Homework 2. Logistic Regression

Francesc Roy, Marcel Pons

November 2020

1. JYB Dataset

In this project, data from the *JYB dataset* will be analyzed. This dataset contains information of 28,846 calls realized by a bank entity in order to sell a product.

For each call, information about 21 variables is recorded, which can be divided in the following categories: about client attributes, call attributes, campaign attributes, indicators of the socio-economical context and the binary response variable y, which represents if a customer has subscribed to a deposit or not.

2. Objective

The objective of the analysis consists of first performing an exploratory data analysis of the JYB dataset in order to get insightful information about the relations that can exist between variables, emphasizing the relationships with the response variable y.

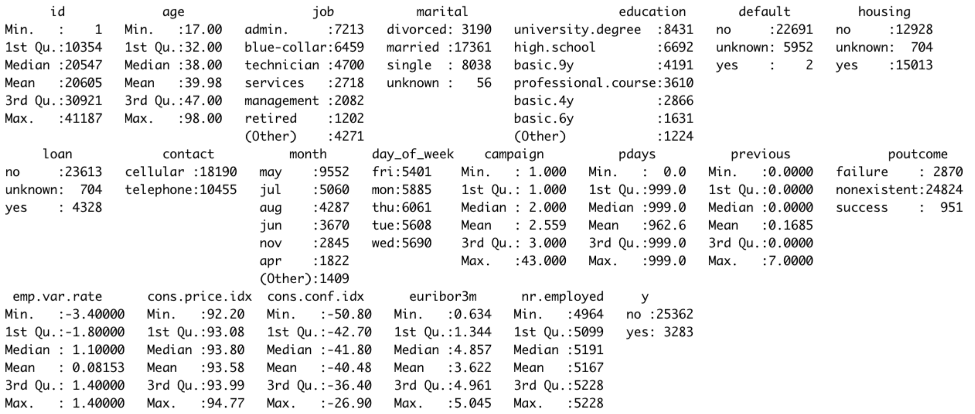
Then, a logistic regression model using the *logit* link function will be built with the purpose of predicting the probability (having a set of explanatory variables) that a client subscribes to the deposit or not as a result of a call. Besides, setting a threshold on the probability value will result in a classifier that would predict a positive or negative subscription from the client.

The model also will be validated and interpreted, trying to understand which explanatory variables are causing more impact (positive or negative) on the response variable, in terms of odds and probabilities.

3. Preprocessing

A data preprocessing stage will be carried out before the analysis, where the categorization of continuous variables will be considered. Furthermore, we will assess if the aggregation of several levels on some categorical variables is necessary and possible. To that end, the summary description of the dataset and the Exploratory Data Analysis will help us to make the final decision.

A basic description of the *JYB dataset* is provided using the `summary` function.

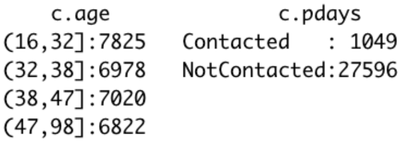


```
       id              age                  job             marital                   education         default          housing     
 Min.   :    1   Min.   :17.00   admin.     :7213   divorced: 3190   university.degree  :8431   no     :22691   no     :12928  
 1st Qu.:10354   1st Qu.:32.00   blue-collar:6459   married :17361   high.school        :6692   unknown: 5952   unknown:  704  
 Median :20547   Median :38.00   technician :4700   single  : 8038   basic.9y           :4191   yes    :    2   yes    :15013  
 Mean   :20605   Mean   :39.98   services   :2718   unknown :   56   professional.course:3610                                  
 3rd Qu.:30921   3rd Qu.:47.00   management :2082                    basic.4y           :2866                                  
 Max.   :41187   Max.   :98.00   retired    :1202                    basic.6y           :1631                                  
                                 (Other)    :4271                    (Other)            :1224                                  
      loan             contact           month       day_of_week    campaign          pdays          previous            poutcome    
 no     :23613   cellular :18190   may    :9552   fri:5401    Min.   : 1.000   Min.   :  0.0   Min.   :0.0000   failure    : 2870  
 unknown:  704   telephone:10455   jul    :5060   mon:5885    1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.0000   nonexistent:24824  
 yes    : 4328                     aug    :4287   thu:6061    Median : 2.000   Median :999.0   Median :0.0000   success    :  951  
                                   jun    :3670   tue:5608    Mean   : 2.559   Mean   :962.6   Mean   :0.1685                      
                                   nov    :2845   wed:5690    3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.0000                      
                                   apr    :1822               Max.   :43.000   Max.   :999.0   Max.   :7.0000                      
                                   (Other):1409                                                                                    
  emp.var.rate       cons.price.idx  cons.conf.idx      euribor3m      nr.employed      y        
 Min.   :-3.40000   Min.   :92.20   Min.   :-50.80   Min.   :0.634   Min.   :4964   no :25362  
 1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.70   1st Qu.:1.344   1st Qu.:5099   yes: 3283  
 Median : 1.10000   Median :93.80   Median :-41.80   Median :4.857   Median :5191              
 Mean   : 0.08153   Mean   :93.58   Mean   :-40.48   Mean   :3.622   Mean   :5167              
 3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.40   3rd Qu.:4.961   3rd Qu.:5228              
 Max.   : 1.40000   Max.   :94.77   Max.   :-26.90   Max.   :5.045   Max.   :5228              
```

