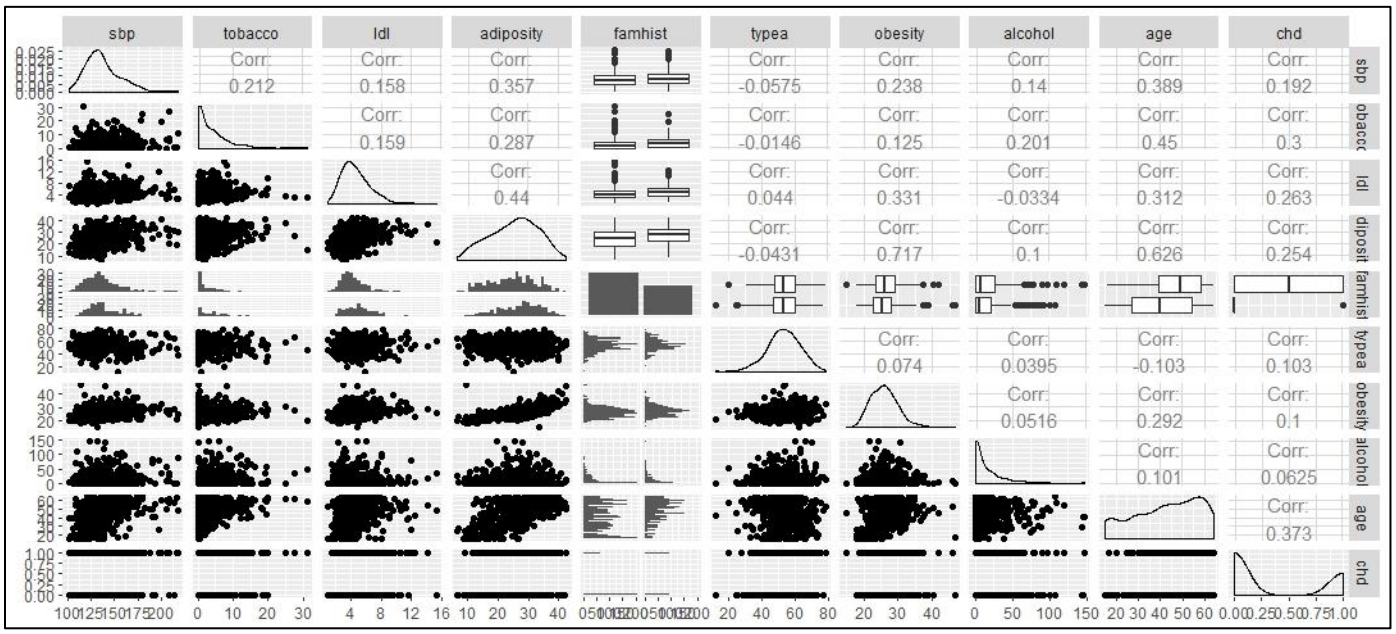
Assignment 2 : Myocardial infarction

1. **Evaluate and compare the variables. Are there missing values? How many observations? Are there correlations between the variables?**



correlationmatrix

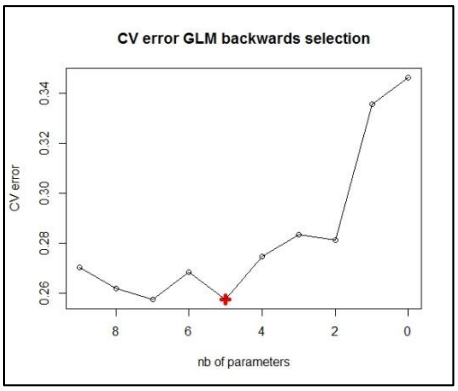
- 9 predictors, 1 response (chd)
- 462 observations
  - 160 persons with mycardial infarction
  - 302 controls
- No missing data
- there is (inevitably) correlation between the predictors. The highest being :
  - adiposity <-> obesity (0.717)
  - age <-> adiposity (0.626)
  - age <-> tobacco (0.45)

