EDA - Part 2 - Predictors exploration

Study on the relationship between claims frequency and driver's age.

To continue our data exploration, we want to verify if the relationship between the outcome and the potential predictors is linear. If not, what would be the methods to be used to assess for non-linearity.

```
# Load the data
df<-read.csv("C:\\Users\\William\\Documents\\Data Science - ML\\Pricing
Project_GLM_vs_GBM\\data.csv")

# Split train / test
set.seed(564738291) # seed
u <- runif(dim(df)[1], min = 0, max = 1)
df$train <- u < 0.7
df$test <- !(df$train)
```

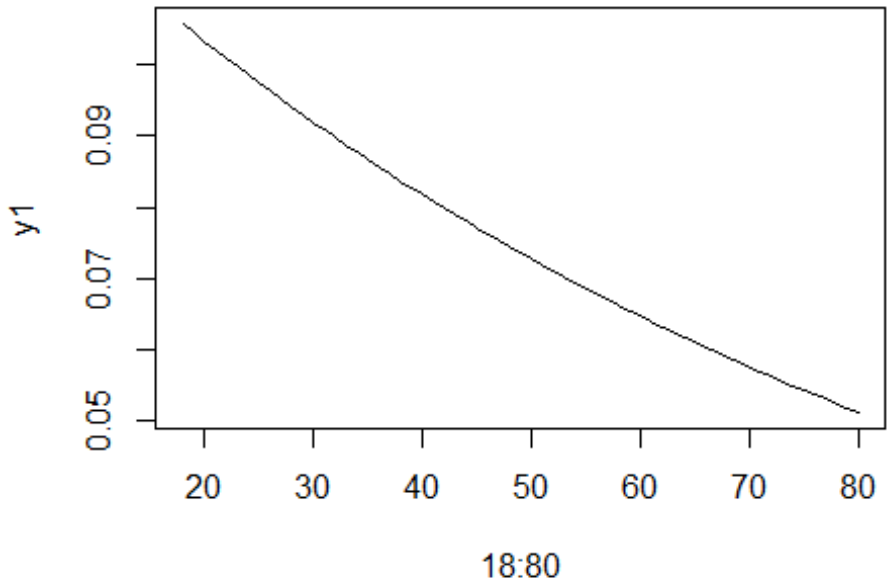
We run a very simple model with the driver age as unique parameter:

```
# Case 1: simple model - one predictor
model_age1 <- glm(formula = ClaimNb ~ DriverAge,
                  family = poisson(link = "log"),
                  data = df,
                  subset = train, offset = log(Exposure))
summary(model_age1)

##
## Call:
## glm(formula = ClaimNb ~ DriverAge, family = poisson(link = "log"),
##     data = df, subset = train, offset = log(Exposure))
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.0366101  0.0299239  -68.06   <2e-16 ***
## DriverAge   -0.0116775  0.0006382  -18.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 80222  on 289487  degrees of freedom
## Residual deviance: 79878  on 289486  degrees of freedom
## AIC: 103087
##
## Number of Fisher Scoring iterations: 6
```

Is the relationship between driver's age and claims frequency linear?

Let's check it by doing a prediction for each age between 18 and 80 years old:

```
y1 <- predict(model_age1, type = "response",
              newdata = data.frame(DriverAge = 18:80, Exposure = 1))

plot(18:80, y1, type = "l") # It is not linear, there is an exponential here
```




```
#plot(18:80, y1, type = "l", log="y") # in log, it is linear
```

On the first graph, we observe a light curve that indicates that the relation between the age of the driver and the claims frequency is not linear. In the second graph, we take the logarithm Using a linear model does not make a lot of sense. In fact, we can expect that driver will have less accident as they get more experienced, but is it the case for all the age band?

