Simple predictive models: Linear and logistic regression

Montserrat Guillen

6 de diciembre 2017

Contents

We have previously analysed the data. Just recall that the data contain 41188 cases and 21 variables. The variable names are: age, job, marital, education, default, housing, loan, contact, month, day_of_week, duration, campaign, pdays, previous, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed, y.

Linear regression

We will study the duration of the telephone call as a function of age.

Linear model (quantitative regressors)

We introduce two variables: **age** and **euribor.3m**.

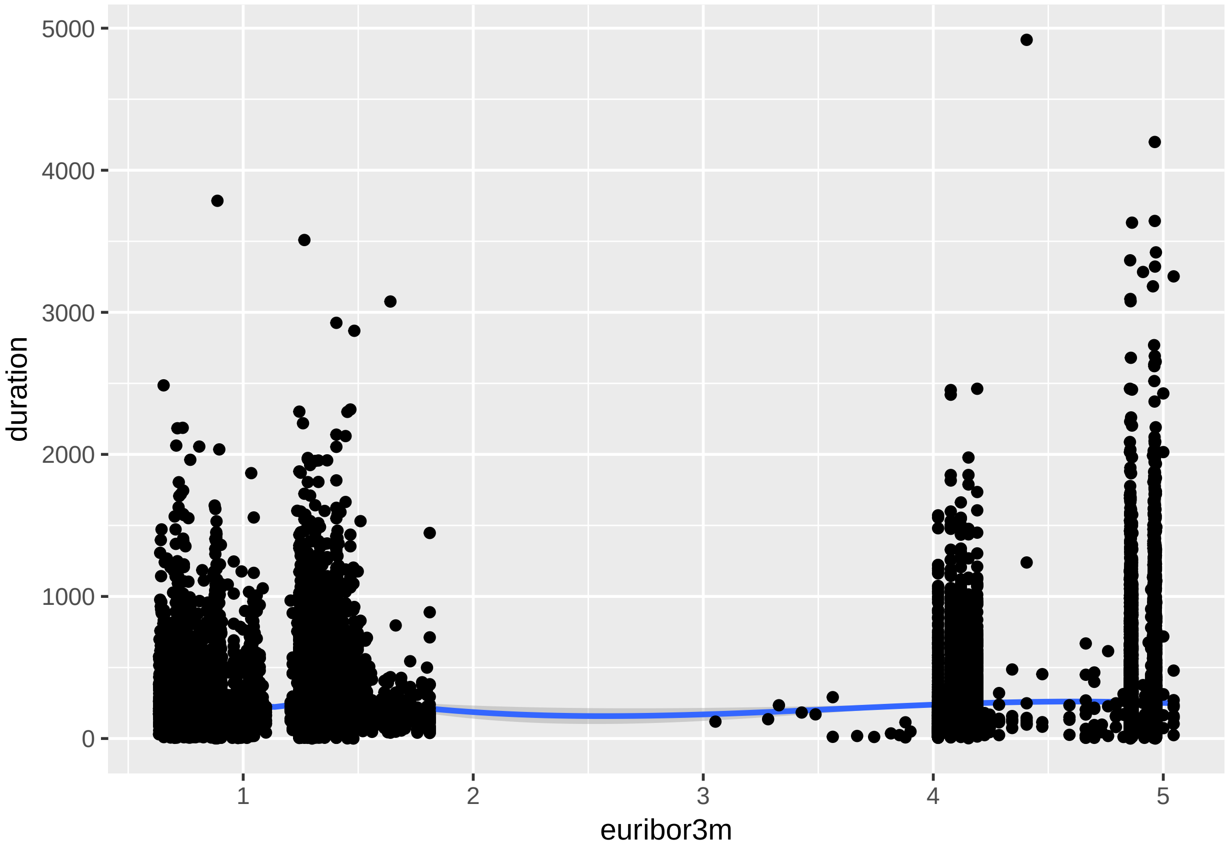
```r
# Model estimation
attach(mydata)

Model.1.1<- lm(duration~age+ euribor3m, data=mydata )
summary(Model.1.1)
```

```
##
## Call:
## lm(formula = duration ~ age + euribor3m, data = mydata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -270.8 -155.5  -78.5   60.1 4663.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 276.59970    5.70184  48.511  < 2e-16 ***
## age          -0.01273    0.12254  -0.104    0.917
## euribor3m    -4.91684    0.73625  -6.678 2.45e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 259.1 on 41185 degrees of freedom
## Multiple R-squared:  0.001082,   Adjusted R-squared:  0.001034
## F-statistic: 22.31 on 2 and 41185 DF,  p-value: 2.061e-10
```

