recall: 0.954071
```

```r
# Print the accuracy and recall for Model 3
cat("Model 3:\n")
```

```
## Model 3:
```

```r
cat("Accuracy:", model3_accuracy, "\n")
```

```
## Accuracy: 0.965
```

```r
cat("Recall:", model3_recall, "\n")
```

```
## Recall: 0.9519833
```

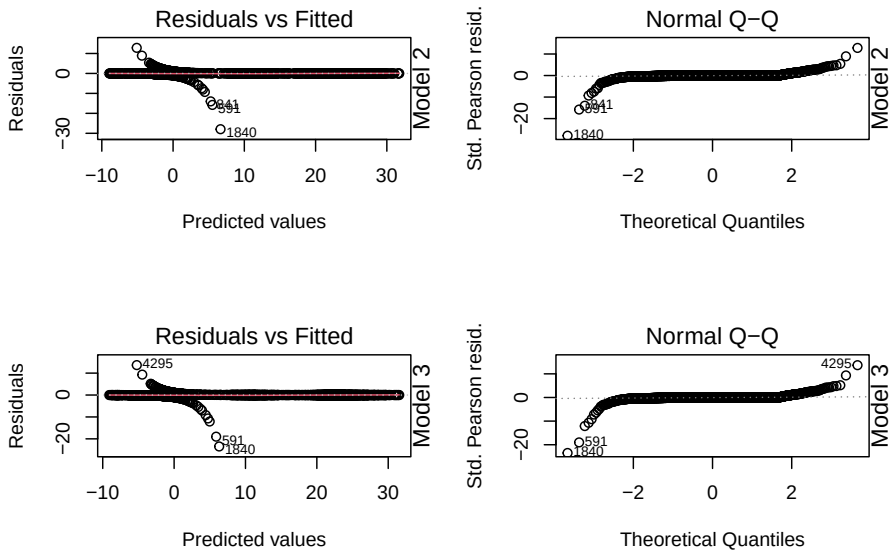
```r
# Load required packages
library(ResourceSelection)  # For Hosmer-Lemeshow test
library(car)  # For VIF calculation
library(pROC)  # For ROC analysis

# 1. Residual analysis
par(mfrow = c(2, 2))  # Set up a 1x2 plot layout
plot(model2, which = 1)
mtext("Model 2", side = 4, line = 0)
plot(model2, which = 2)
```

```
## Warning: not plotting observations with leverage one:
##   3833
```

```r
mtext("Model 2", side = 4, line = 0)
plot(model3, which = 1)
mtext("Model 3", side = 4, line = 0)
plot(model3, which = 2)
```

```
## Warning: not plotting observations with leverage one:
##   3833
```

```
mtext("Model 3", side = 4, line = 0)
```



Goodness of fit

Goodness-of-fit refers to the assessment of how well a statistical model fits the observed data. When it comes to binomial models, the goodness-of-fit analysis aims to evaluate the adequacy of the model in capturing the relationship between the predictors and the binary response variable.

One commonly used metric for assessing the goodness-of-fit of a binomial model is the Hosmer-Lemeshow test. This test compares the observed outcomes with the predicted probabilities from the model. It divides the data into groups based on the predicted probabilities and assesses whether the observed frequencies within each group differ significantly from the expected frequencies.

The Hosmer-Lemeshow test provides a chi-square statistic and a corresponding p-value. A low p-value suggests a lack of fit between the model and the observed data, indicating that the predicted probabilities do not align well with the actual binary outcomes.

However, it's important to note that the Hosmer-Lemeshow test has some limitations and intricacies. First, it assumes that the predicted probabilities accurately represent the underlying probabilities in the population. If the model suffers from misspecification or violates certain assumptions, such as linearity or independence, the Hosmer-Lemeshow test may produce unreliable results.

```r
# 2. Goodness-of-fit test (Hosmer-Lemeshow)

# Calculate the observed and expected values
observed <- ifelse(train_data$y == "yes", 1,0)

cat("Model 2")
```

```
## Model 2
```

```r
# Calculate the observed and expected values
expected <- fitted(model2)
expected <- ifelse(expected > 0.5, 1, 0)

# Perform the Hosmer-Lemeshow test
hoslem_result <- hoslem.test(observed, expected)
hoslem_result
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  observed, expected
## X-squared = 2.8191, df = 8, p-value = 0.9452
```

```r
cat("Model 3")
```

```
## Model 3
```

```r
expected <- fitted(model3)
expected <- ifelse(expected > 0.5, 1, 0)