Proposal 1: Addition of splines

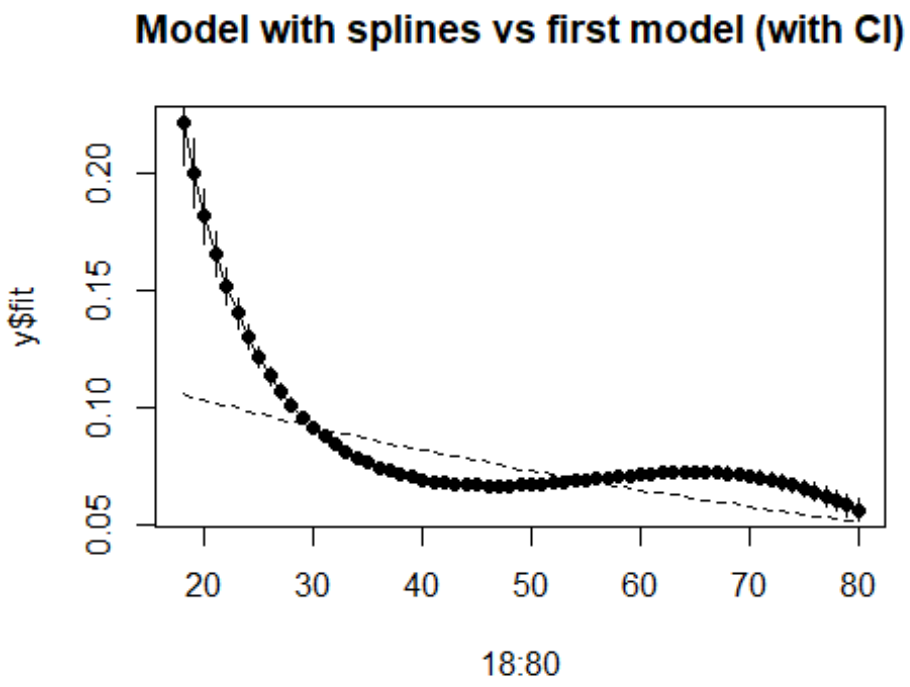
In statistics, a "spline" refers to a function constructed by piecing together different polynomials across specific intervals of the data, called "knots," resulting in a smooth curve that can be used to approximate relationships between variables. Here, we want to check if there are some knots and interpolate around them.

```
library(splines)

# Addition of splines
reg2 <- glm(ClaimNb ~ bs(DriverAge), data = df,
            subset = train, family = poisson,
```

```
            offset = log(Exposure))
summary(reg2) # We get a linear model on each of the transformation.

##
## Call:
## glm(formula = ClaimNb ~ bs(DriverAge), family = poisson, data = df,
##     subset = train, offset = log(Exposure))
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.50485    0.04165 -36.128  < 2e-16 ***
## bs(DriverAge)1 -2.85894    0.14414 -19.834  < 2e-16 ***
## bs(DriverAge)2  0.80901    0.15796   5.122 3.03e-07 ***
## bs(DriverAge)3 -3.01070    0.23287 -12.928  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 80222  on 289487  degrees of freedom
## Residual deviance: 79529  on 289484  degrees of freedom
## AIC: 102741
##
## Number of Fisher Scoring iterations: 6
```

3 splines are proposed by the function bs(), visible in the below graph.

```
# We compare with the previous linear model, without the splines
# Same process as above
nd <- data.frame(DriverAge = 18:80, Exposure = 1)
y <- predict(reg2, type = "response", newdata = nd, se.fit=TRUE)
plot(18:80, y$fit, type = "l")
#plot(18:80, y$fit, type = "l", log="y")

y1 <- predict(model_age1, type = "response", newdata = nd)
lines(nd$DriverAge,y1,lty=2 ) + title("Model with splines vs first model
(with CI)")

## integer(0)

points(nd$DriverAge, y$fit, pch=19)
segments(nd$DriverAge, y$fit-2*y$se.fit, nd$DriverAge, y$fit+2*y$se.fit)
```



The shape show an overall decrease as the more the driver is experienced, but with some variations. The frequency rate driven by experience translates into a high decrease on the young drivers, then stabilizes between 40 and 50 years old. Then we observe an increase after 50 until 65yo, followed by a decrease. We observe that we are 100% more in term of frequency on the young driver. The linear model is not adequate, so we need to take that non-linear effect into account.

