ASM Homework 2

Generalized Linear Model for JYB data

Maria Gkotsopoulou & Ricard Monge Calvo & Amalia Vradi

22/10/2019

Exploratory Data Analysis

We check for any missing values or attributes without a value and find none nor NAs. Next, we present a summary of the variables, basic statistics and levels:

Table 1: Numerical variables

variable	class	min	mean	median	max
age	integer	17.00	39.98	38.00	98.00
campaign	integer	1.00	2.56	2.00	43.00
pdays	integer	0.00	962.63	999.00	999.00
previous	integer	0.00	0.17	0.00	7.00
emp.var.rate	numeric	-3.40	0.08	1.10	1.40
cons.price.idx	numeric	92.20	93.58	93.80	94.77
cons.conf.idx	numeric	-50.80	-40.48	-41.80	-26.90
euribor3m	numeric	0.63	3.62	4.86	5.04
nr.employed	numeric	4963.60	5167.00	5191.00	5228.10

Table 2: Categorical variables

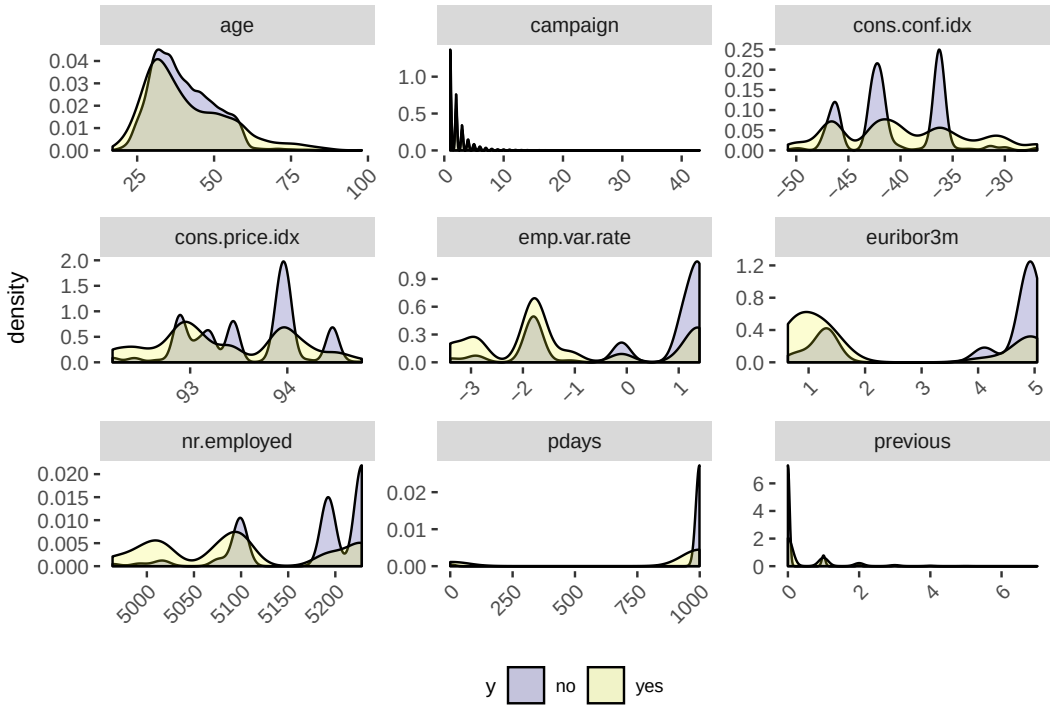
attribute	# levels
job	12
marital	4
education	8
default	3
housing	3
loan	3
contact	2
month	10
day_of_week	5
poutcome	3
y	2

attribute	level_1	level_2	level_3	level_4	level_5	level_6	level_7	level_8	level_9	level_10	level_11	level_12
job	admin.	blue-collar	entrepreneur	housemaid	management	retired	self-employed	services	student	technician	unemployed	unknown
marital	divorced	married	single	unknown								
education	basic.4y	basic.6y	basic.9y	high.school	illiterate	professional.course	university.degree	unknown				
default	no	unknown	yes									
housing	no	unknown	yes									
loan	no	unknown	yes									
contact	cellular	telephone										
month	apr	aug	dec	jul	jun	mar	may	nov	oct	sep		
day_of_week	fri	mon	thu	tue	wed							
poutcome	failure	nonexistent	success									
y	no	yes										