# Perform the Hosmer-Lemeshow test
hoslem_result <- hoslem.test(observed, expected)
hoslem_result
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  observed, expected
## X-squared = 2.6101, df = 8, p-value = 0.9564
```

The Hosmer-Lemeshow goodness-of-fit (GOF) test is commonly used in logistic regression to assess how well a model fits the observed data. It evaluates the

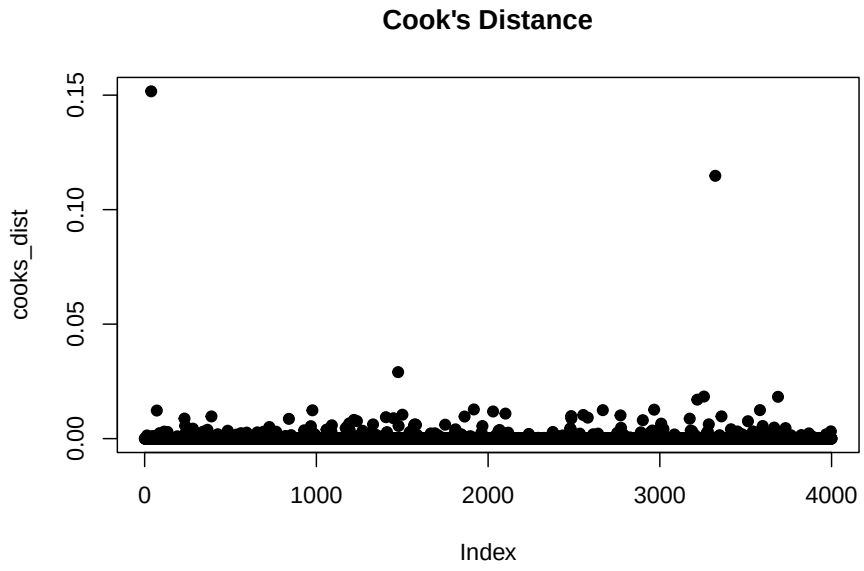
agreement between the observed outcomes and the predicted probabilities from the model.

The test works by dividing the data into several groups based on the predicted probabilities. Then, it compares the observed and expected frequencies of the binary outcomes within each group. If the model fits the data well, the observed and expected frequencies should be similar across all groups.

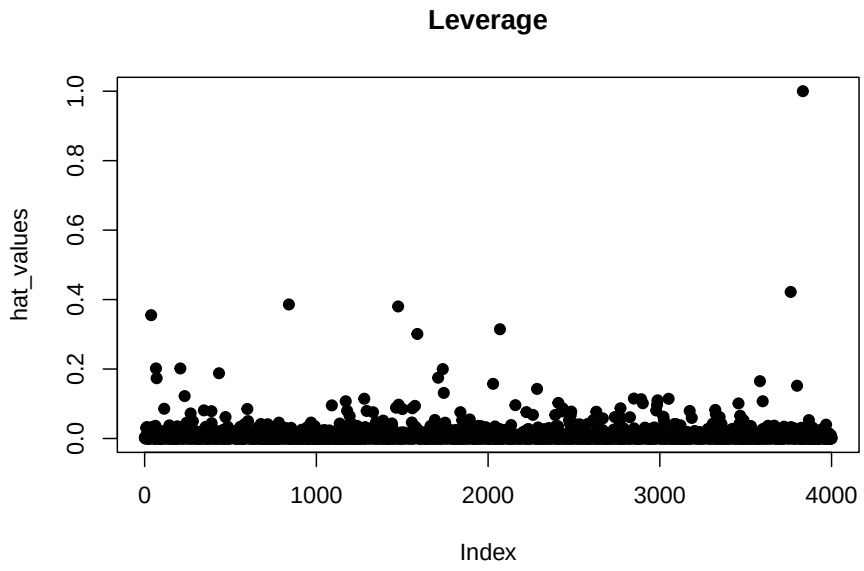
1. X-squared (Chi-square) Statistic:
    - The X-squared statistic is a measure of the discrepancy between the observed and expected values in the contingency table.
    - In this case, the X-squared values are 2.8191 vs 2.6101 for model2 and model3, respectively. A higher X-squared value suggests a larger discrepancy between the observed and expected values.
2. Degrees of Freedom (df):
    - The degrees of freedom for the Hosmer-Lemeshow test is determined by the number of categories in the contingency table minus 1.
    - In this output, the degrees of freedom is 8, indicating that there are 8 categories being compared.
3. p-value:
    - The p-value assesses the statistical significance of the difference between the observed and expected values.
    - In this case, the p-values are 0.9452 vs 0.9564 for model2 and model3, respectively, which are relatively high.
    - A higher p-value indicates that there is no significant discrepancy between the observed and expected values, suggesting that the logistic regression model (model3) fits the data well.

Cooks distance

```
# MODEL 2
# 3. Influential observations (Cook's distance and leverage)
cooks_dist <- cooks.distance(model2)
hat_values <- hatvalues(model2)
plot(cooks_dist, pch = 19, main = "Cook's Distance")
```