2. **Using only family history (famlist) as a predictor, how much does the risk of infarction increase for someone with a family history, compared with someone without family history?**

```
> lrmod1<-glm(chd~famhist,data=heart,family = 'binomial')
> coefFamhist<-lrmod1$coefficients[[2]]
> Intercept<-lrmod1$coefficients[[1]]
> RiseInP<-(exp(coefFamhist+Intercept)/(1+exp(coefFamhist+Intercept))) -
+         (exp(Intercept)/(1+exp(Intercept)))
> RiseInP
[1] 0.262963
```

(the chance rises from 0.237037 to 0.50000, logodds ratio rises with 1.168993)

3. **Now going back to the full dataset, select an optimal generalized linear model using Backward Selection. Test the variables that you dropped out of the model for association with the response, and explain your observations.**



picking the number of predictors based on cross-validation

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age
9pred	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"
8pred	"*"	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"
7pred	"*"	"*"	"*"	" "	"*"	"*"	"*"	" "	"*"
6pred	" "	" "	"*"	" "	"*"	"*"	"*"	" "	"*"
5pred	" "	" "	"*"	" "	"*"	"*"	" "	" "	"*"
4pred	" "	" "	"*"	" "	"*"	" "	" "	" "	"*"
3pred	" "	" "	" "	" "	"*"	" "	" "	" "	"*"
2pred	" "	" "	" "	" "	"*"	" "	" "	" "	"*"
1pred	" "	" "	" "	" "	" "	" "	" "	" "	"*"

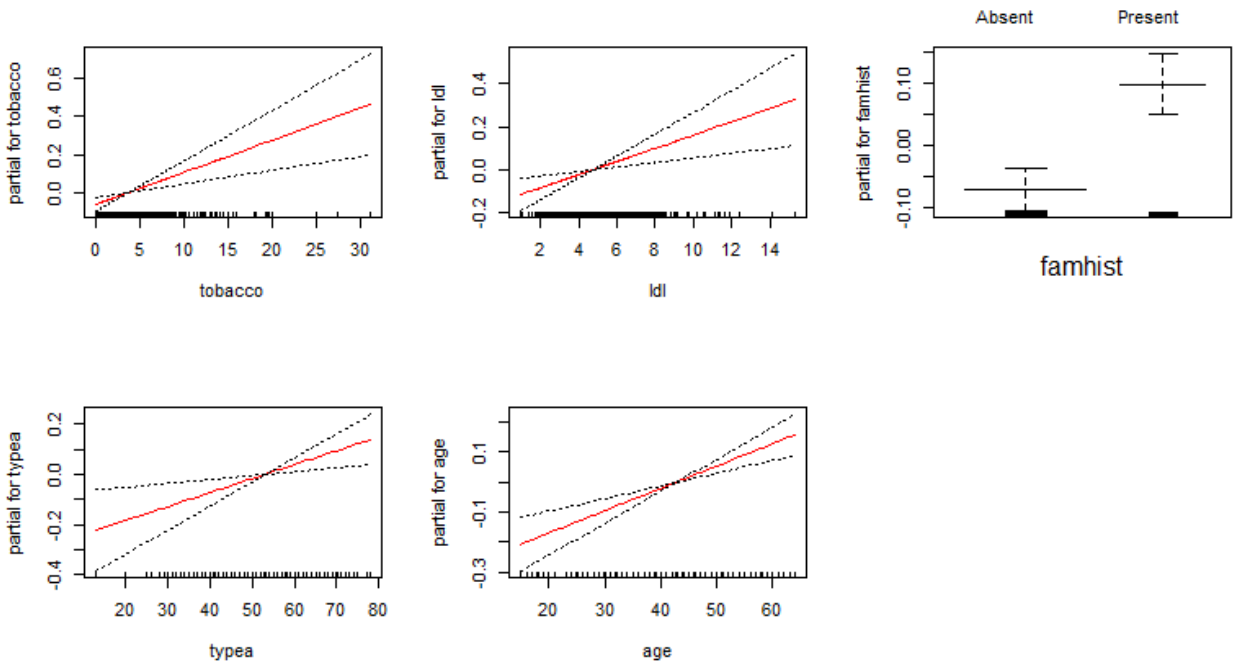
Backward selection matrix for glm

The lowest cross-validation error is seen in the models with 5 and 7 predictors. The model with the least number of predictors is preferred so I select the model with these 5 predictors

```
chd ~ 1 + tobacco + ldl + famhist + typea + age
```