Variable *pdays* has a value of 999 if the customer was not previously contacted. This makes the basic statistics of the variable meaningless. We could consider discretizing it.

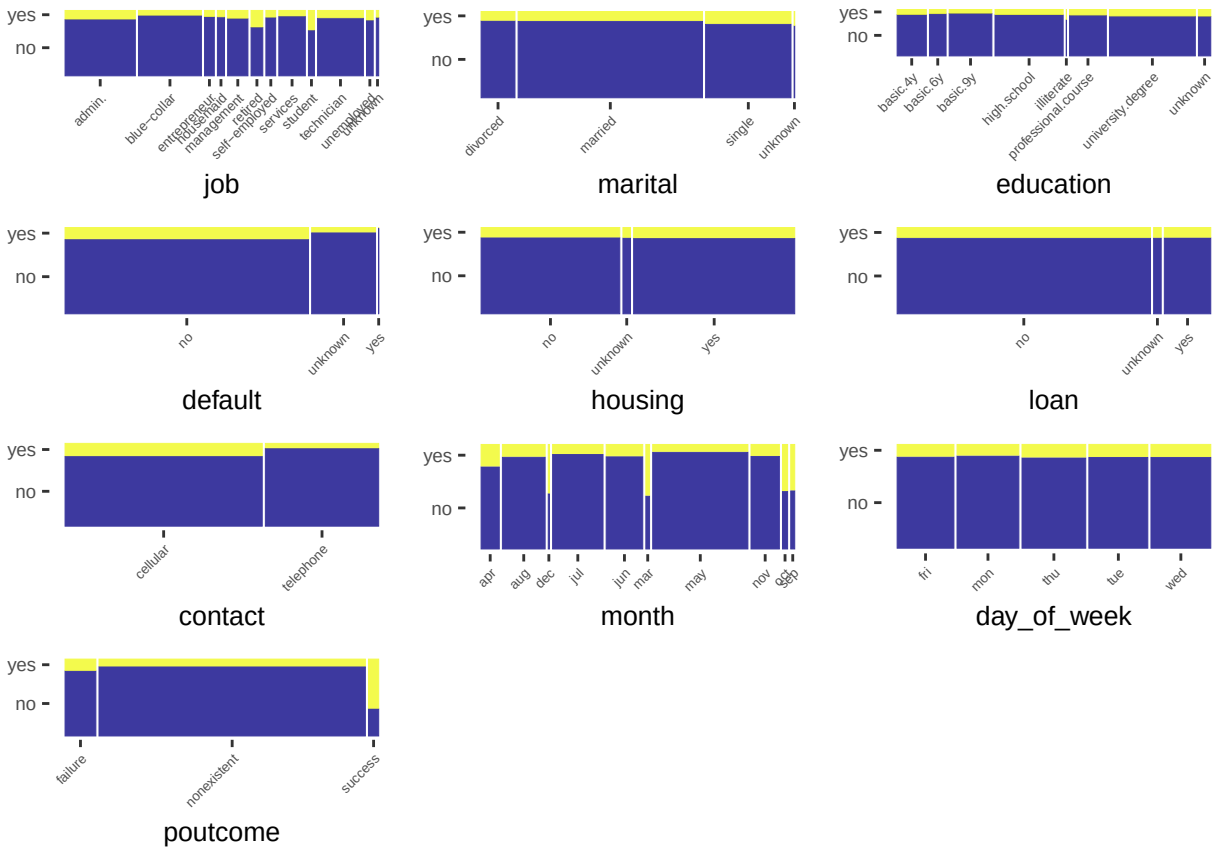
We are interested in predicting whether the customer subscribed to the deposit, so our target variable y is a binary one.