Initially, the dataset contains 11 variables of type *factor*, 5 of type *integer* and 5 of type *numeric*. Moreover, there is no missing data so no technique of imputation is needed.

With regard to the continuous variables, we will categorize the variable *pdays*, which stores the number of days that have passed since the customer was contacted for the last time. We create a new variable *c.pdays* with two levels: Contacted, which represents all the values that range from 1 to 30 in *pdays*, and NotContacted, which includes all the 999s that represented the clients not previously contacted.

On the other hand, the *Age* variable has been factorized into 4 levels in a balanced way. We will decide in the EDA which type of variable representing age we keep for modelling.



```
     c.age              c.pdays     
 (16,32]:7825   Contacted   : 1049  
 (32,38]:6978   NotContacted:27596  
 (38,47]:7020                       
 (47,98]:6822                       
```

4. Exploratory Data Analysis

In order to get more insight about the different variables and their relations on the JYB dataset, we will perform several visualizations, mainly boxplots and barplots. Regarding the barplots, we are interested in assessing the frequency of positive subscriptions in the different levels of the categorical variables. To that end, we count the number of *yes* subscriptions in the *y* binary variable for the different levels of each variable, and then we divide it by the total of subscriptions on each level due to the fact that some levels have more instances (calls) than others. For this reason, if we used the total number of subscriptions instead of frequencies, we would reach biased conclusions about which levels have more tendency of subscribing to the deposit.

4.1. Age Variables

First, we start exploring the the *Age* variable, both in continuous type and categorical type.

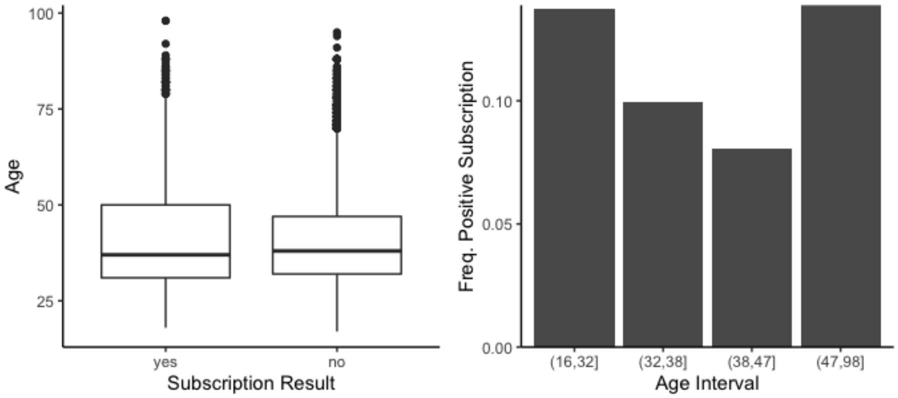


Figure 1: Relation of the binary response variable y and age

We can see that there are not significant differences on ages when it comes to whether to subscribe to the deposit or not. It seems that the categorical approach for age is a little bit more informative, where it is appreciated that younger and older people have a little more tendency to say ”yes” than middle aged people (32-47 years).

