Simple predictive models: Linear and logistic regression

Montserrat Guillen

14 de diciembre 2017

Contents

In this document we continue with simple Data Analysis linked to the article by S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014. The two datasets ontain similar information, but not exactly the same.. Here we will analyse the smaller data set (called **bank.csv**). The file can be downloaded from: https://archive.ics.uci.edu/ml/datasets/bank+marketing

or (for this course)

http://www.ub.edu/rfa/docs/DATA/bank.csv

We will see linear regression and logistic regression.

Introduction

Here we set up some options for the Rmarkdown ouput. We want to see the R programme (echo=TRUE), but sometime we do not want to see the output, then we set include=FALSE.

Reading the data

Here we read the data and check the names.

```
# read data
#setwd("..")
mydata <- read.csv2("bank.csv", header=TRUE, sep=";", dec=".")
n.var <- names(mydata)
glimpse(mydata)
```

```
## Observations: 4,521
## Variables: 17
## $ age       <int> 30, 33, 35, 30, 59, 35, 36, 39, 41, 43, 39, 43, 36, ...
```

```
## $ job       <fctr> unemployed, services, management, management, blue-...
## $ marital   <fctr> married, married, single, married, married, single,...
## $ education <fctr> primary, secondary, tertiary, tertiary, secondary, ...
## $ default   <fctr> no, no, no, no, no, no, no, no, no, no, no, no, no,...
## $ balance   <int> 1787, 4789, 1350, 1476, 0, 747, 307, 147, 221, -88, ...
## $ housing   <fctr> no, yes, yes, yes, yes, no, yes, yes, yes, yes, yes...
## $ loan      <fctr> no, yes, no, yes, no, no, no, no, no, yes, no, no, ...
## $ contact   <fctr> cellular, cellular, cellular, unknown, unknown, cel...
## $ day       <int> 19, 11, 16, 3, 5, 23, 14, 6, 14, 17, 20, 17, 13, 30,...
## $ month     <fctr> oct, may, apr, jun, may, feb, may, may, may, apr, m...
## $ duration  <int> 79, 220, 185, 199, 226, 141, 341, 151, 57, 313, 273,...
## $ campaign  <int> 1, 1, 1, 4, 1, 2, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 5, 1...
## $ pdays     <int> -1, 339, 330, -1, -1, 176, 330, -1, -1, 147, -1, -1,...
## $ previous  <int> 0, 4, 1, 0, 0, 3, 2, 0, 0, 2, 0, 0, 0, 0, 1, 0, 0, 2...
## $ poutcome  <fctr> unknown, failure, failure, unknown, unknown, failur...
## $ y         <fctr> no, no, no, no, no, no, no, no, no, no, no, no, no,...
```

We have previously analysed the data. Just recall that the data contain 4521 cases and 17 variables. The variable names are: age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome, 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**, **balance** and days since last call (**pdays**).