We now look closer into the relation between y and all the numerical variables.



In general, we see that for the *no* class we have higher and more skewed distribution of values, for instance for *cons.conf.idx*, *cons.price.idx*, *euribor3m* and *nr.employed*. We expect them to be relevant in predicting the response variable.

Furthermore, we look into the relation between y and all the factor variables.



From the mosaic plots we see the variables which are more differenciated in some levels are *job*, *default*, *contact*,

month and *poutcome*. We expect them to be relevant in predicting the response variable.

Table 3: description by categorical variables

	p.value	df
month	0.000000e+00	9
poutcome	0.000000e+00	2
contact	0.000000e+00	1
job	0.000000e+00	11
default	0.000000e+00	2
education	6.666612e-26	7
marital	5.822164e-15	3
day_of_week	2.940879e-02	4

Table 4: description by continuous variables

	Eta2	P-value
pdays	0.101421477	0.000000e+00
previous	0.051869032	0.000000e+00
emp.var.rate	0.089296876	0.000000e+00
euribor3m	0.094811061	0.000000e+00
nr.employed	0.124452342	0.000000e+00
cons.price.idx	0.019873250	0.000000e+00
campaign	0.004582956	1.860000e-30
cons.conf.idx	0.003391874	5.907421e-23
age	0.001232525	2.791189e-09

Based on the p-value, all variables seem signifcant. However, the most important ones seems to be *month* and *poutcome*, as categorical variables, and *pdays* and *previous* as numerical variables. As before, we expect them to be relevant in the following models.

Complete Model

We build a complete model with all the variables and compare it with the null model.

```
m.null <- glm(y~1, family = binomial(link = "logit"),
              data = jyb)
m.full <- glm(y~., family = binomial(link = "logit"),
              data = jyb)
# Compare full and null model by anova test
anova(m.null, m.full, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ 1
## Model 2: y ~ age + job + marital + education + default + housing + loan +
##     contact + month + day_of_week + campaign + pdays + previous +
##     poutcome + emp.var.rate + cons.price.idx + cons.conf.idx +
##     euribor3m + nr.employed
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     28644      20398
## 2     28593      16112 51     4286 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# View model summary
glance(m.full)
```

```
## # A tibble: 1 x 7
##   null.deviance df.null logLik    AIC    BIC deviance df.residual
##           <dbl>   <int>  <dbl>  <dbl>  <dbl>    <dbl>       <int>
## 1        20398.   28644 -8056. 16216. 16646.   16112.       28593
```

```r
# See significance of factor variables
anova(m.full, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: y
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           28644      20398
## age             1    34.57     28643      20363 4.119e-09 ***
## job            11   501.43     28632      19862 < 2.2e-16 ***
## marital         3    52.93     28629      19809 1.893e-11 ***
## education       7    41.89     28622      19767 5.463e-07 ***
## default         2   285.19     28620      19482 < 2.2e-16 ***
## housing         2     2.31     28618      19480 0.3153336
## loan            1     0.45     28617      19479 0.5009470
## contact         1   468.35     28616      19011 < 2.2e-16 ***
## month           9   944.52     28607      18066 < 2.2e-16 ***
## day_of_week     4    19.41     28603      18047 0.0006541 ***
## campaign        1    76.87     28602      17970 < 2.2e-16 ***
## pdays           1   887.84     28601      17082 < 2.2e-16 ***
## previous        1     0.28     28600      17082 0.5965456
## poutcome        2    12.65     28598      17069 0.0017947 **
## emp.var.rate    1   628.50     28597      16441 < 2.2e-16 ***
## cons.price.idx  1   290.55     28596      16150 < 2.2e-16 ***
## cons.conf.idx   1    26.48     28595      16124 2.666e-07 ***
## euribor3m       1    10.08     28594      16114 0.0015028 **
## nr.employed     1     1.74     28593      16112 0.1869030
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Evaluation First Order Interactions

Considering that a model with all the variables and all the factor-factor and factor-covariable first order interactions would be too big to compute and handle in a reasonable amount of time, we perform a model with the interactions with each factor separately. We regard as signficant the interactions which give a p-value lower than 0.001 in the Anova Chi-squared test.