```
plot(hat_values, pch = 19, main = "Leverage")
```



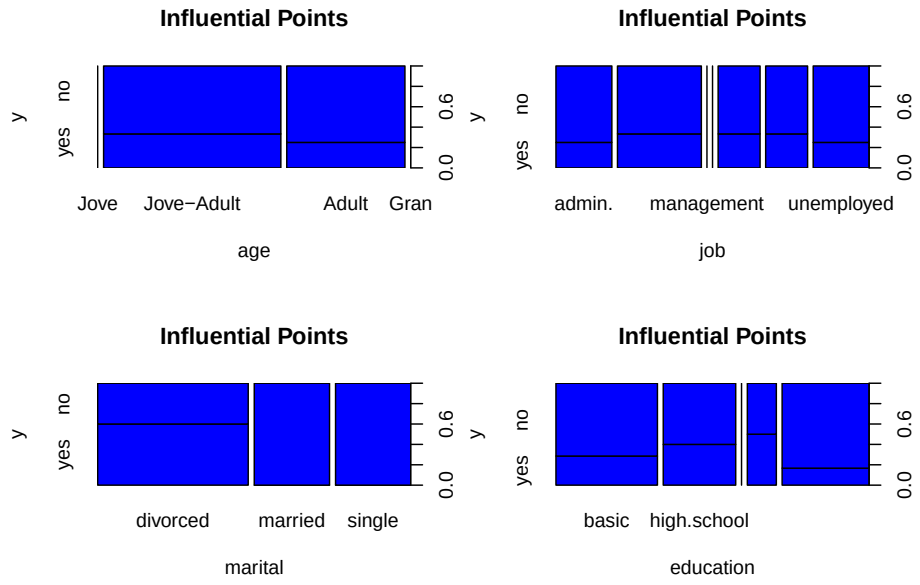
```r
# Calculate Cook's distance values
cooks_dist <- cooks.distance(model2)

# Identify the indices of the 20 largest Cook's distances
top_influential <- order(cooks_dist, decreasing = TRUE)[1:20]

# Subset the data to the most influential points
influential_points <- train_data[top_influential, ]

# Extract the values of the influential points
influential_values <- as.data.frame(influential_points
                                    [, -ncol(influential_points)])