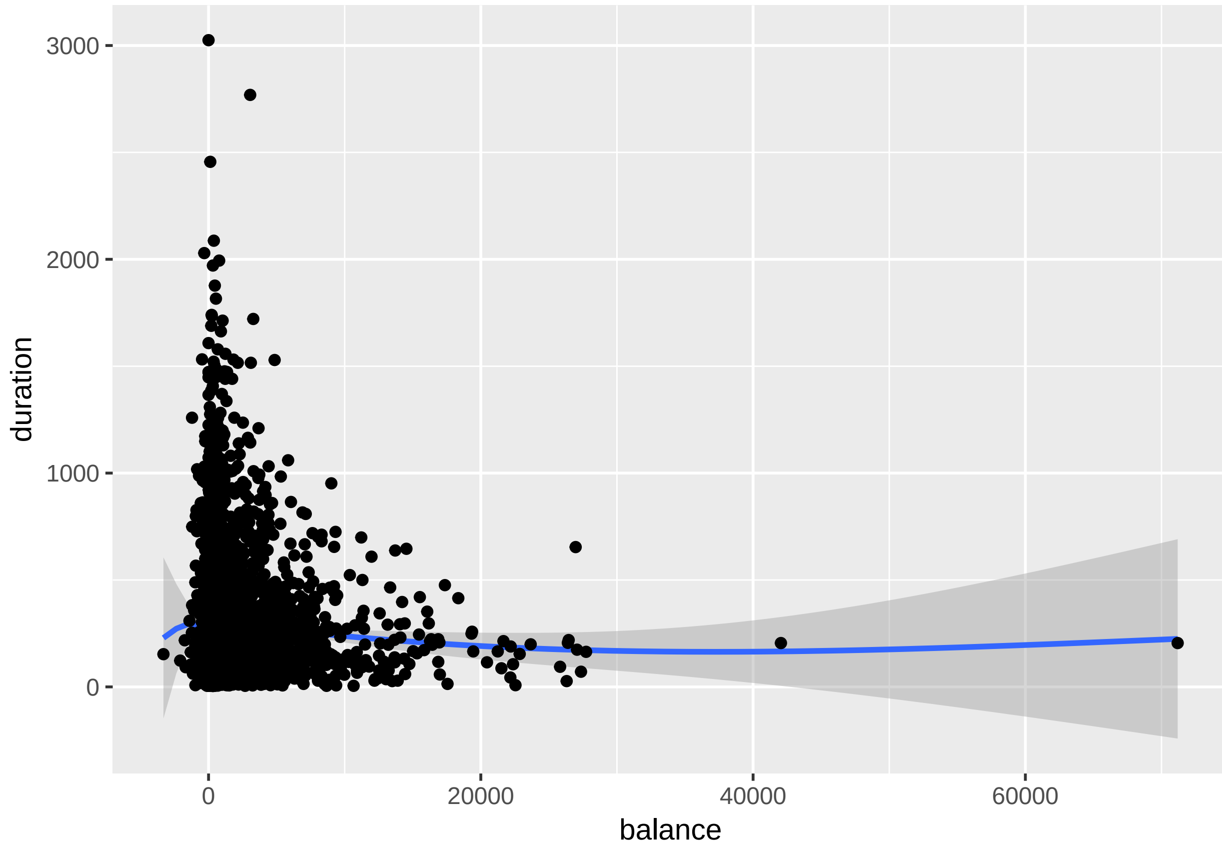
```r
# Model estimation
attach(mydata)

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