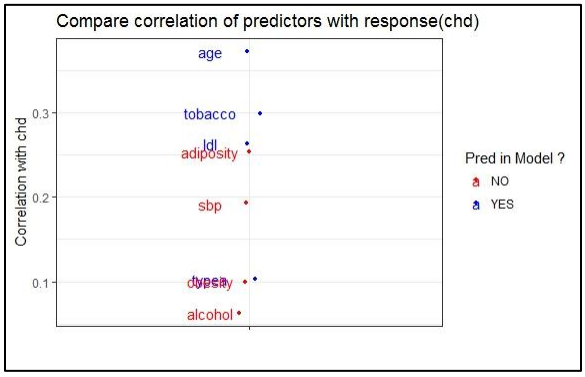
All of these predictors are significant

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.545452	0.130431	-4.182	3.47e-05	***
tobacco	0.016792	0.004783	3.511	0.000491	***
ldl	0.030494	0.010031	3.040	0.002501	**
famhistPresent	0.170682	0.041185	4.144	4.07e-05	***
typea	0.005588	0.002018	2.769	0.005858	**
age	0.007439	0.001606	4.632	4.72e-06	***

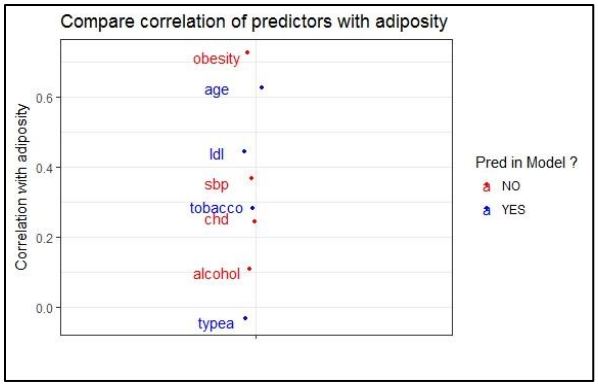


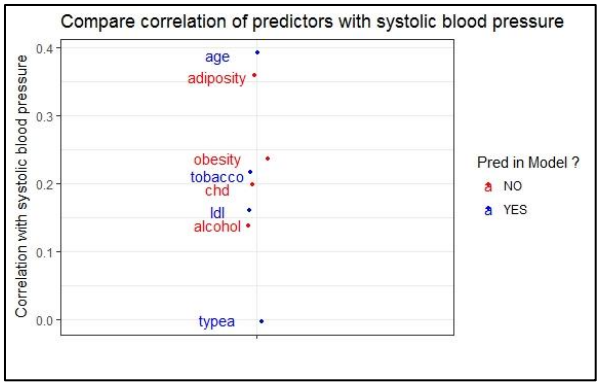
glm model

Broadly speaking these are the predictors that show the highest correlation with the response, which is to be expected. But there are exceptions eg *adiposity* has a correlation of 0.25 which is higher then *typea* (0.10*)*. Still *typea* is selected, and *adiposity* not. This can be explained by the collinearity of the predictors. When *adiposity* was dropped from the model during the backward selection, 7 other predictors were present. At that point much of the association of *adiposity* with the response was already incorporated in these 7 predictors, so *adiposity* itself did not provide much additional predictive power, and therefore was dropped from the model. Conversely, you could construct a model with *obesity* included, which would most likely cause another correlated variable to be dropped.