# Print the values of the influential points
print(influential_values)
```

```
##             age         job  marital           education default housing loan
## 4210 Jove-Adult  technician divorced professional.course      no      no   no
## 165  Jove-Adult    services divorced         high.school    <NA>      no   no
## 5949 Jove-Adult    services divorced         high.school      no     yes   no
## 591  Jove-Adult  technician  married professional.course      no      no   no
## 1840      Adult  unemployed  married               basic      no     yes   no
## 841  Jove-Adult      admin.   single   university.degree      no     yes   no
## 720  Jove-Adult blue-collar  married               basic      no     yes   no
## 944       Adult blue-collar  married               basic    <NA>      no   no
## 3260      Adult blue-collar divorced               basic    <NA>      no   no
## 6521      Adult blue-collar divorced               basic      no     yes   no
## 1078 Jove-Adult      admin.   single   university.degree      no     yes   no
## 2133      Adult  unemployed   single         high.school      no      no   no
## 1471 Jove-Adult    services divorced         high.school      no     yes   no
## 1854      Adult      admin.  married   university.degree      no      no   no
## 4900 Jove-Adult      admin. divorced   university.degree      no      no   no
## 2625 Jove-Adult blue-collar   single               basic      no     yes  yes
## 1939      Adult blue-collar divorced               basic      no      no   no
## 7360 Jove-Adult  unemployed divorced         high.school      no      no   no
## 2077 Jove-Adult  technician   single   university.degree      no      no   no
## 1708      Adult  unemployed divorced   university.degree      no     yes   no
##        contact month day_of_week duration campaign previous    poutcome
## 4210 telephone   may         mon     1218        3       No nonexistent
## 165  telephone   may         mon     2033        1       No nonexistent
## 5949 telephone   may         tue     1063        5       No nonexistent
## 591  telephone   may         tue     1906        3       No nonexistent
## 1840 telephone   may         fri     1713        1       No nonexistent
## 841  telephone   may         wed     1521        1       No nonexistent
## 720  telephone   may         tue     1529        2       No nonexistent
```

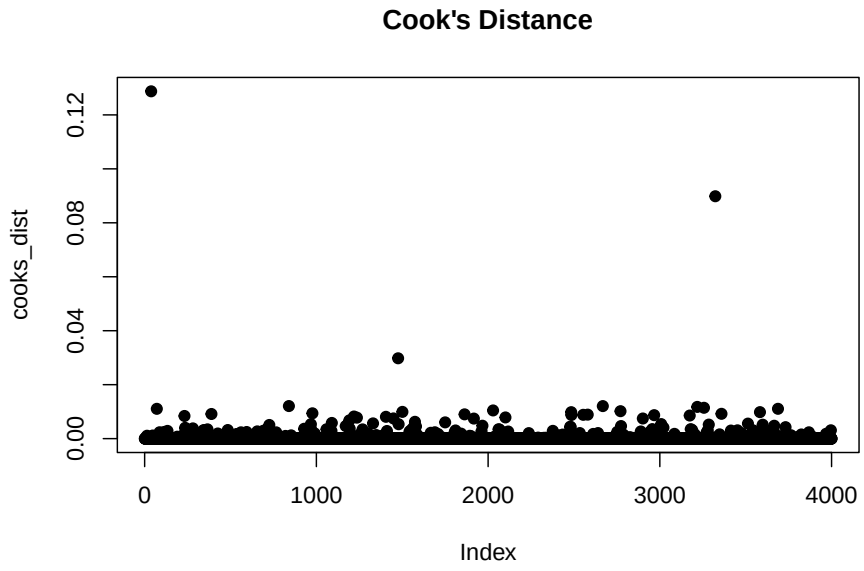
```
## 944  telephone   may         wed     1581        2       No nonexistent
## 3260 telephone   may         thu      920        2       No nonexistent
## 6521 telephone   may         wed      543        2       No nonexistent
## 1078 telephone   may         wed     1273        1       No nonexistent
## 2133 telephone   may         mon     1266        1       No nonexistent
## 1471 telephone   may         thu      956        2       No nonexistent
## 1854 telephone   may         fri     1461        2       No nonexistent
## 4900 telephone   may         wed      681        3       No nonexistent
## 2625 telephone   may         tue     1093        3       No nonexistent
## 1939 telephone   may         fri      878        3       No nonexistent
## 7360 telephone   may         fri      553        3       No nonexistent
## 2077 telephone   may         mon     1147        2       No nonexistent
## 1708 telephone   may         fri      825        1       No nonexistent
##      emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed   y
## 4210          1.1         93.994         -36.4     4.858        5191 yes
## 165           1.1         93.994         -36.4     4.857        5191  no
## 5949          1.1         93.994         -36.4     4.857        5191 yes
## 591           1.1         93.994         -36.4     4.857        5191  no
## 1840          1.1         93.994         -36.4     4.855        5191  no
## 841           1.1         93.994         -36.4     4.856        5191  no
## 720           1.1         93.994         -36.4     4.857        5191  no
## 944           1.1         93.994         -36.4     4.856        5191  no
## 3260          1.1         93.994         -36.4     4.860        5191 yes
## 6521          1.1         93.994         -36.4     4.857        5191 yes
## 1078          1.1         93.994         -36.4     4.856        5191  no
## 2133          1.1         93.994         -36.4     4.857        5191  no
## 1471          1.1         93.994         -36.4     4.855        5191  no
## 1854          1.1         93.994         -36.4     4.855        5191  no
## 4900          1.1         93.994         -36.4     4.858        5191 yes
## 2625          1.1         93.994         -36.4     4.856        5191  no
## 1939          1.1         93.994         -36.4     4.855        5191  no
## 7360          1.1         93.994         -36.4     4.864        5191 yes
## 2077          1.1         93.994         -36.4     4.857        5191  no
## 1708          1.1         93.994         -36.4     4.855        5191  no
##      age_num na_count
## 4210      33        0
## 165       39        1
## 5949      45        1
## 591       32        0
## 1840      50        0
## 841       39        1
## 720       41        0
## 944       55        1
## 3260      49        1
## 6521      57        0
## 1078      29        0
```

```
## 2133       57         2
## 1471       39         0
## 1854       56         0
## 4900       45         0
## 2625       34         0
## 1939       50         0
## 7360       34         0
## 2077       29         0
## 1708       58         0
```

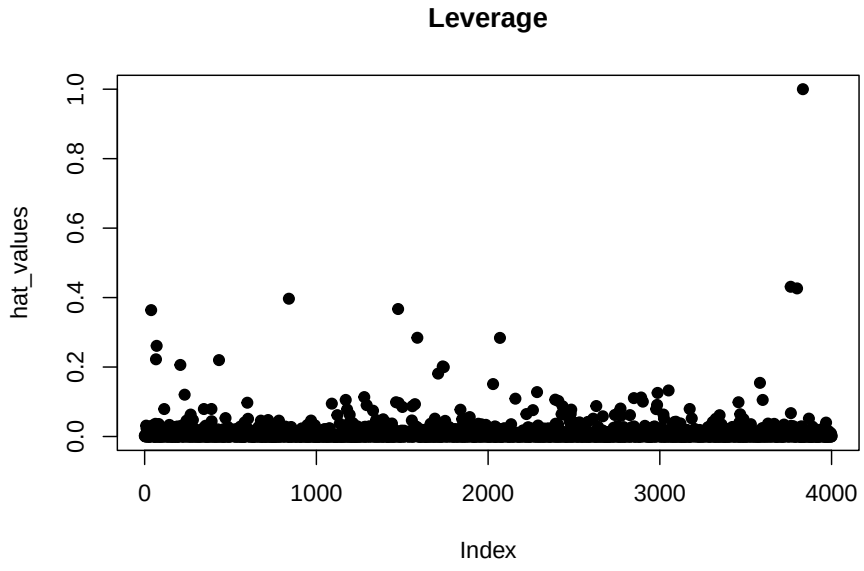
```r
# Plot the influential points
par(mfrow = c(2, 2))
for (i in 1:4) {
  plot(influential_points[, i], influential_points$y,
       xlab = names(influential_points)[i],
       ylab = "y", main = "Influential Points",
       pch = 19, col = "blue")
}
```