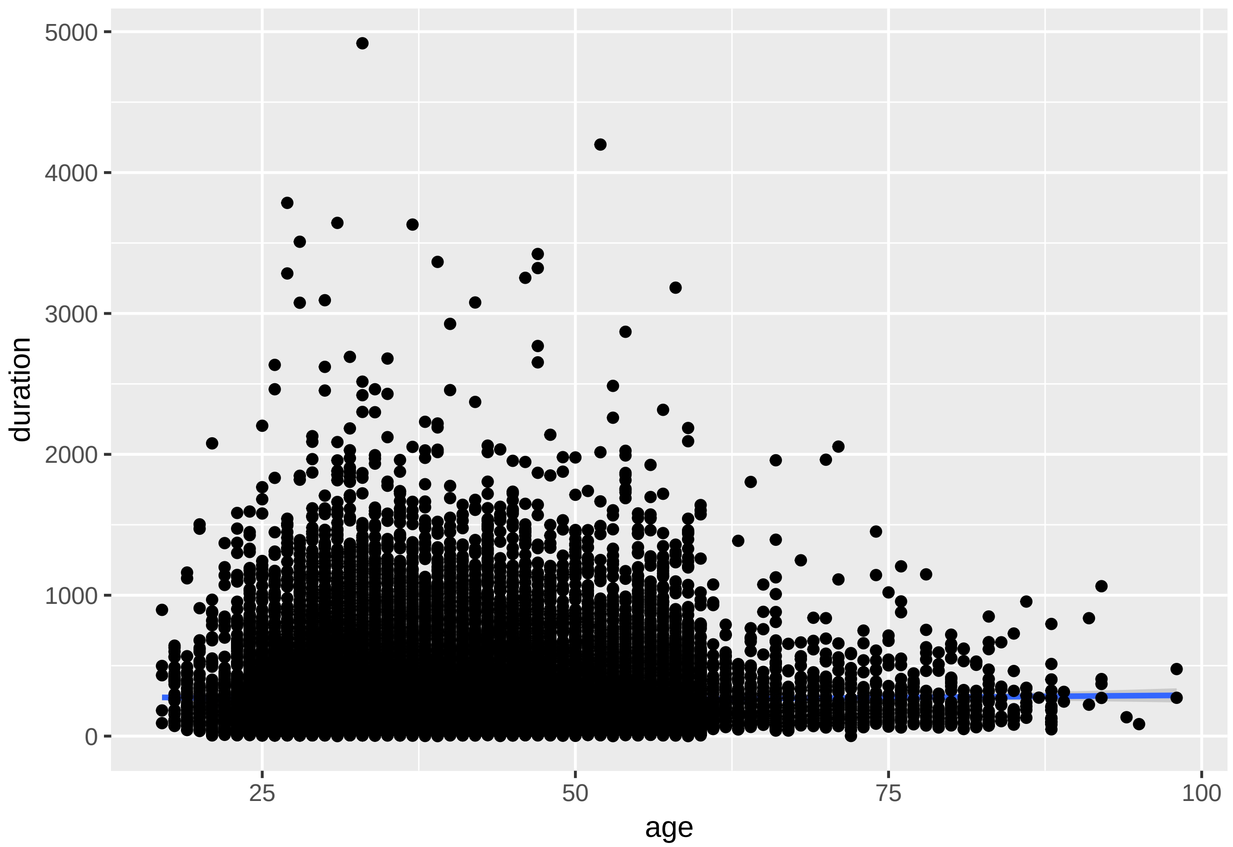
```
qplot(euribor3m,duration, data = mydata,geom = c("smooth", "point"))
```

```
## `geom_smooth()` using method = 'gam'
```



```
qplot(age,duration, data = mydata,geom = c("smooth", "point"))
```

```
## `geom_smooth()` using method = 'gam'
```