4.1. Variables about economical context

Next, we will analyze if the 5 continuous variables that represent the economical status along the months may have some influence on the decision on subscribing. In that case, we build line plots to help with visualizations, being the first one the frequency of positive subscription on every month (from March to December). This first plot will be used as the reference to reach conclusions about the relationship of positive subscription with economical indexes.

The plots of the different context variables are performed by obtaining the average value for each variable each month (since some variables have little fluctuations in the same month)

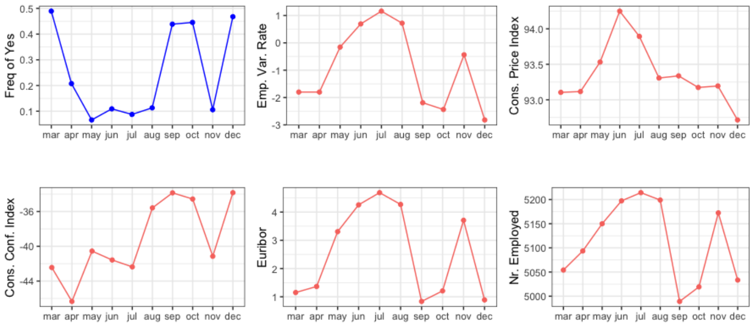


Figure 2: Relation of the freq. of positive subscription and the context variables along the months

In these 6 plots of *Figure 2* we can see interesting relationships among explanatory variables. On May - July, when the number of employed people is at its maximum, the relative number of subscriptions decreases. In contrast, when the number of employed people decreases (Sept - Oct), the frequency of subscriptions increases. This is maybe due to the fact that when people is earning a salary (i.e., is working), they tend to be less encouraged to buy bank credit.

Moreover, as we can observe, *Emp. Var. Rate* and *Euribor*[1] follow similar trends, also being considerably related with each other. Besides, the two indexes (consumer price and confidence) seem to be strongly related with the ups and downs on the frequency of subscription.

4.1. Day of the Week Variable

We can also study if the day of the week on which a call is realized can lead to more positive subscriptions. From the *Figure 3*, it seems that the frequency of subscriptions does not change considerably depending on the day of the week, maybe a little less on Mondays[2] and a little more on Thursdays.

[1] In fact it is logic that the frequency of subscriptions increases as Euribor decreases, because Euribor is a daily reference rate, published by the European Money Markets Institute, based on the averaged interest rates at which Eurozone banks offer to lend unsecured funds.

[2] Which can be due to some psychological explanation.

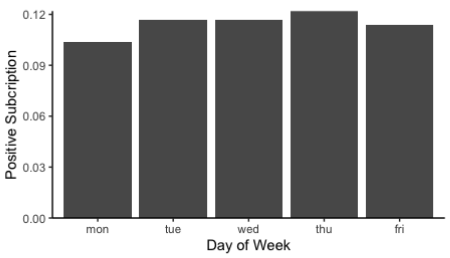


Figure 3: Relation of the frequency of positive subscription with the day of the week

4.2. Job & Education Variables

For *Job* and *Education* variables, we encounter a considerable number of levels, which depending on their similarity on the frequency of positive subscription, we could decide to aggregate them to a new level.

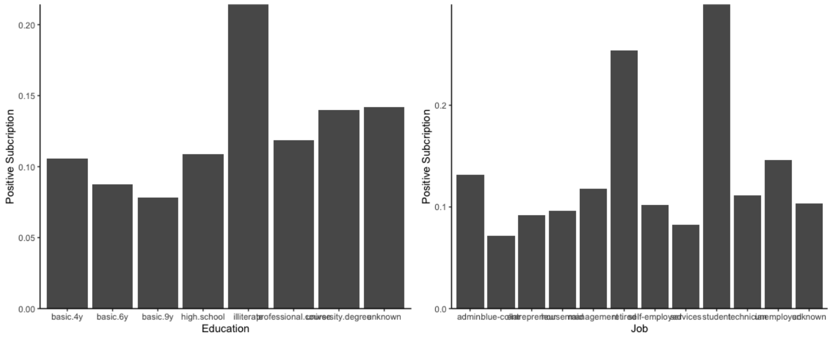


Figure 4: Relation of the frequency of positive subscription with job and education

In these 2 plots of *Figure 4* we can note that people without any kind of education tend to be much more prone to buy bank products. Also we could point from here that students are the ones that buy more credit and that is normal because the don't usually have an monthly income.