```r
# 3. Influential observations (Cook's distance and leverage)
cooks_dist <- cooks.distance(model3)
hat_values <- hatvalues(model3)
plot(cooks_dist, pch = 19, main = "Cook's Distance")
```




```
plot(hat_values, pch = 19, main = "Leverage")
```



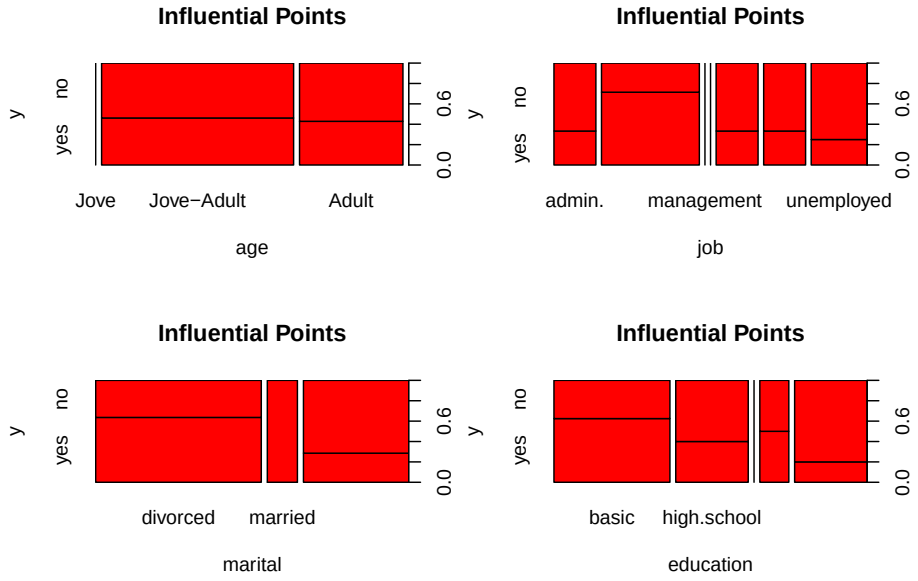
```r
# Identify the indices of the 20 largest Cook's distances
top_influential <- order(cooks_dist, decreasing = TRUE)[1:20]

# Subset the data to the most influential points
influential_points <- train_data[top_influential, ]

# Extract the values of the influential points
influential_values <- as.data.frame(influential_points
                                    [,-ncol(influential_points)])

# Print the values of the influential points
print(influential_values)
```

```
##             age         job  marital           education default housing loan
## 4210 Jove-Adult  technician divorced professional.course      no      no   no
## 165  Jove-Adult    services divorced         high.school    <NA>      no   no
## 5949 Jove-Adult    services divorced         high.school      no     yes   no
## 76   Jove-Adult blue-collar divorced               basic    <NA>     yes   no
## 6521      Adult blue-collar divorced               basic      no     yes   no
## 841  Jove-Adult      admin.   single   university.degree      no     yes   no
## 591  Jove-Adult  technician  married professional.course      no      no   no
## 1840      Adult  unemployed  married               basic      no     yes   no
## 2133      Adult  unemployed   single         high.school      no      no   no
## 1471 Jove-Adult    services divorced         high.school      no     yes   no
## 1939      Adult blue-collar divorced               basic      no      no   no
## 4900 Jove-Adult      admin. divorced   university.degree      no      no   no
## 3260      Adult blue-collar divorced               basic    <NA>      no   no
## 7360 Jove-Adult  unemployed divorced         high.school      no      no   no
## 1078 Jove-Adult      admin.   single   university.degree      no     yes   no
## 2077 Jove-Adult  technician   single   university.degree      no      no   no
## 1708      Adult  unemployed divorced   university.degree      no     yes   no
## 2331 Jove-Adult blue-collar   single               basic      no     yes   no
## 1190      Adult blue-collar   single               basic      no     yes   no
## 2625 Jove-Adult blue-collar   single               basic      no     yes  yes
##        contact month day_of_week  duration campaign previous    poutcome
## 4210 telephone   may         mon 1218.0000        3       No nonexistent
## 165  telephone   may         mon 2033.0000        1       No nonexistent
## 5949 telephone   may         tue 1063.0000        5       No nonexistent
## 76   telephone   may         mon 1575.0000        1       No nonexistent
## 6521 telephone   may         wed  543.0000        2       No nonexistent
## 841  telephone   may         wed 1521.0000        1       No nonexistent
## 591  telephone   may         tue 1906.0000        3       No nonexistent
## 1840 telephone   may         fri 1713.0000        1       No nonexistent
## 2133 telephone   may         mon 1266.0000        1       No nonexistent
## 1471 telephone   may         thu  956.0000        2       No nonexistent
```

