GAMs report

Introduction:

GAMs is the addition is smooth functions of covariates and linear effects. GAMs provides a tradeoff between the interpretability of linear models or generalized linear models and the flexibility of non parametric complex approaches like Neural networks or trees. They can easily fot nonlinear functions to get predictions and obtain some inference about the model as well. Data is fitted using smooth functions(made up from basis functions) or splines to accommodate for the nonlinearity in the relationship between variables. To work with GAMs in R the mgcv library was used. The model is fitted using the gam() function with the formula similar to that of LM or GLM however we introduce the s() smoothing function for the nonlinear covariate. The command looks something like this:

gams\_model <- gam(response ~ s(nonlinear covariate 1)+ s(nonlinear covariate 2)+ linear covariate, data = mydataset, family = poisson(link= "log"), method="REML"))

The number of basis in the smooth obtained impacts the model shape greatly. This could be changed by changing the value of k defined for the smooth as follows: s(x1, k= no# of splines). As k increases the smooth fits the data more. We can also change the smoothing parameter to balance between likelihood and wiggliness optimizing model fit. Increasing the smoothing parameter sp greatly will give us a linear effect.

Linear and non linear terms could be added.

Practical work:

First, a GAM was fitted using all the covariates. By using the gam.check() function, we were able to decide on which variables are already linear and does not need smoothing. This is indicated by the covariates with edf =1 (Estimated Degrees of Freedom). The higher the edf, the higher the complexity of the smooth: the curve becomes more wiggly. This function provides us with 4 plots as well: qq-plot, residual Vs. Linear predictor, histogram of residuals and response Vs. fitted. Low values of p indicate that rssdiuals are not randomly distributed, which means there is not enough basis function. In addition, a plot of residuals of smoothed functions vs. its covariate was plotted.

Summary(): we can determine which covariates are not statistically significant; thus removing them in the next fitted model.

Concurvity is the similar term to co-linearity in GLMs. It occurs when a predictor could be represented as a smooth of another predictor. The value ranges from 0 causing no problem to 1 indicating an issue.

Avoid putting multiple collinear variables in the same model.

How to determine a good model?

1. qq-plot: compare residuals of the model to a normal distribution. Residuals of a well fitted model would be almost a straight line
2. Residual vs. linear predictor: evenly distributed around zero
3. Histogram of residuals: if it is a symmetrical bell shape
4. Response vs. fitted values: formation of a straight line or clustering around one to one line
5. AIC (Akiake Information Criterion): estimate of prediction error, the lower the better
6. Plot of partial residuals: shows the non-linearity/linearity of the smooths obtained. Also, shows the significance of smooth terms (a horizontal line can pass through the 95% confidence interval)

Results and conclusion:

* First model: A model with the most of the variables with the following residual plot.

Graphical user interface, application

Description automatically generated

From this residual plot, we can conclude that the following variables should be smoothed are: total\_patients\_hospitalized and death. Whereas, these variables are not significant: positive and total\_cases

The gam.check() produced the following plots:

The QQ plot shows that the deviance residuals and theoretical residuals do not coincide. However, we have an almost normal distribution for the residual histogram. Residuals vs linear predictor shows that there is somehow an evenly distribution around 0 and finally, the response over fitted value nearly form a straight line.

AIC=904.0313 R-sq.(adj) = 0.997 Deviance explained = 99.7%.

The Root Mean Square Error(RMSE) of predicted values for patients in intensive care for 14 days compared to the actual values is 12.43594

For the next model, will start by linearizing some variables and remove the least significant ones.

Chart

Description automatically generated

* Second model:

The residual plots are as follows:

Chart

Description automatically generated

The gam.check() produced the following plots:

Diagram

Description automatically generated

It shows very little enhancement than the previous model.

AIC=904.3962 R-sq.(adj) = 0.997 Deviance explained = 99.7%.

The Root Mean Square Error(RMSE) of predicted values for patients in intensive care for 14 days compared to the actual values is12.434509352

We can get a new model by removing some least significant variables, with p>0.5

* Third model:

The residual plots are as follows:

Chart, scatter chart

Description automatically generated

The gam.check() produced the following plots:

Diagram

Description automatically generated

The histogram of residuals is worse than the previous model.

AIC=896.283 R-sq.(adj) = 0.997 Deviance explained = 99.7%.

The Root Mean Square Error(RMSE) of predicted values for patients in intensive care for 14 days compared to the actual values is12.272515

We remove the variable with the p>0.5

* Fourth model:

The residual plots are as follows:

Chart, scatter chart

Description automatically generated

The gam.check() produced the following plots:

Chart

Description automatically generated

The histogram of residuals is not the best but better than the previous model.

AIC=894.4661 R-sq.(adj) = 0.997 Deviance explained = 99.7%.

The Root Mean Square Error(RMSE) of predicted values for patients in intensive care for 14 days compared to the actual values is 9.85764

We can check the interaction between the 2 smoothed variables

* Fifth model:

The residual plots are as follows:

Diagram

Description automatically generated

The gam.check() produced the following plots:

Chart

Description automatically generated

AIC=920.273 R-sq.(adj) = 0.996 Deviance explained = 99.5%.

The Root Mean Square Error(RMSE) of predicted values for patients in intensive care for 14 days compared to the actual values is 16.447

CONCLUSION:

From all the models we have tried the fourth model has the lowest AIC and lowest RMSE.