Applied Statistics

Final coursework

Robin Mathelier, Imperial College London 7^{th}-11^{th} december 2020

Question 1

The data set was acquired in a psychophysical study concerned with the sensitivity of the human eye to colour. The response variable is the measure of the frequency level required to detect a change of chromatic (colour) stimulus. The data set gives different information on the experimental subjects which are listed below :

- Age – age in years of the experimental subject.
- Axis – a categorical variable indicating in which of three directions (in colour space) the change occurred. The levels are the axes of the colour space: Protan, Deutan and Tritan.
- IQ – The intelligence quotient of the subject, measured in various ways for different ages, and normalised.
- Gender – The gender of the subject.

The data set is composed of observations made on 459 experimental subjects. The main characteristics of the data set are summarised in table 1. There is a total of 109 missing values that we need to handle, this issue will be dealt in question 2.

Age	Thresh	Axis	Gender	IQ
Min. : 0.25	Min. :0.00041	Deutan:151	Female:218	Min. : 65
1st Qu.: 0.75	1st Qu.:0.00178	Protan:143	Male :223	1st Qu.: 94
Median : 1.00	Median :0.00705	Tritan:144	NA's : 18	Median :100
Mean : 9.85	Mean :0.00857	NA's : 21		Mean :100
3rd Qu.: 9.00	3rd Qu.:0.01089			3rd Qu.:106
Max. :86.00	Max. :0.05940			Max. :126
NA's :22	NA's :27			NA's :21

Table 1: Summary of the data

First of all, it is observed that the age's median is equal to 1 year old and the age's third quantile is equal to 9. Therefore, the sample is mainly constituted of children. One should be cautious concerning the conclusions drawn from the study of this data set that may be biased by this asymmetry of age. In particular, an attention is needed if one wants to generalize properties of the data set to the whole population.

For the other features, the samples seem to be representative of the population. The range of IQ in the data set is usual, the Axis values are balanced between the three levels possible, and we have approximately as many male as women. Nevertheless, it could seem weird to have IQ observations for infant a few months old. It should be checked how the IQs were measured and to what extent these measures are relevant. The IQ measures will not play an important role in our study so we do not deal more with this potential issue.

It is also noticed that the scales of the variables are different, in the order of 10^{-3} for $Thresh$ and in the order of 10^{1} for Age. It is chosen not to rescale the data in order to keep a simple interpretability of our results. Nevertheless, the plots will be often presented with a logarithm scale to gain clarity.

Then, a pairplot gives an idea of the relationships between the variables of the data set (Figure 1).
The pairplot suggests a dependency between the age and the oberved threshold for the experimental subjects. The dependency is not linear, a transformation of the variable Age may be considered before fitting a linear model to the data. The pattern of $Thresh$ against Age shows that a lot of observations are concentrated

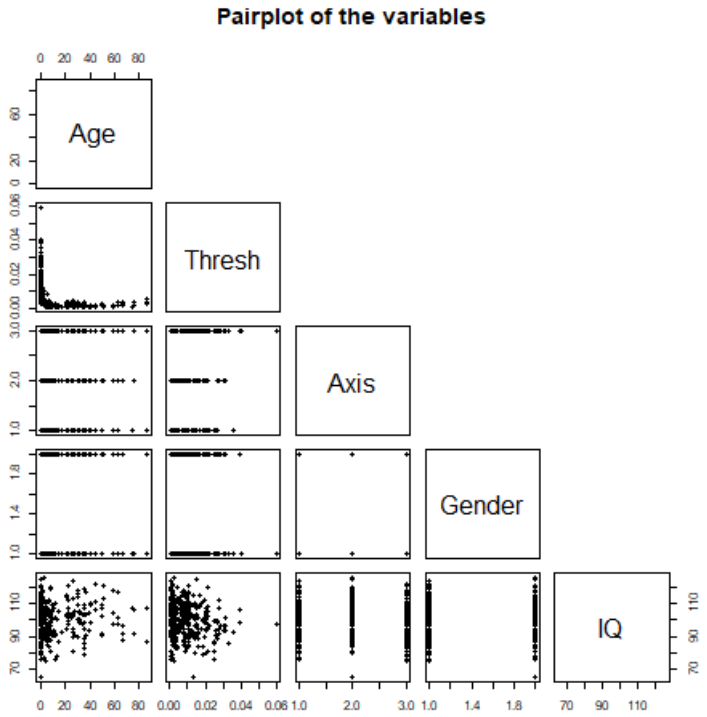

Figure 1: Pairplot of the different variables

with low age (less than 1 year) while some subjects are much older (more than 50 years). It leads us to introduce a new variable *Age_log* which is the base e logarithm of the variable *Age*. It is observed that the variable *Thresh* is more linear dependent to *Age_log* than to *Age* (figure 2) . It also helps to spread the observations making the plot more easy to interpret. The transformation is appropriate and the variable *Age_log* may be included in the data set.