```
##
## Call:
## lm(formula = duration ~ age + balance + pdays, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -281.86 -159.95  -78.95   64.60 2760.60
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 265.782426  15.635701  16.998   <2e-16 ***
## age          -0.022964   0.366818  -0.063    0.950
## balance      -0.001379   0.001289  -1.070    0.285
## pdays         0.027311   0.038614   0.707    0.479
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 259.9 on 4517 degrees of freedom
## Multiple R-squared:  0.0003662,  Adjusted R-squared:  -0.0002977
## F-statistic: 0.5515 on 3 and 4517 DF,  p-value: 0.6471
```

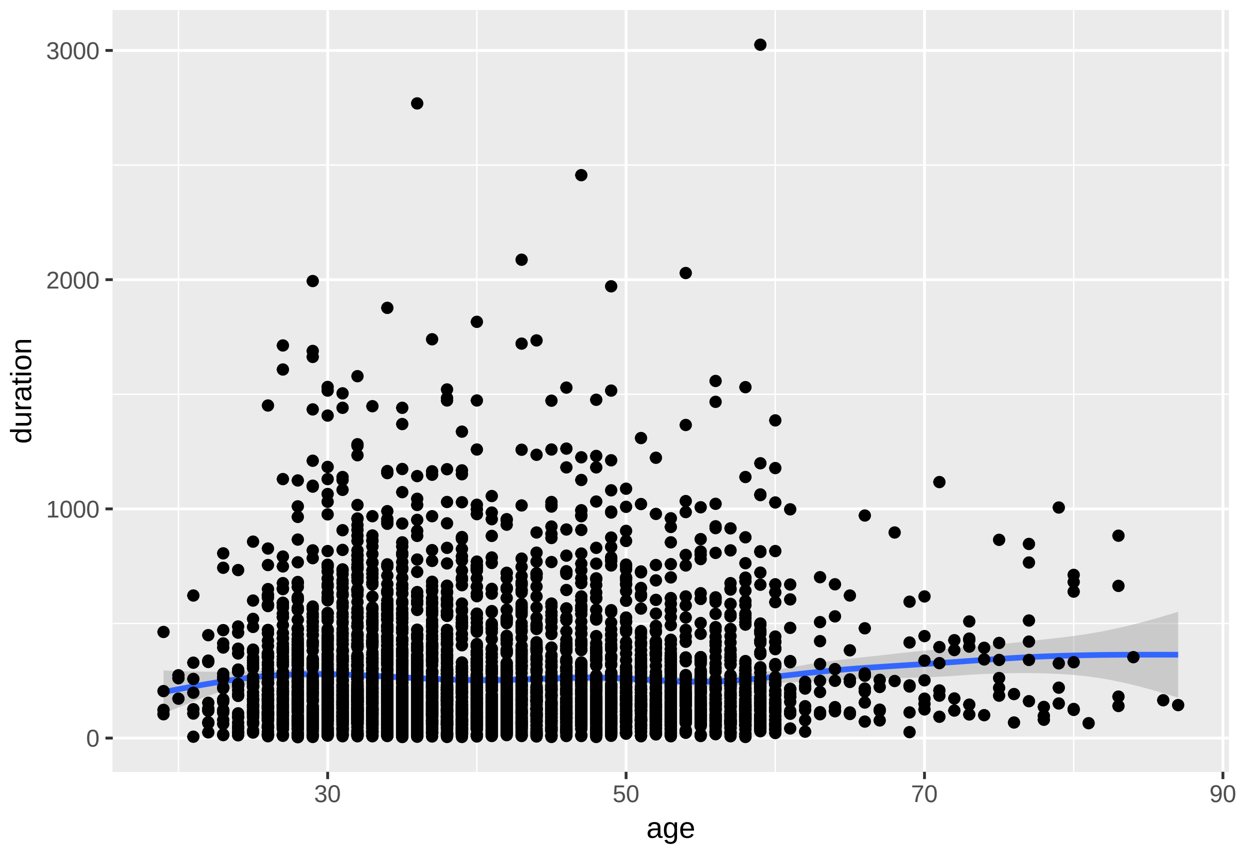
```r
qplot(balance,duration, data = mydata,geom = c("smooth", "point"))
```

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



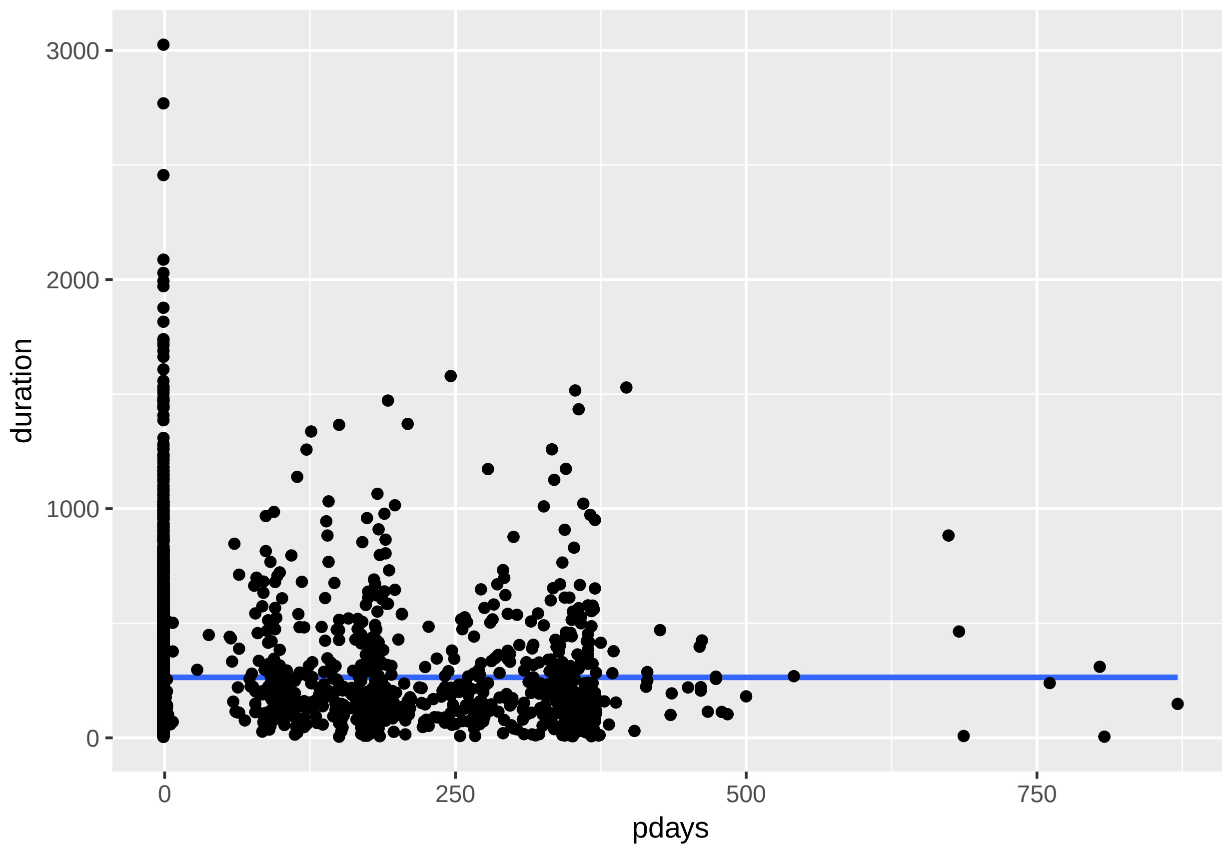

```r
qplot(age,duration, data = mydata,geom = c("smooth", "point"))
```

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




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

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



The goodness-of-fit coefficient is 0.00037

Linear model (quantitative and qualitative regressors)

We now also include **month**, **loan** (yes/no) and **contact** (telephone/cellular/other).