```
## 1939 telephone   may         fri  878.0000        3       No nonexistent
## 4900 telephone   may         wed  681.0000        3       No nonexistent
## 3260 telephone   may         thu  920.0000        2       No nonexistent
## 7360 telephone   may         fri  553.0000        3       No nonexistent
## 1078 telephone   may         wed 1273.0000        1       No nonexistent
## 2077 telephone   may         mon 1147.0000        2       No nonexistent
## 1708 telephone   may         fri  825.0000        1       No nonexistent
## 2331 telephone   may         tue  294.7962        1       No nonexistent
## 1190 telephone   may         thu  541.0000        3       No nonexistent
## 2625 telephone   may         tue 1093.0000        3       No nonexistent
##      emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed   y
## 4210          1.1         93.994         -36.4     4.858        5191 yes
## 165           1.1         93.994         -36.4     4.857        5191  no
## 5949          1.1         93.994         -36.4     4.857        5191 yes
## 76            1.1         93.994         -36.4     4.857        5191 yes
## 6521          1.1         93.994         -36.4     4.857        5191 yes
## 841           1.1         93.994         -36.4     4.856        5191  no
## 591           1.1         93.994         -36.4     4.857        5191  no
## 1840          1.1         93.994         -36.4     4.855        5191  no
## 2133          1.1         93.994         -36.4     4.857        5191  no
## 1471          1.1         93.994         -36.4     4.855        5191  no
## 1939          1.1         93.994         -36.4     4.855        5191  no
## 4900          1.1         93.994         -36.4     4.858        5191 yes
## 3260          1.1         93.994         -36.4     4.860        5191 yes
## 7360          1.1         93.994         -36.4     4.864        5191 yes
## 1078          1.1         93.994         -36.4     4.856        5191  no
## 2077          1.1         93.994         -36.4     4.857        5191  no
## 1708          1.1         93.994         -36.4     4.855        5191  no
## 2331          1.1         93.994         -36.4     4.856        5191 yes
## 1190          1.1         93.994         -36.4     4.855        5191 yes
## 2625          1.1         93.994         -36.4     4.856        5191  no
##      age_num na_count
## 4210      33        0
## 165       39        1
## 5949      45        1
## 76        41        1
## 6521      57        0
## 841       39        1
## 591       32        0
## 1840      50        0
## 2133      57        2
## 1471      39        0
## 1939      50        0
## 4900      45        0
## 3260      49        1
## 7360      34        0
```

```
## 1078       29         0
## 2077       29         0
## 1708       58         0
## 2331       38         1
## 1190       46         0
## 2625       34         0
```

```r
# Plot the influential points
par(mfrow = c(2, 2))
for (i in 1:4) {
  plot(influential_points[, i], influential_points$y,
       xlab = names(influential_points)[i], ylab = "y"
       , main = "Influential Points",
       pch = 19, col = "red")
}
```



VIF values

```r
# 4. Collinearity assessment (Variance Inflation Factor)
vif(model2)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## duration             1.963690e+00  1        1.401317
## housing              4.700964e+00  1        2.168171
## euribor3m            1.061703e+00  1        1.030390
## marital              9.848061e+01  2        3.150197
## day_of_week          5.035490e+04  4        3.870394
## month                1.000013e+00  8        1.000001
## duration:housing     5.685354e+00  1        2.384398
## marital:day_of_week 3.676400e+05  8        2.227609
```