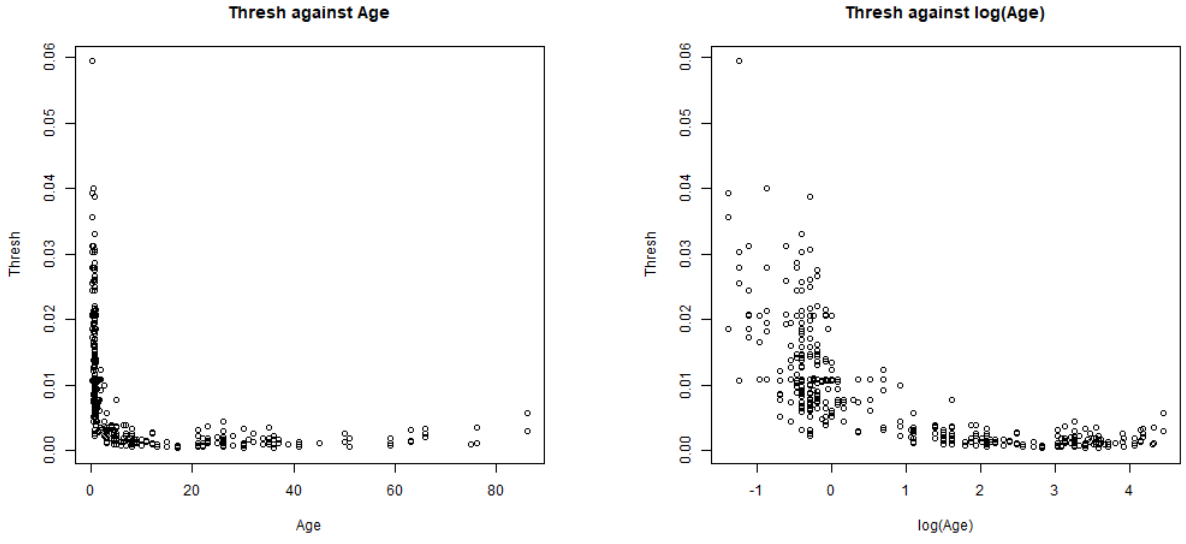

Figure 2: Observed thresholds against the age (left) and the logarithm of the age (right) of the subjects

The pairplot does not suggests a correlation between the measured IQ of the subjects and their observed threshold. The inclusion of this variable in our models may be uninformative and should be tested. For the discrete variables *Axis* and *Gender*, a boxplot and a scatter plot are more appropriate to detect a dependency with *Thresh*. The observed threshold does not seem to be impacted by the gender of the subject (figure 3). In the contrary, the axis influences the observed threshold. The subjects with axis Tritan are observed to get higher threshold than the subjects with axis Deutan or Protan (figure 4).