In the education plot, the *basic* levels tend to be similar on subscribing. On the other hand, we could appreciate some similar behaviour on people with higher education. Likewise, in the *Job* plot we can also see similar tendencies between levels. Therefore, in the modelling part we will consider the aggregation of some of these levels.

4.3. Other Variables

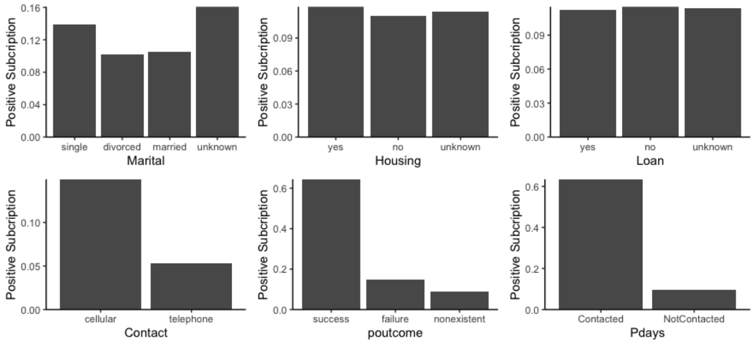


Figure 5: Relation of the frequency of positive subscription with the remaining categorical variables

In these last 6 plots of *Figure 5* what is more interesting to remark is that being contacted previously has a huge impact on subscribing to the product and that people that subscribe to the product last campaign are more likely to repeat and buy it again (*poutcome*).

5. Aggregation in Categorical Variables

As we have seen in the EDA, some categorical that have a considerable amount of levels can be modified by the aggregation of those levels that happened to be more similar on their tendency to subscribe to the deposit (also considering that they can grouped into a general hierarchy, i.e., are related[3]).

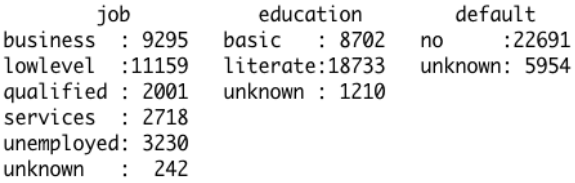
Here there is an example for the aggregation process of the *Job* variable levels.

```
df$job <- as.character(df$job)
df$job[which(df$job == "student")] <-  "unemployed"
df$job[which(df$job == "retired")] <-  "unemployed"
df$job[which(df$job == "housemaid")] <-  "unemployed"
df$job[which(df$job == "blue-collar")] <-  "lowlevel"
df$job[which(df$job == "technician")] <-  "lowlevel"
df$job[which(df$job == "admin.")] <-  "business"
df$job[which(df$job == "management")] <-  "business"
df$job[which(df$job == "entrepreneur")] <-  "qualified"
df$job[which(df$job == "self-employed")] <-  "qualified"
df$job <- as.factor(df$job)
```

[3] For example, basic and higher education could not be aggregated due to their difference in semantics

Furthermore, in the *Default* variable, there is an huge unbalanced level with only 2 instances (”yes” level), therefore we also aggregate the instances of this level into the ”Unknown” level of this variable.

Here is the resulting summary of these categorical variables.



```
        job           education        default     
 business  : 9295   basic    : 8702   no     :22691  
 lowlevel  :11159   literate:18733   unknown: 5954  
 qualified : 2001   unknown : 1210                  
 services  : 2718                                   
 unemployed: 3230                                   
 unknown   :  242                                   
```

6. Logistic Regression Analysis

On the modelling part, we start by performing a Logistic Regression with all the variables (excluding continuous age and pdays). Also, any interaction is considered for the moment.

```
# Model 1 all variables
model1 <- glm(y ˜ ., data = jyb_dataset,
                     family = binomial)
```

With this model, a lot of non-significant coefficients are present, perhaps because of the multicolinearity effect between variables. Moreover, even there are coefficients with NA values due to being completely correlated with each other, being for example the case of *loan*[4]. Therefore, presumably the step function is going to remove them.