```r
# First order interactions with one factor housing
factors <- jyb %>% dplyr::select_if(is.factor) %>% colnames()
formula.inter <- as.formula(
```

```
  paste0("y~.*(", paste0(factors[5], collapse = "+"),")"))
m.inter <- glm(formula.inter, family = binomial(link = "logit"), data = jyb)
glance(m.inter)
# Comparison with full model
anova(m.full, m.inter, test = "Chisq")
# Significance of interaction terms
anova(m.inter, test = "Chisq")
```

In the case of the *contact* variable interactions based on the p-value of the anova test comparison we can reject the H_0 so the models are different. The AIC is 16150 while the AIC of the full model is 16216 so the new model is better. From the *ANOVA* output we see that the interactions *contact:month*, *contact:cons.price.idx* and *contact:euribor3n* are significant.

In the case of the *day_of_week* variable interactions based on the p-value of the anova test comparison we can reject the H_0 so the models are different. The AIC is 16262 while the AIC of the full model is 16216 so the full model is slightly better. From the *ANOVA* output we see that the interaction *day_of_week:month* is significant.

In the case of the *default* variable interactions based on the p-value of the anova test comparison we can reject the H_0 so the models are different. The AIC is 16241 while the AIC of the full model is 16216 so the full model is better. From the *ANOVA* output we see that *default:euribor3m* is significant.

In the case of the *education* variable interactions based on the p-value of the anova test comparison we can't reject the H_0 so the models are the same. From the*ANOVA* output we see that *education:previous*, *education:poutcome* is significant.

In the case of *housing* and *loan* variables interactions, based on the p-value of the anova test comparison we can't reject the H_0 so the models are the same. Furthermore, we do not get any significant interaction terms.

In the case of the *marital* variable interactions based on the p-value of the anova test comparison we can't reject the H_0 so the models are the same. From the *ANOVA* output we see that *marital:day_of_week* is significant.

In the case of the *month* variable interactions, based on the p-value of the anova test comparison we can reject the H_0 so the models are different. The AIC is 16208 while the AIC of the full model is 16216 so the new model is slightly better. From the *ANOVA* output we see that *month:emp.var.rate*, *month:cons.price.idx*, *month:euribor3n* is significant.

In the case of *poutcome* variable interactions, based on the p-value of the anova test comparison we can reject the H_0 so the models are different. However, we do not get any significant interaction terms.

After reviewing the full model together with the obtained significant interaction terms, we can conclude that variables *job*, *housing*, *loan* and *nr.employed* can be removed. Next, we build our complete model with all the interaction terms.

```
formula.inter <- as.formula("y ~ . - (job + housing + loan + nr.employed) +
                            contact:month + contact:cons.price.idx +
                            contact:euribor3m + day_of_week:month +
                            default:euribor3m + education:previous +
                            education:poutcome + marital:day_of_week +
                            month:emp.var.rate + month:cons.price.idx +
                            month:euribor3m")
m.inter <- glm(formula.inter, family = binomial(link = "logit"), data = jyb)
```

With this complete model we get an AIC of 1.5997802×10^4, better than the 1.6215852×10^4 of the full model.

Automatic Variable Selection process

We use the stepwise procedure, by using the *AIC* & *BIC* criterion, to select our final model. Since our objective is the interpretability of the model we choose as starting point the null model, in contrast to starting form the complete. We place as an upper bound the previous complete model with first order interaction terms.