Proposal 2: Tentative to model with a polynomial of order 3

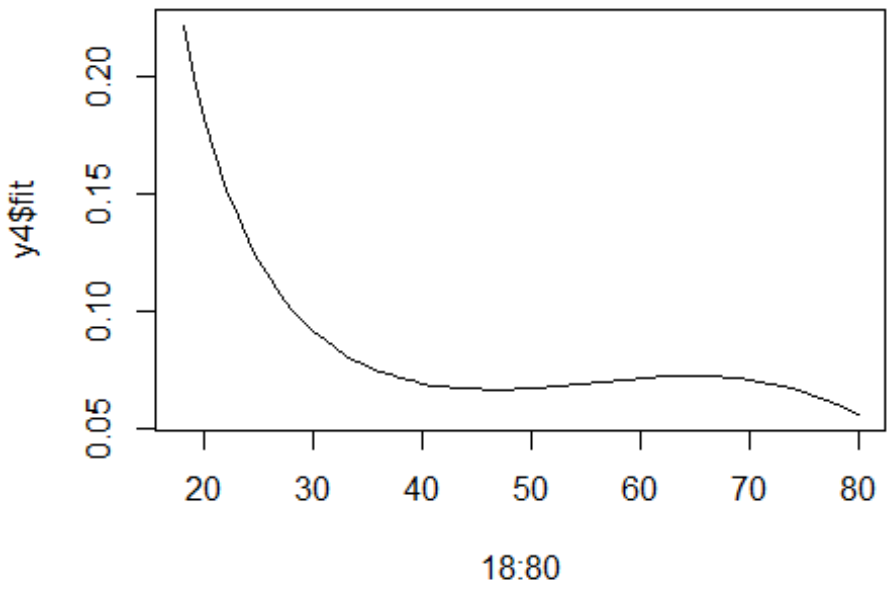
The cubic shape seen above tends to let us think that an order 3 could work:

```
reg_age3 <- glm(ClaimNb ~ poly(DriverAge,3), data = df,
                subset = train, family = poisson,
                offset = log(Exposure))
summary(reg_age3)
```

```
##
## Call:
## glm(formula = ClaimNb ~ poly(DriverAge, 3), family = poisson,
##     data = df, subset = train, offset = log(Exposure))
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.575e+00  9.063e-03 -284.11   <2e-16 ***
## poly(DriverAge, 3)1 -1.016e+02  5.531e+00  -18.36   <2e-16 ***
## poly(DriverAge, 3)2  6.541e+01  5.738e+00   11.40   <2e-16 ***
## poly(DriverAge, 3)3 -7.747e+01  5.894e+00  -13.14   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 80222  on 289487  degrees of freedom
## Residual deviance: 79529  on 289484  degrees of freedom
## AIC: 102741
##
## Number of Fisher Scoring iterations: 6
```

```r
y4 <- predict(reg_age3, type = "response", newdata = nd, se.fit = TRUE)

# With the IC
plot(18:80, y4$fit, type = "l") # same
```



Following the same methodology, we obtain a very similar shape as using the splines. Here all the coefficients are statistically significant.

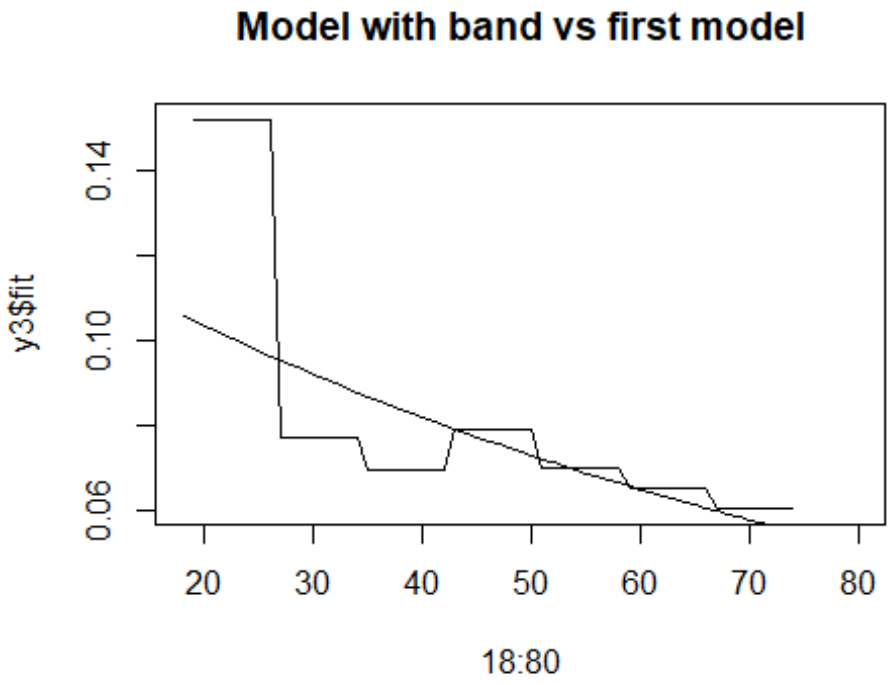
Proposal 3: Binning the age?