```
monthR=relevel(month, ref = 'mar')
loanR=relevel(loan, ref = 'no')
contactR=relevel(contact, ref = 'telephone')

Model.1.2<- lm(duration~age+ balance+factor(monthR)+factor(loanR)+factor(contactR), data=mydata )
summary(Model.1.2)
```

```
##
## Call:
## lm(formula = duration ~ age + balance + factor(monthR) + factor(loanR) +
##     factor(contactR), data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -353.60 -158.81  -78.40   62.85 2746.30
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              171.242123  44.030023   3.889 0.000102 ***
## age                        0.197277   0.377299   0.523 0.601092
```

```
## balance                    -0.001513   0.001311  -1.154 0.248443
## factor(monthR)apr          92.949051  40.207779   2.312 0.020838 *
## factor(monthR)aug          40.641928  38.611697   1.053 0.292590
## factor(monthR)dec         218.572502  68.940605   3.170 0.001532 **
## factor(monthR)feb          54.871529  41.062446   1.336 0.181520
## factor(monthR)jan          68.562270  42.866071   1.599 0.109790
## factor(monthR)jul          72.012452  38.542344   1.868 0.061771 .
## factor(monthR)jun          52.273360  40.104623   1.303 0.192496
## factor(monthR)may          66.457346  38.499819   1.726 0.084385 .
## factor(monthR)nov          73.249744  39.458837   1.856 0.063468 .
## factor(monthR)oct          73.640188  47.112642   1.563 0.118107
## factor(monthR)sep          14.211828  51.794399   0.274 0.783798
## factor(loanR)yes           -7.471544  10.950277  -0.682 0.495075
## factor(contactR)cellular  27.257131  16.167606   1.686 0.091882 .
## factor(contactR)unknown   24.269113  19.035798   1.275 0.202403
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 259.5 on 4504 degrees of freedom
## Multiple R-squared:  0.005938,   Adjusted R-squared:  0.002407
## F-statistic: 1.682 on 16 and 4504 DF,  p-value: 0.04299
```

Compare goodnes-of-fit

```r
summary(Model.1.1)$adj.r.squared*100
```

```
## [1] -0.02977405
```

```r
summary(Model.1.2)$adj.r.squared*100
```

```
## [1] 0.2406728
```

The goodness-of-fit coefficient is in the first model -3e-04 and in the second model 0.0024.

Prediction

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

```r
newdata=data.frame(age=30, balance=100.0, monthR='jun', loanR='yes', contactR='cellular', pdays=30)
predict(Model.1.1, newdata)
```

```
##        1
## 265.7749
```

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

```
##        1
## 249.0681
```

Logistic regression model

Estimation of the model

We estimate the model for the dependent variable **y** = **Term Diposit**. We only consider age and duration of the call.

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

```
##
## Call:
## glm(formula = y ~ age + duration, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9345  -0.4331  -0.3550  -0.3053   2.5960
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.8606442  0.2141070  -18.03  < 2e-16 ***
## age          0.0144683  0.0046083    3.14  0.00169 **
## duration     0.0035526  0.0001713   20.73  < 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: 3231.0  on 4520  degrees of freedom
## Residual deviance: 2692.1  on 4518  degrees of freedom
## AIC: 2698.1
##
## Number of Fisher Scoring iterations: 5
```

Prediction with this model

```
newdata=data.frame(age=30, balance=100.0, monthR='jun', loanR='yes', contactR='cellular', pdays=30, dura

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

```
##          1
## 0.07320616
```

The prediction for that custmer and the logistic model is 0.073.