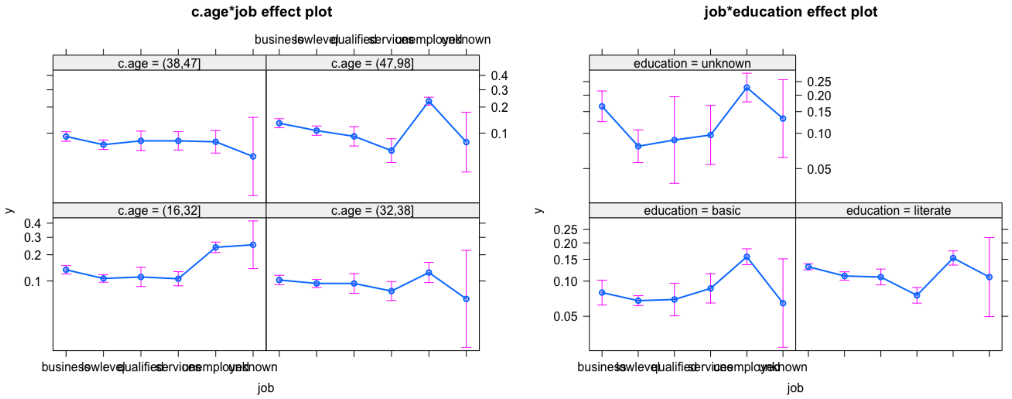
6.1. Assessing Interactions

We propose two possible interactions to be added to the previous model, the one between age and job, and the other between education and job. We choose these two because these variables tend to be related in our society and the fact of making decisions.

```
# Possible interactions
model2 <- glm(y ˜ c.age*job + education*job,
                            data=jyb_dataset,
                            family=binomial)
plot(allEffects(model2, ask=FALSE))
```

[4] Executing the vif function of this model warns us that there are not aliased coefficients in the model, which means a complete colinearity. If we remove loan this problem disappears

In the summary function of the model, some significance can be appreciated between levels in the interaction, for example age(38,47]:unemployed and unemployed:literate. If we visualize the interactions with the allEffects plot, we can see indeed that several changes on the probability of subscription can be influenced by the combination of levels.

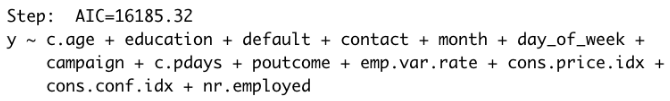


6.2. Feature Selection and Final Model

We include the previous proposed interactions on first model, and then execute the step procedure for feature selection, either with the AIC Criterion and with the BIC criterion.

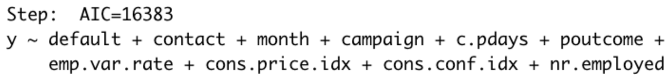
```
model3 <- glm(y ~ . + c.age*job + education*job,
                                    data=jyb_dataset,
                                    family=binomial)
model3_aic <- step(model3, direction='backward', k=2)
model3_bic <- step(model3, direction='backward',
                             k=log(nrow(jyb_dataset)))
```

We can see the final output (step) of these two procedures.



```
Step:  AIC=16185.32
y ~ c.age + education + default + contact + month + day_of_week +
    campaign + c.pdays + poutcome + emp.var.rate + cons.price.idx +
    cons.conf.idx + nr.employed
```

Figure 6: Last step with AIC



```
Step:  AIC=16383
y ~ default + contact + month + campaign + c.pdays + poutcome +
    emp.var.rate + cons.price.idx + cons.conf.idx + nr.employed
```

Figure 7: Last step with BIC

As it can be appreciated, the output of the step procedure with the BIC considers less final variables than the one with the AIC. Using the Anova function to compare the initial model with these two new models, we see that the comparison with the model3_aic is not significant (p-value > 0.05). However, model3_bic appears to be significantly different from the initial model.