```
# Stepwise model selection with AIC
m.step.aic <- step(m.null, scope=list(upper=m.inter), direction="both",
                   k=2, trace = 0)
m.step.aic$formula
```

```
## y ~ euribor3m + month + poutcome + contact + pdays + cons.price.idx +
##     cons.conf.idx + day_of_week + campaign + default + previous +
##     euribor3m:month + month:cons.price.idx + month:day_of_week +
##     contact:cons.price.idx + euribor3m:contact + euribor3m:default
```

```
# Stepwise model selection with BIC
m.step.bic <- step(m.null, scope=list(upper=m.inter), direction="both",
                   k=log(nrow(jyb)), trace = 0)
m.step.bic$formula
```

```
## y ~ euribor3m + month + poutcome + contact + pdays + cons.price.idx +
##     cons.conf.idx + campaign + emp.var.rate + default + contact:cons.price.idx +
##     euribor3m:contact
```

```
# Anova test comaprison between models
anova(m.step.aic, m.step.bic, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ euribor3m + month + poutcome + contact + pdays + cons.price.idx +
##     cons.conf.idx + day_of_week + campaign + default + previous +
##     euribor3m:month + month:cons.price.idx + month:day_of_week +
##     contact:cons.price.idx + euribor3m:contact + euribor3m:default
## Model 2: y ~ euribor3m + month + poutcome + contact + pdays + cons.price.idx +
##     cons.conf.idx + campaign + emp.var.rate + default + contact:cons.price.idx +
##     euribor3m:contact
##   Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
## 1     28563      15774
## 2     28622      16134 -59  -360.32 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

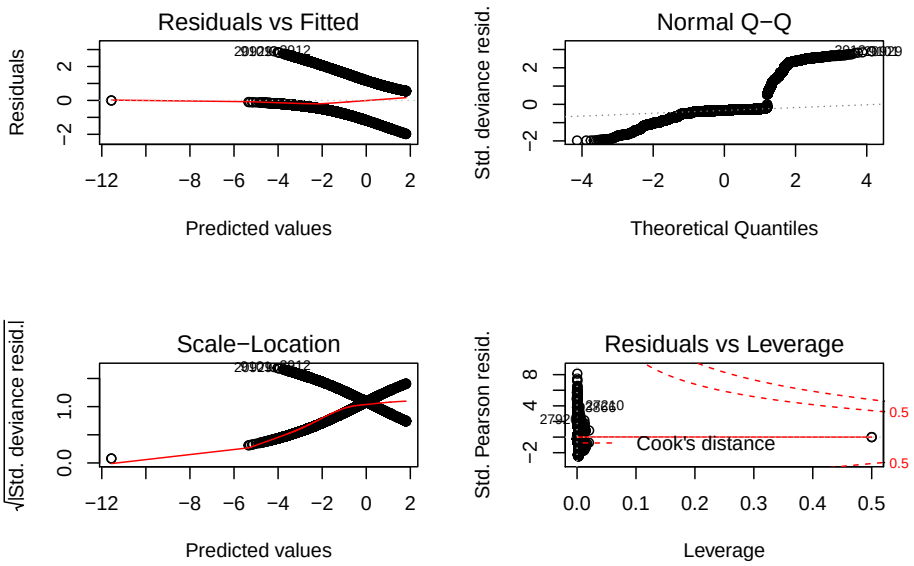
Table 5: Models Summary

model	AIC	BIC
Complete_interactions	15997.80	17146.32
stepwise_AIC	15938.20	16615.74
stepwise_BIC	16180.52	16370.56

We see the models are significantly different through the *Anova* test, but their used predictors are similar. In particular, all but one predictor (*emp.var.rate*) of the *BIC* selected model are included in the *AIC* model. This is due to the difference in criteria, penalizing more additional parameters with *BIC*. As expected, the *AIC* selected model has a better *AIC* measure while the *BIC* selected model has better *BIC* measure. In order to better interpret it, we choose the more succint model based on the parsimony criteria.