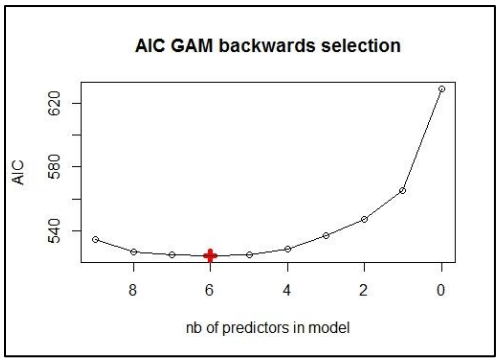
Variables that have high correlation with chd are likely to be part of the model





adiposity is not included in the model, but age and ldl are, which are both highly correlated with adiposity
same reasoning for systolic blood pressure (correlation with age)

4. <u>Now similarly optimize a gam with non-linear trends through Backward Selection. Write down the equation of your model. Do you see evidence of non-linear effects? Do you select the same variables as in the linear model? Explain your results.</u>



Picking the number of predictors based on AIC

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age
9pred	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"
8pred	"*"	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"
7pred	" "	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"
6pred	" "	"*"	"*"	" "	"*"	"*"	"*"	" "	"*"
5pred	" "	"*"	"*"	" "	"*"	"*"	" "	" "	"*"
4pred	" "	"*"	" "	" "	"*"	"*"	" "	" "	"*"
3pred	" "	"*"	" "	" "	"*"	" "	" "	" "	"*"
2pred	" "	" "	" "	" "	"*"	" "	" "	" "	"*"
1pred	" "	" "	" "	" "	" "	" "	" "	" "	"*"

Backward selection matrix for gam

For the splines I used 4 <u>effective degrees of freedom</u>. 4 df is the same as for a cubic polynomial, so it allows for a reasonable amount of flexibility ('S'-like shapes).

The gam-model with the lowest AIC-score contains 6 predictors :

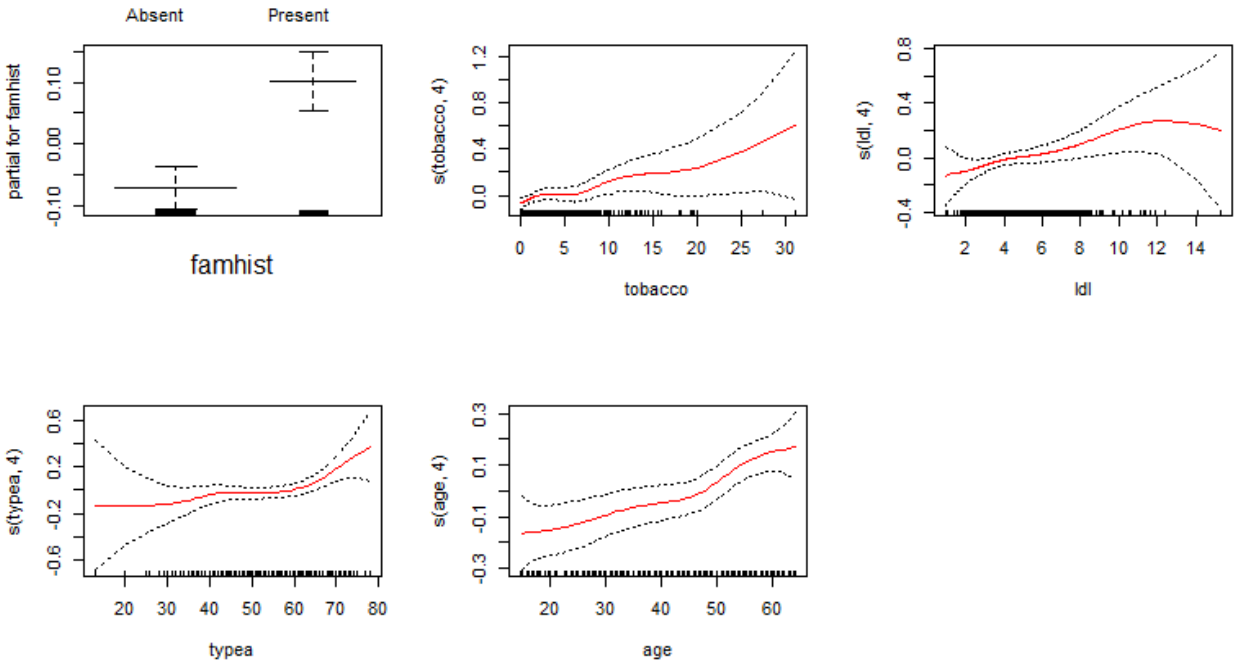
```
chd ~ 1 + famhist + s(tobacco, 4) + s(ldl, 4) + s(typea, 4) + s(obesity, 4) + s(age, 4)
```