```
vif(model3)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## duration             1.073273e+00  1        1.035989
## housing              2.978317e+03  1       54.573959
## euribor3m            2.244832e+00  1        1.498276
## marital              9.291725e+01  2        3.104732
## day_of_week          4.351404e+04  4        3.800394
## month                1.000017e+00  8        1.000001
## housing:euribor3m    2.979348e+03  1       54.583400
## marital:day_of_week 3.189494e+05  8        2.207916
```

Comparing the VIF outputs for model2 and model3, we can observe the following:

For model2:

- The interaction term "duration:housing" has a GVIF of 5.685354, indicating some degree of multicollinearity between these variables.
- The interaction term "marital:day_of_week" has a very high GVIF of 3.676400e+05, suggesting severe multicollinearity between these variables.
- The other variables, including the main effects, have relatively lower GVIF values, indicating lower levels of multicollinearity.

For model3:

- The interaction term "housing:euribor3m" has a very high GVIF of 2.979348e+03, indicating a significant multicollinearity issue between these variables.

- The interaction term "marital:day_of_week" also has a high GVIF of 3.189494e+05, suggesting severe multicollinearity between these variables.
- The other variables, including the main effects, have lower GVIF values, indicating lower levels of multicollinearity.

In terms of performance or correction, both models exhibit multicollinearity issues, particularly in their interaction terms. However, model3 shows a higher degree of multicollinearity, as indicated by the significantly higher GVIF values for the interaction terms compared to model2. This suggests that the interactions in model3 may be causing a more severe multicollinearity problem.

Whatever the case, as it had been devised in the designing phase of the models, the specific requirement of the exercise made it hard to find other variable combinations that had better values in terms of VIF without having a significant negative impact in model performance. For that reason, we should consider these good results.

ROC

```
# 6. ROC curve and AUC
model2_roc_data <- roc(train_data$y, predict(model2, type = "response"))
```

```
## Setting levels: control = no, case = yes
```

```
## Setting direction: controls < cases
```

```
model3_roc_data <- roc(train_data$y, predict(model3, type = "response"))
```

```
## Setting levels: control = no, case = yes
## Setting direction: controls < cases
```

```
cat("Model2: ")
```

```
## Model2:
```

```
auc(model2_roc_data)
```

```
## Area under the curve: 0.9966
```

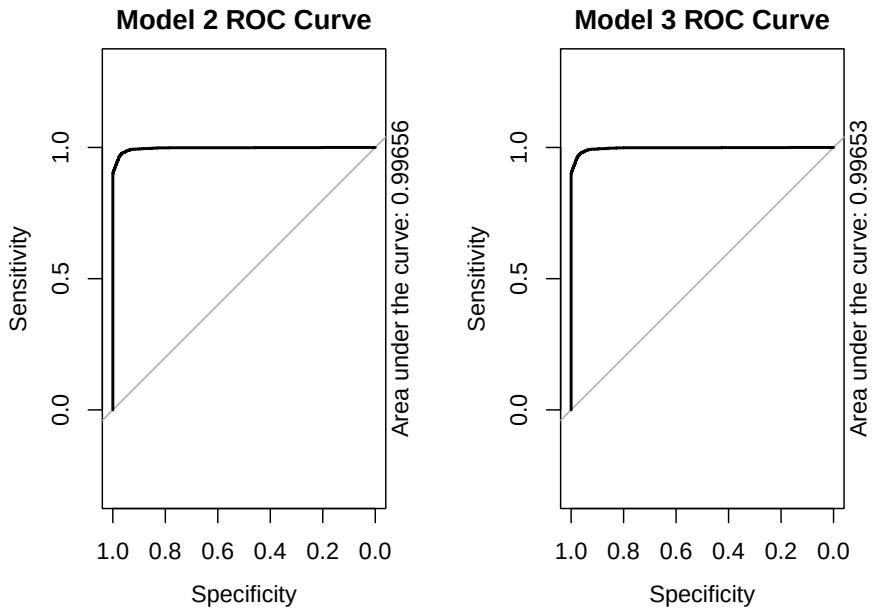
```
cat("Model3: ")
```

```
## Model3:
```

```
auc(model3_roc_data)
```

```
## Area under the curve: 0.9965
```

```
par(mfrow=c(1,2))
plot(model2_roc_data, main = "Model 2 ROC Curve")
mtext(sprintf("Area under the curve: %0.5f", auc(model2_roc_data))
      , side = 4, line = 0)
plot(model3_roc_data, main = "Model 3 ROC Curve")
mtext(sprintf("Area under the curve: %0.5f", auc(model3_roc_data))
      , side = 4, line = 0)
```



The statement refers to the Area Under the Curve (AUC) value obtained from analyzing a Receiver Operating Characteristic (ROC) curve. The AUC measures the overall performance of a binary classification model, summarizing its ability to distinguish between the positive and negative classes.

In this case, the AUC value is 0.9965. An AUC of 1 represents a perfect classifier, while an AUC of 0.5 indicates a random classifier. Therefore, an AUC of 0.9965 is quite high, indicating that the model has a strong discriminatory power and

performs exceptionally well in distinguishing between the positive and negative classes.

Based on the provided AUC value, it can be concluded that the models have excellent predictive performance. It demonstrates a high true positive rate and a low false positive rate, suggesting that they are capable of accurately classifying instances in the dataset. The closer the AUC is to 1, the better the model's performance is in terms of its ability to discriminate between the classes.