Model validation

After selecting our final model, we check the assumptions by looking at the following plots:

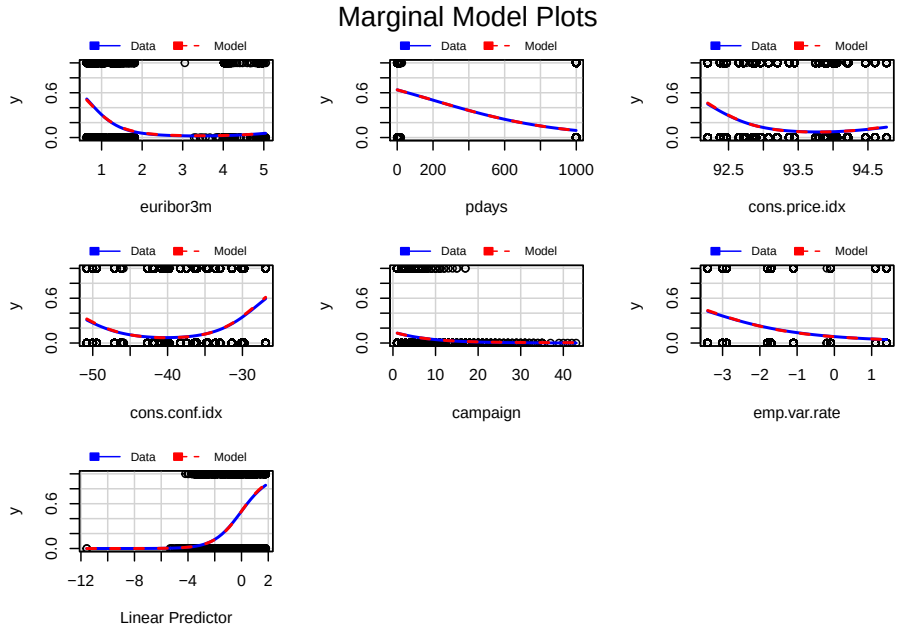


From the *Normal Q-Q* plot we see that there is assymetry in the distribution and we can conclude that normality of the residuals is not met.

Furthermore, from the *Scale-Location* plot, we seek to validate the assumption of homoskedasticity, which does not seem to hold in our case.

Finally, from the *Residuals-Fitted* plot we see that the residuals seem to have some kind of tendency. This indicates that we should consider quadratic, or higher order, terms. However, this is out of the scope of this study.

On the other hand, we can see the marginal residual plots to compare the mean values of the data and our model:

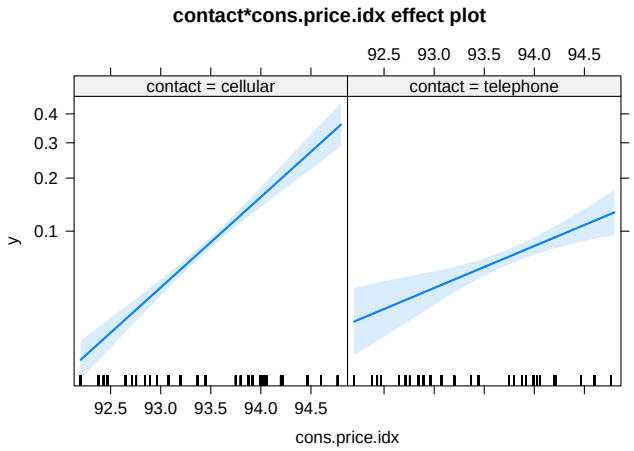


For each variable we see the data and model's mean values coincide very well.