these are the same predictors as obtained with glm **+ *obesity***
You can also see that leaving *obesity* out of the model only gives a marginal increase in AIC. So picking this 5-predictor model might be a good idea. If I then check the p-values of the coefficients I see that obesity is not significant

```
Anova for Parametric Effects
               Df Sum Sq Mean Sq F value    Pr(>F)
famhist         1  8.213  8.2126 47.4025 2.003e-11 ***
s(tobacco, 4)   1  8.500  8.5000 49.0609 9.346e-12 ***
s(ldl, 4)       1  4.168  4.1676 24.0550 1.320e-06 ***
s(typea, 4)     1  0.793  0.7933  4.5788   0.03292 *
s(obesity, 4)   1  0.109  0.1090  0.6292   0.42808
s(age, 4)       1  4.720  4.7201 27.2438 2.770e-07 ***
Residuals     440 76.231  0.1733
---
```

So, I drop *obesity* and I'm left with the final model that has the same 5 predictors as in the glm-fit

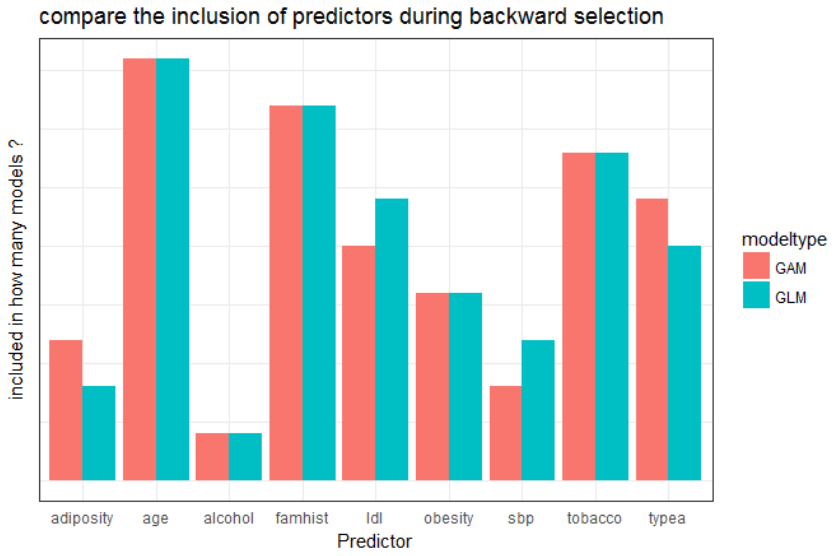
```
chd ~ 1 + famhist + s(tobacco, 4) + s(ldl, 4) + s(typea, 4) + s(age, 4)
```



the GAM model

All in all, the gam and glm backward selection fitting seem very similar (cfr figure below) :

- The 3 predictor model contains : *age, famhist, tobacco.* Selected in the same order in glm and gam
- *Typea* shows a fairly low correlation(0.1) with *chd*. But it still present in the 4-predictor gam model (not in the 4-predictor glm-model). This probably indicates a more non-linear relationship with the response (correlation only captures a linear relationship)



which predictors are present in the models during backward selection ?