Improve the model

We can improve the model now with more information

```
Model.2.2=glm(y~age+duration+factor(month), family=binomial)
summary(Model.2.2)
```

```
##
## Call:
## glm(formula = y ~ age + duration + factor(month), family = binomial)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1203  -0.4102  -0.3028  -0.2340   2.8635
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.0065925  0.2583671 -11.637  < 2e-16 ***
## age               0.0049704  0.0047033   1.057 0.290612
## duration          0.0039210  0.0001856  21.131  < 2e-16 ***
## factor(month)aug -0.3975137  0.2136124  -1.861 0.062757 .
## factor(month)dec  1.0696658  0.5424684   1.972 0.048627 *
## factor(month)feb  0.0273859  0.2553147   0.107 0.914580
## factor(month)jan -0.6592617  0.3442742  -1.915 0.055501 .
## factor(month)jul -1.0758984  0.2245574  -4.791 1.66e-06 ***
## factor(month)jun -0.7430330  0.2331552  -3.187 0.001438 **
## factor(month)mar  1.7134677  0.3447136   4.971 6.67e-07 ***
## factor(month)may -1.3384223  0.2038424  -6.566 5.17e-11 ***
## factor(month)nov -0.8687405  0.2532264  -3.431 0.000602 ***
## factor(month)oct  1.5898738  0.2928645   5.429 5.68e-08 ***
## factor(month)sep  1.1940001  0.3496861   3.414 0.000639 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3231.0  on 4520  degrees of freedom
## Residual deviance: 2465.5  on 4507  degrees of freedom
## AIC: 2493.5
##
## Number of Fisher Scoring iterations: 6
```

The Akaike Information Criterion (AIC) in the first model was 2698 ans now it is 2494.

ROC curve

Predictive performance

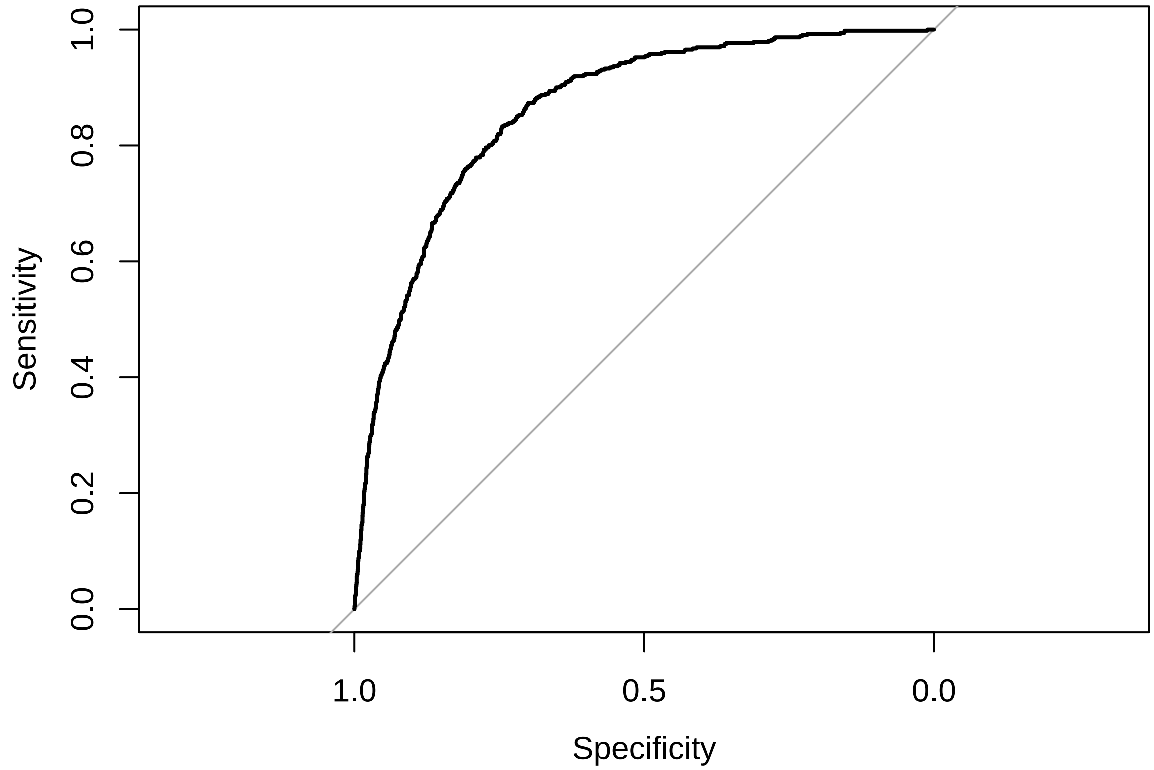
```
#install.packages("pROC")
library(pROC)
```

```
## 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)
```




```r
auc(g)
```

```
## Area under the curve: 0.8618
```

The AUROC is 0.86.