Model interpretation

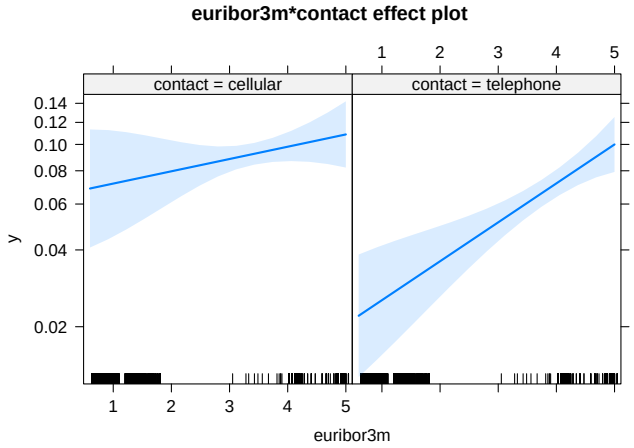
For the model interpretation, we will use the *effect* plots, taking into account the interaction terms. As we are building a logistic regression model, we can not directly interpret the coefficients and response values, since they do not represent the response variable class but the value of the link function.

```
m.effects <- effects::allEffects(m.final)
plot(m.effects, "contact:cons.price.idx")
```



We observe the lines are not parallel, so we have different behaviour of the clients depending on the contact method. In general terms, the increase on the the consumer price index for a client gives higher probability of him/her subscribing to the deposit when being contacted by cellular instead of telephone.

```
plot(m.effects, "euribor3m:contact")
```



In the same way as before the increase in the *Euribor* gives higher probability of subscribing for clients contacted by cellular than for clients contacted by phone. Nevertheless, the slope of the increase due to a change in *Euribor* is bigger in the case of telephone contact.

To complete the interpretation of our numerical variables, we see the odds ratios increase per unit increase of each variable. We obtain the odds ratios by exponentiating the coefficients.

```
## odds ratios and 95% CI
odds <- exp(cbind(OR = coef(m.final), confint(m.final)))
```

Table 6: Odds ratio increase by model coefficient.

Var	OR	2.5 %	97.5 %
contacttelephone	3.254051e+29	8.073728e+19	1.575399e+39
cons.price.idx	3.984000e+00	3.167000e+00	5.015000e+00
monthmar	3.375000e+00	2.613000e+00	4.353000e+00
poutcomesuccess	2.069000e+00	1.322000e+00	3.224000e+00
poutcomenonexistent	1.698000e+00	1.487000e+00	1.943000e+00
monthaug	1.563000e+00	1.242000e+00	1.967000e+00
monthdec	1.317000e+00	8.630000e-01	2.006000e+00
euribor3m:contacttelephone	1.281000e+00	1.170000e+00	1.404000e+00
monthjul	1.132000e+00	9.180000e-01	1.395000e+00
euribor3m	1.121000e+00	9.260000e-01	1.357000e+00
monthsep	1.077000e+00	8.130000e-01	1.426000e+00
monthoct	1.073000e+00	8.240000e-01	1.398000e+00
cons.conf.idx	1.021000e+00	1.009000e+00	1.034000e+00
pdays	9.990000e-01	9.980000e-01	9.990000e-01
campaign	9.470000e-01	9.260000e-01	9.680000e-01
monthjun	8.370000e-01	6.710000e-01	1.044000e+00
defaultunknown	7.270000e-01	6.360000e-01	8.270000e-01
monthnov	7.090000e-01	5.640000e-01	8.900000e-01
monthmay	5.910000e-01	5.060000e-01	6.900000e-01
contacttelephone:cons.price.idx	4.770000e-01	3.750000e-01	6.050000e-01
emp.var.rate	3.850000e-01	3.020000e-01	4.900000e-01
defaultyes	0.000000e+00	NA	4.553719e+05
(Intercept)	0.000000e+00	0.000000e+00	0.000000e+00

We see the highest factor of increase in odds ratio is given by the *cons.price.idx* variable. We interpret this as a higher probability of subscribing to the deposit the higher the consumer price index is. Contrary, the lowest factor of decrease in odds ratio is for the case of *emp.var.rate*. We interpret this as a lower probability of subscribing to the deposit the higher the employer variation rate is.