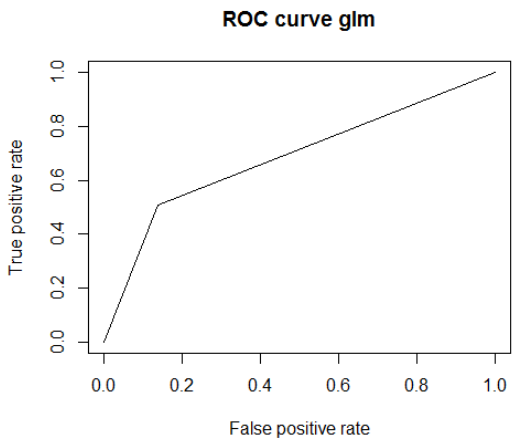
Signs of non-linearity :

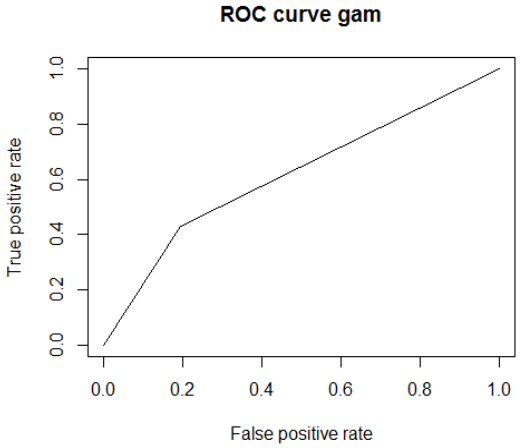
```
Anova for Nonparametric Effects
                Npar Df  Npar F   Pr(F)
(Intercept)
famhist
s(tobacco, 4)         3 1.67025 0.17267
s(ldl, 4)             3 0.57038 0.63477
s(typea, 4)           3 2.56850 0.05388
s(obesity, 4)         3 1.99535 0.11395
s(age, 4)             3 0.97333 0.40513
```

The signs of non-linearity is weak. Out of all the predictors only *typea* seems to show a reasonable amount of non-linearity. Using a smoothing spline for the other predictors seems not necessary here, a linear fit is probably sufficient.

5. Compare the performance of the linear and nonlinear models. Discuss your result

To compare the performance I used a single validationset (1 fold) with a train/test ratio of 2/3



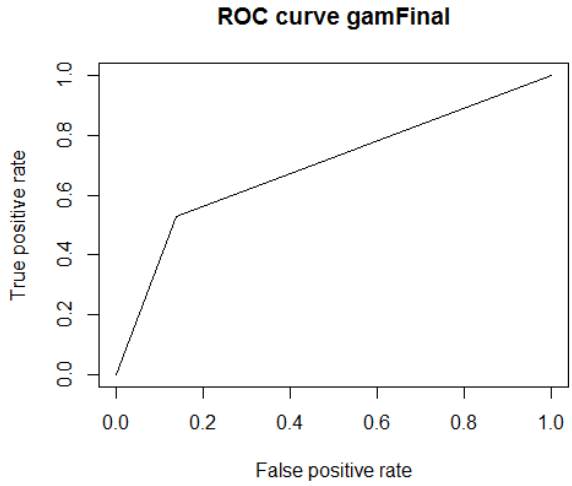


Misclassification rate GLM : 25.3%
Misclassification rate GAM : 31.8%
AUC GLM : 0.69
AUC GAM : 0.62

The GLM-fit seems to perform better than the GAM-fit. This comes at no surprise, because the non-linear association is very weak (cfr question 4). If the underlying association is indeed fairly linear, a linear fit will (almost always) outperform the non-linear fit because the variance of this linear fit is smaller (less parameters to fit)
Overall, the performance is not fantastic. The number of false positives is reasonable (14 out of 103), but the glm-fit could only detect half of the people with mycardial infarction (26 out of 51).