The goodness-of-fit coefficient is 0.0010825

Linear model (quantitative and qualitative regressors)

We now also include **month**, **day_of_week** and **contact**

```
monthR=relevel(month, ref = 'mar')
day_of_weekR=relevel(day_of_week, ref = 'mon')
contactR=relevel(contact, ref = 'telephone')

Model.1.2<- lm(duration~age+ euribor3m+factor(monthR)+factor(day_of_weekR)+factor(contactR), data=mydat
summary(Model.1.2)
```

```
##
## Call:
## lm(formula = duration ~ age + euribor3m + factor(monthR) + factor(day_of_weekR) +
##     factor(contactR), data = mydata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -340.7 -154.0  -77.0   58.8 4704.0
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               211.91697   13.30118  15.932  < 2e-16 ***
## age                         0.06294    0.12351   0.510 0.610337
## euribor3m                   0.99360    1.13216   0.878 0.380159    
## factor(monthR)apr          46.62301   12.17291   3.830 0.000128 ***
## factor(monthR)aug         -19.58197   12.13000  -1.614 0.106461    
## factor(monthR)dec         100.67585   22.14814   4.546 5.49e-06 ***
## factor(monthR)jul          23.33980   12.12062   1.926 0.054158 .  
## factor(monthR)jun           8.82025   11.98862   0.736 0.461906    
## factor(monthR)may          22.36805   11.46354   1.951 0.051036 .  
## factor(monthR)nov          -4.40525   12.13417  -0.363 0.716573    
## factor(monthR)oct          42.88593   14.69139   2.919 0.003512 ** 
## factor(monthR)sep          50.88047   15.49524   3.284 0.001026 ** 
## factor(day_of_weekR)fri     7.36957    4.05461   1.818 0.069136 .  
## factor(day_of_weekR)thu    18.69792    3.95446   4.728 2.27e-06 ***
## factor(day_of_weekR)tue    16.05568    4.02523   3.989 6.65e-05 ***
## factor(day_of_weekR)wed    21.17355    4.02114   5.266 1.40e-07 ***
## factor(contactR)cellular   21.14074    4.20038   5.033 4.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 258.4 on 41171 degrees of freedom
## Multiple R-squared:  0.006776,   Adjusted R-squared:  0.00639 
## F-statistic: 17.55 on 16 and 41171 DF,  p-value: < 2.2e-16
```

The goodness-of-fit coefficient is in the first model 0.001 and in the second model 0.0064.

Prediction

Assume we have a new observation and want to predict the duration.

```
newdata=data.frame(age=30, euribor3m=1.0, monthR='jun', day_of_weekR='fri', contactR='cellular')
predict(Model.1.1, newdata)
```

```
##       1 
## 271.301
```

```
predict(Model.1.2, newdata)
```

```
##        1 
## 252.1293
```

Logistic regression model

Estimation of the model

We estimate the model for the dependent variable **y** = **Term Diposit**

```
Model.2.1=glm(y~age+euribor3m, family=binomial)
summary(Model.2.1)
```

```
## 
## Call:
```

```
## glm(formula = y ~ age + euribor3m, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0056  -0.3953  -0.3010  -0.2857   2.5801
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.801623   0.062184  -12.89  < 2e-16 ***
## age          0.008145   0.001371    5.94 2.85e-09 ***
## euribor3m   -0.536241   0.009540  -56.21  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28999  on 41187  degrees of freedom
## Residual deviance: 25308  on 41185  degrees of freedom
## AIC: 25314
##
## Number of Fisher Scoring iterations: 5
```

Prediction with this model

```r
predict(Model.2.1, newdata, type="response")
```

```
##         1
## 0.2509548
```

The prediction for that custmer and the logistic model is 0.25.