A table with the degrees of freedom and the value of both indexes is presented below.

	df	**AIC**	**BIC**
model3_aic	29	16185.32	16424.94
model3_bic	20	16217.74	16383.00

7. Validation of the Final Model

In order to validate the model we could use the same validation plots that we used for the Linear Regression models. However, since now our model has as response variable a binary one (the residuals are discrete), the traditional residual plots are not very helpful in order to asses a logistic model fit (normality, homoscedasticity and independence).

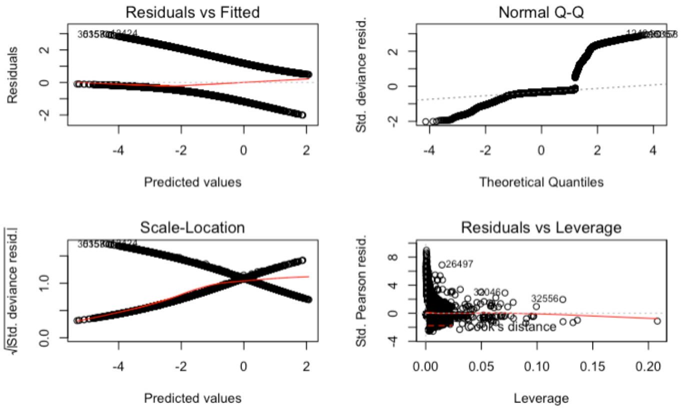


Figure 8: Traditional Validation Plots

A binned residual plot[5] is better in order to assess the fit of the model.

[5] Available on the *arm* package. This plot divides the data into categories (bins) based on their fitted values, the average residual versus the average fitted value for each bin. Gelman and Hill p.97

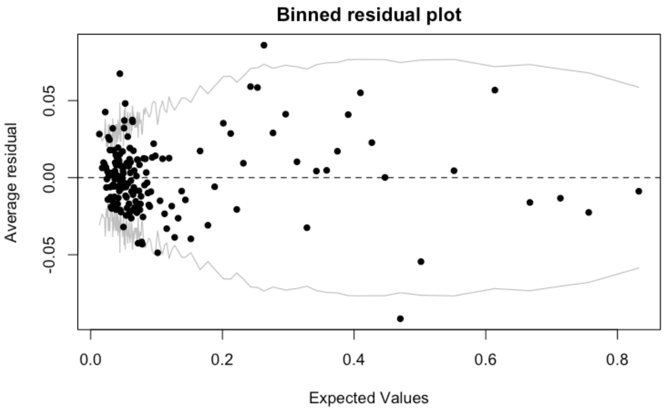


Figure 9: Validation Plots

In this plot, the grey lines represent ±2 Standard Deviation bands, which we expect that contain about 95% of the instances. We can see that the majority of fitted values appear to fall within the SE bands, although some of them are outside.