Final model
As stated above, only *typea* seems to have a (borderline) significant non-linear relationship with *chd*. So let's see how this model will perform

```
chd ~ 1 + tobacco + ldl + famhist + s(typea,4) + age
```



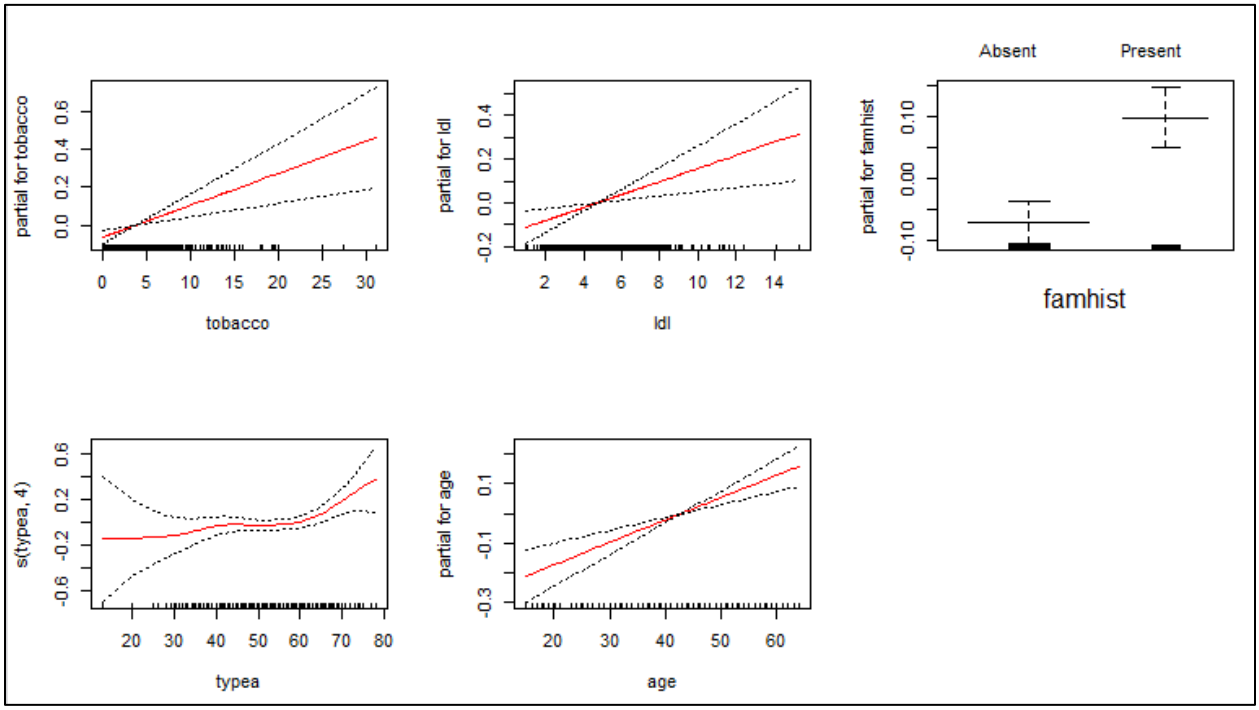
Misclassification rate Final model : 24.7%
AUC Final model : 0.70

A very slight improvement over the glm-fit, but improvement nevertheless.

Doing an anova test on all models (going from linear to non-linear) we see the same picture : the non-linear model does not explain a significant amount of the variance in *chd* . Making *typea* non-linear can be considered (borderline case)

```
Analysis of Deviance Table

Model 1: chd ~ 1 + tobacco + ldl + famhist + typea + age
Model 2: chd ~ 1 + tobacco + ldl + famhist + s(typea, 4) + age
Model 3: chd ~ 1 + famhist + s(tobacco, 4) + s(ldl, 4) + s(typea, 4) +
    s(age, 4)
  Resid. Df Resid. Dev     Df Deviance      F  Pr(>F)
1       456     80.686
2       453     79.405 3.0002   1.2810 2.4435 0.06352 .
3       444     77.582 8.9999   1.8232 1.1594 0.31950
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



the final model