Improve the model

We can improve the model now with more information

```r
Model.2.2=glm(y~age+euribor3m+factor(day_of_weekR), family=binomial)
summary(Model.2.2)
```

```
##
## Call:
## glm(formula = y ~ age + euribor3m + factor(day_of_weekR), family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0542  -0.4029  -0.3069  -0.2780   2.6561
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.003265   0.070895 -14.151  < 2e-16 ***
## age                      0.008226   0.001372   5.995 2.03e-09 ***
## euribor3m               -0.539354   0.009562 -56.404  < 2e-16 ***
## factor(day_of_weekR)fri  0.125124   0.053762   2.327   0.0199 *
## factor(day_of_weekR)thu  0.276230   0.051449   5.369 7.92e-08 ***
## factor(day_of_weekR)tue  0.302954   0.052632   5.756 8.61e-09 ***
## factor(day_of_weekR)wed  0.319297   0.052673   6.062 1.35e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28999  on 41187  degrees of freedom
## Residual deviance: 25252  on 41181  degrees of freedom
## AIC: 25266
##
## Number of Fisher Scoring iterations: 5
```

The Akaike Information Criterion (AIC) in the first model was 25314 ans now it is 25266.

ROC curve

Predictive performance

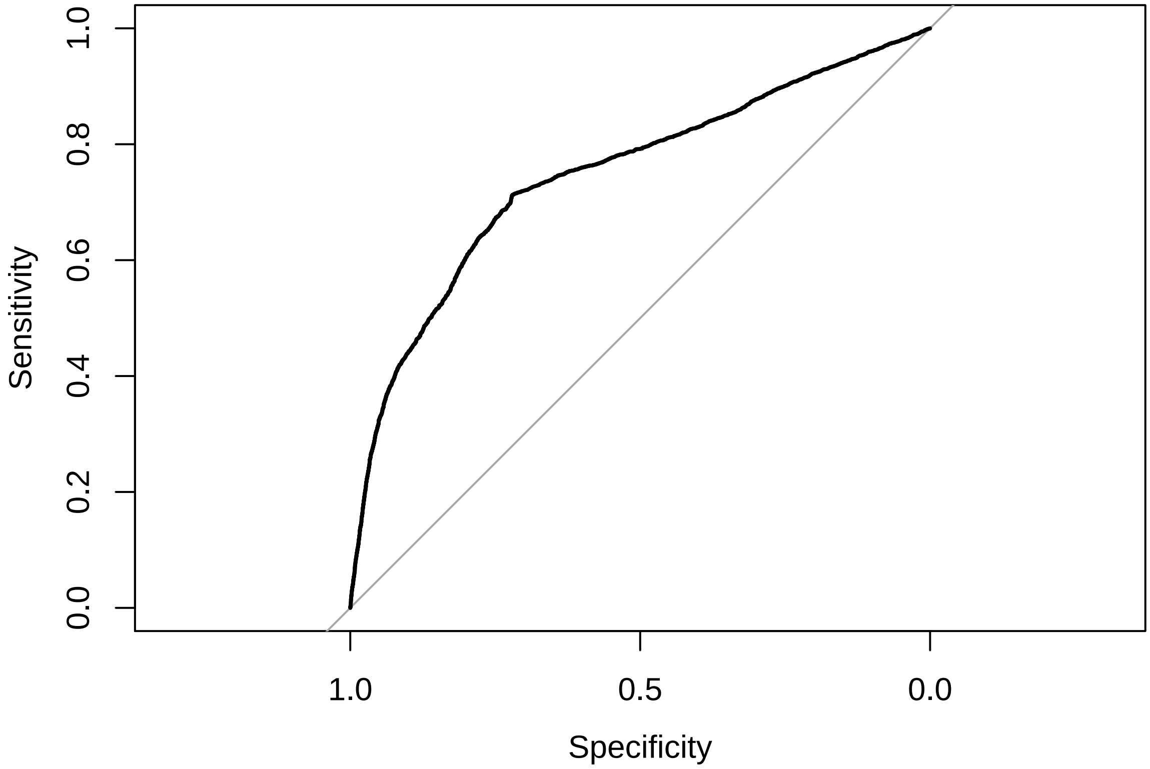
```
#install.packages("pROC")
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
prob=predict(Model.2.2,type=c("response"))
mydata$prob=prob
g=roc(y,prob, data=mydata)
plot(g)
```



```
auc(g)
```

```
## Area under the curve: 0.7476
```

The AUROC is 0.75.