This solution is often imposed by the necessity of build a rate grid for commercial tariff. The grid is split by customer age band:

```r
reg3 <- glm(ClaimNb ~ cut(DriverAge, breaks = seq(18, 80, by=8)), data = df,
            subset = train, family = poisson,
            offset = log(Exposure))
summary(reg3)
```

```
##
## Call:
## glm(formula = ClaimNb ~ cut(DriverAge, breaks = seq(18, 80, by = 8)),
##     family = poisson, data = df, subset = train, offset = log(Exposure))
##
## Coefficients:
##                                                  Estimate Std. Error z
value
## (Intercept)                                      -1.88393    0.02585  -
72.87
## cut(DriverAge, breaks = seq(18, 80, by = 8))(26,34] -0.68218    0.03426  -
19.91
## cut(DriverAge, breaks = seq(18, 80, by = 8))(34,42] -0.78549    0.03360  -
23.38
## cut(DriverAge, breaks = seq(18, 80, by = 8))(42,50] -0.65677    0.03264  -
20.12
## cut(DriverAge, breaks = seq(18, 80, by = 8))(50,58] -0.78000    0.03401  -
22.93
## cut(DriverAge, breaks = seq(18, 80, by = 8))(58,66] -0.85053    0.04122  -
20.63
## cut(DriverAge, breaks = seq(18, 80, by = 8))(66,74] -0.92393    0.04609  -
20.05
##                                                  Pr(>|z|)
## (Intercept)                                        <2e-16 ***
## cut(DriverAge, breaks = seq(18, 80, by = 8))(26,34]   <2e-16 ***
## cut(DriverAge, breaks = seq(18, 80, by = 8))(34,42]   <2e-16 ***
## cut(DriverAge, breaks = seq(18, 80, by = 8))(42,50]   <2e-16 ***
## cut(DriverAge, breaks = seq(18, 80, by = 8))(50,58]   <2e-16 ***
## cut(DriverAge, breaks = seq(18, 80, by = 8))(58,66]   <2e-16 ***
## cut(DriverAge, breaks = seq(18, 80, by = 8))(66,74]   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 76835  on 278797  degrees of freedom
## Residual deviance: 76154  on 278791  degrees of freedom
##   (10690 observations deleted due to missingness)
## AIC: 98405
##
## Number of Fisher Scoring iterations: 6

y3 <- predict(reg3, type = "response", newdata = nd, se.fit = TRUE)
#print(y3)

# Replace the NA by 0
#y3_corrected <- y3 %>% replace(is.na(.), 0)

#plot(18:80, y3_corrected, type = "l")
#plot(18:80, y3_corrected, type = "l", log="y")
```

```r
# With the IC
plot(18:80, y3$fit, type = "l") # same

lines(nd$DriverAge,y1) + title("Model with band vs first model")
```




```
## integer(0)
```

Here the automatic split defines 6 age bands. In this case, we chose to obtain a constant claims frequency for each band. Moreover, we observe the same trend as in the previous models. The increase between 40 and 50yo is clearly visible.

Conclusion

It is obvious that the we need to take into account the non-linear effect of age on the claims frequency. In practice, a commercial tariff will propose a different price for different age band to account for that effect. For a unconstrained model, using splines or a GAM could be considered as good candidates.