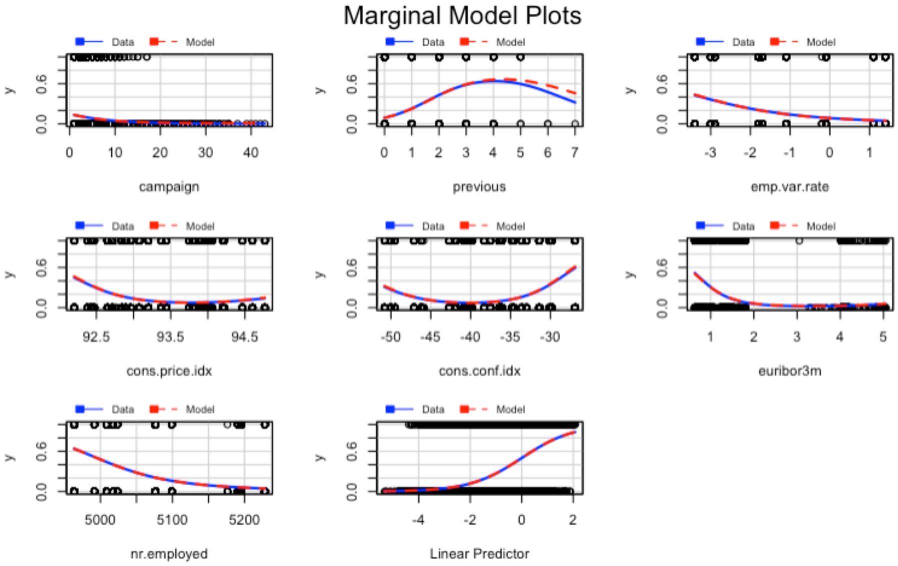
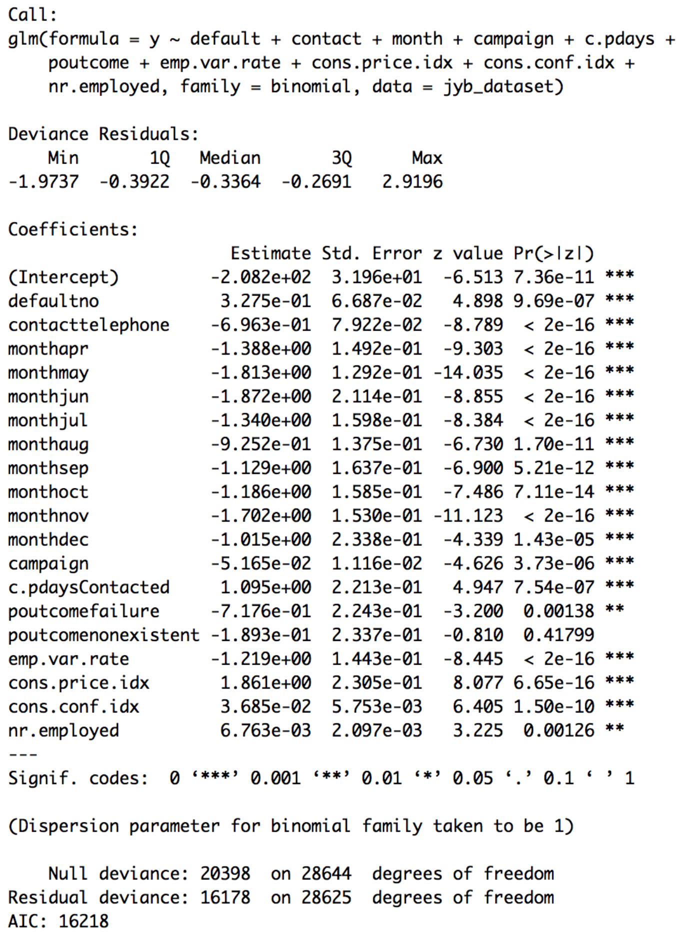


Figure 10: Marginal Model Plots

With regard to the marginal plots, it seems that the data and the model lines coincides almost perfectly.

8. Interpretation

Let's proceed with the interpretation of the model:



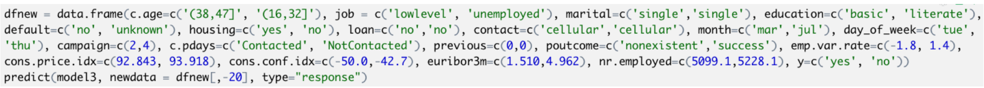
```
Call:
glm(formula = y ~ default + contact + month + campaign + c.pdays +
    poutcome + emp.var.rate + cons.price.idx + cons.conf.idx +
    nr.employed, family = binomial, data = jyb_dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9737  -0.3922  -0.3364  -0.2691   2.9196

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.082e+02  3.196e+01  -6.513 7.36e-11 ***
defaultno            3.275e-01  6.687e-02   4.898 9.69e-07 ***
contacttelephone    -6.963e-01  7.922e-02  -8.789  < 2e-16 ***
monthapr            -1.388e+00  1.492e-01  -9.303  < 2e-16 ***
monthmay            -1.813e+00  1.292e-01 -14.035  < 2e-16 ***
monthjun            -1.872e+00  2.114e-01  -8.855  < 2e-16 ***
monthjul            -1.340e+00  1.598e-01  -8.384  < 2e-16 ***
monthaug            -9.252e-01  1.375e-01  -6.730 1.70e-11 ***
monthsep            -1.129e+00  1.637e-01  -6.900 5.21e-12 ***
monthoct            -1.186e+00  1.585e-01  -7.486 7.11e-14 ***
monthnov            -1.702e+00  1.530e-01 -11.123  < 2e-16 ***
monthdec            -1.015e+00  2.338e-01  -4.339 1.43e-05 ***
campaign            -5.165e-02  1.116e-02  -4.626 3.73e-06 ***
c.pdaysContacted     1.095e+00  2.213e-01   4.947 7.54e-07 ***
poutcomefailure     -7.176e-01  2.243e-01  -3.200  0.00138 **
poutcomenonexistent -1.893e-01  2.337e-01  -0.810  0.41799
emp.var.rate        -1.219e+00  1.443e-01  -8.445  < 2e-16 ***
cons.price.idx       1.861e+00  2.305e-01   8.077 6.65e-16 ***
cons.conf.idx        3.685e-02  5.753e-03   6.405 1.50e-10 ***
nr.employed          6.763e-03  2.097e-03   3.225  0.00126 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20398  on 28644  degrees of freedom
Residual deviance: 16178  on 28625  degrees of freedom
AIC: 16218
```

Figure 11: Summary of final model

As we can see from the summary table, from the numerical variables, *cons.price.idx* is the one augmenting the odds in greater measure (the odds are multiplied by six ($e^{1.861}$)) by each unit incremented.

Regarding categorical variables we see that being contacted by telephone decreases the odds respecting being contacted by cellular. Also we can observe that if a person has been contacted previously, the odds that he subscribes to the product also increases ($e^{1.095} \approx 3$). Finally we can also detect that if a person didn't subscribe to the last product campaign the odds that he subscribes in this one decreases by $e^{-7.17e-01} \approx 0.48$.

We can evaluate how the model predicts new data by using the `predict` function. To that end, we create two new instances, each one with a different type of subscription outcome.



```
dfnew = data.frame(c.age=c('(38,47]', '(16,32]'), job = c('lowlevel', 'unemployed'), marital=c('single','single'), education=c('basic', 'literate'),
default=c('no', 'unknown'), housing=c('yes', 'no'), loan=c('no','no'), contact=c('cellular','cellular'), month=c('mar','jul'), day_of_week=c('tue',
'thu'), campaign=c(2,4), c.pdays=c('Contacted', 'NotContacted'), previous=c(0,0), poutcome=c('nonexistent','success'), emp.var.rate=c(-1.8, 1.4),
cons.price.idx=c(92.843, 93.918), cons.conf.idx=c(-50.0,-42.7), euribor3m=c(1.510,4.962), nr.employed=c(5099.1,5228.1), y=c('yes', 'no'))
predict(model3, newdata = dfnew[,-20], type="response")
```

Since we specify type='response', we get the probability of positive subscription. The result is $probability = 0.6$ for the instance that was actually positive and 0.066 for the negative one. Therefore, it seems that the model did a good job at assessing the probability of subscription for these two instances.

9. Conclusions

To sum up let's state all the path done until the final model:

We started building the complete model with all explanatory variables.

After fitting the complete model *model1*, we build *model3* that was the complete plus two interactions that seemed interesting.

Next phase has been to use step procedure in order to improve this *model3*.

The step procedure using the AIC criterion was first analysed but the resulting model was not significantly better than model 3, so we discarded it.

At the end the best model that we could achieve is the model resulting from the step procedure with BIC criterion. This model is formed by 10 explanatory variables (4 numerical and 6 categorical) and without taking into account any interaction (they were removed). This model reduces the deviance from the null model (20398) to 16178.

After building all this nested models we realized that:

- Performing a intensive exploratory data analysis can bring useful insights on the data set that sometimes coincides with the model and other times not.
- The study of interactions is complex due to the considerable number of levels in categorical variables.
- The validation and interpretation of the logistic model is a little bit more tricky than in linear models.