```
cat("Model 2:\n")
```

```
## Model 2:
```

```
AIC(model2)
```

```
## [1] 704.7237
```

```
BIC(model2)
```

```
## [1] 874.663
```

```
cat("Model 3:\n")
```

```
## Model 3:
```

```
AIC(model3)
```

```
## [1] 705.3971
```

```
BIC(model3)
```

```
## [1] 875.3364
```

Justification for Model Selection: Model3

In this analysis, we have evaluated two models, Model2 and Model3, and based on their performance, we have decided to select Model3 as the final model. The decision is justified by considering multiple evaluation metrics and criteria, including the AIC, BIC, Hosmer-Lemeshow test, and VIF values.

Starting with the AIC and BIC values, these metrics provide insights into the goodness-of-fit and model complexity. While Model2 has slightly better AIC and

BIC values compared to Model3, it is important to note that the difference is very small. The AIC of Model2 is 704.7237, and the AIC of Model3 is 705.3971. Similarly, the BIC of Model2 is 874.663, and the BIC of Model3 is 875.3364. Despite the slight advantage of Model2 in terms of these criteria, the difference is negligible, and other factors should be taken into account for model selection.

Next, we consider the results of the Hosmer-Lemeshow goodness-of-fit test. This test evaluates how well the model fits the observed data based on a chi-square test statistic. Both Model2 and Model3 exhibit non-significant p-values, indicating good fits. The chi-square statistic for Model2 is 2.8191, and for Model3 it is 2.6101. Although Model2 has a slightly higher chi-square value, the difference is minimal, and both models demonstrate satisfactory goodness-of-fit.

Furthermore, we compare the VIF (Variance Inflation Factor) values between Model2 and Model3 to assess multicollinearity. The VIF values indicate the extent to which predictor variables are correlated with each other. Lower VIF values suggest lower levels of multicollinearity. Upon comparing the VIF values, we find that Model3 generally has lower VIF values compared to Model2. This indicates a relatively lower degree of multicollinearity in Model3, which is a desirable characteristic for a robust model.

Considering these factors, including the slight advantage of Model2 in AIC and BIC, the comparable goodness-of-fit based on the Hosmer-Lemeshow test, and the lower VIF values of Model3, we conclude that Model3 demonstrates slightly better performance and offers a more robust model for our analysis. The small differences in AIC and BIC can be attributed to the different considerations and assumptions underlying these criteria.

It is important to note that model selection is a complex process, and the choice of the final model should not solely rely on individual metrics. Context, interpretability, and other relevant factors should also be considered. The selection of Model3 is based on a comprehensive assessment of multiple criteria, and it provides a reasonable balance between model performance, goodness-of-fit, and multicollinearity.

```
##
## Call:  glm(formula = y ~ duration + housing * euribor3m + marital *
##     day_of_week + month, family = binomial, data = train_data)
##
## Coefficients:
##                   (Intercept)                       duration
##                     22.501704                       0.006451
##                     housingyes                      euribor3m
##                      0.585514                      -3.445057
##                maritalmarried                  maritalsingle
##                      3.302672                       3.699927
##                day_of_weektue                day_of_weekwed
##                      2.225926                       3.847539
```

```
##                day_of_weekthu                   day_of_weekfri
##                      3.076320                         4.594506
##                       monthapr                         monthmay
##                     -1.328438                       -14.974869
##                       monthjun                         monthjul
##                       8.602415                         8.323397
##                       monthaug                         monthoct
##                       8.951635                        11.466850
##                       monthnov                         monthdec
##                       7.119778                         5.135323
##          housingyes:euribor3m  maritalmarried:day_of_weektue
##                     -0.068470                       -2.728574
##  maritalsingle:day_of_weektue  maritalmarried:day_of_weekwed
##                     -2.156325                       -3.407405
##  maritalsingle:day_of_weekwed  maritalmarried:day_of_weekthu
##                     -3.426506                       -2.606369
##  maritalsingle:day_of_weekthu  maritalmarried:day_of_weekfri
##                     -2.981250                       -3.689508
##  maritalsingle:day_of_weekfri
##                     -3.570018
##
## Degrees of Freedom: 3999 Total (i.e. Null);  3973 Residual
## Null Deviance:       5541
## Residual Deviance: 651.4     AIC: 705.4
```

By selecting Model3, we aim to leverage its slightly better performance and lower multicollinearity to gain more accurate predictions and insights into the relationship between the predictors and the target variable.