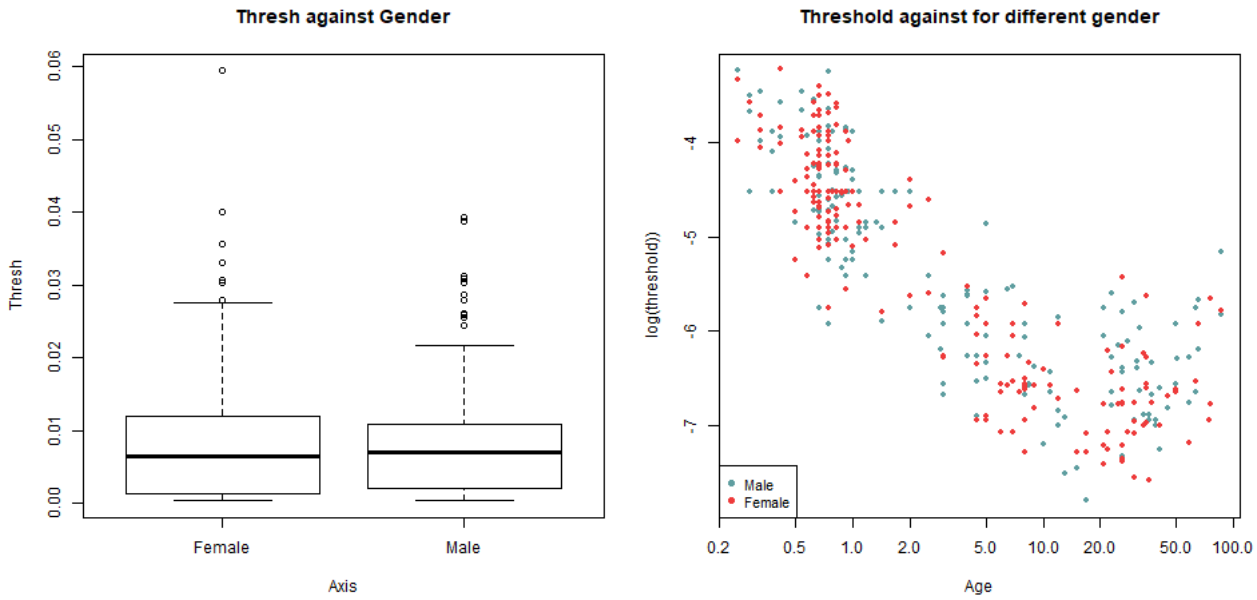

Figure 3: Observed thresholds against the gender of the subjects

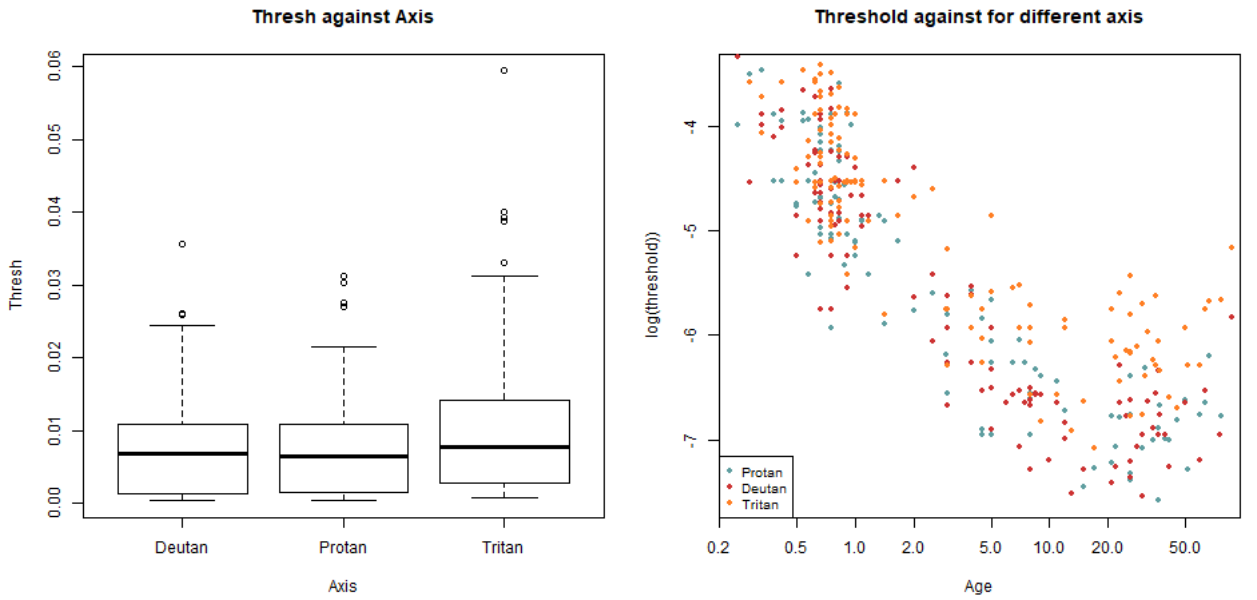

Figure 4: Observed thresholds against the axis of the subjects

From the different plots, it is observed that one observation has a very high threshold (0.06) compared to the others. It will be studied whether one should consider this observation as an outlier or not.

Question 2

As stated in question 1, an important stage before fitting a model is to handle the 109 missing values of our data set. Two methods are proposed and described below.

The listwise deletion consists in deleting the observations that have at least one missing value. This method is really simple to execute in practice, because it is only needed to detect the presence of a missing value row by row. The listwise deletion causes a lost of data that could be a major issue for our study. For example, an observation with only one missing value for IQ will be deleted of the data set. Processing like that, the information brought by this observation for the other variables of the data set is lost. In the studied data set, 93 observations have at least 1 missing value. the lost of data with listwise deletion represents 20%

of the whole data set. Another issue would occur if the missing values concern only a specific part of the studied population. For example, some old people may choose not to give their age and the observations for this part of the population would be lost. It could bring a bias in our model. Before completing a listwise deletion, one has to check that the missing values are randomly distributed among the studied population.

The (simple) Mean imputation consists in replacing the missing values for a variable by the mean value for that variable. This method helps counter the lost of potentially important observations evocated for list deletion. Another advantage is that keeping more observations in our data set will make our models more robust to outliers. However, a particular attention needs to be paid concerning the relevancy of the data set with replaced values compared to the original studied data. In particular, it has to be checked that the main characteristics and properties of the data set have not been altered by the mean imputation. Another issue is raised for the discrete variables : should one replace the missing values by the main represented class in the data set ? This can be not relevant when the number of observations for each level is the same (or almost the same). In this case, it could proposed to choose the level randomly, but again this may bring a bias.

In both case, a particular attention needs to be paid to the relevancy of the new data created after processing the missing values. For our data set, the mean computation for the *Age* variable produces a bias, because a peak of threshold is observed at the mean age 9.63 years that was not present in the original data set (figure 5).

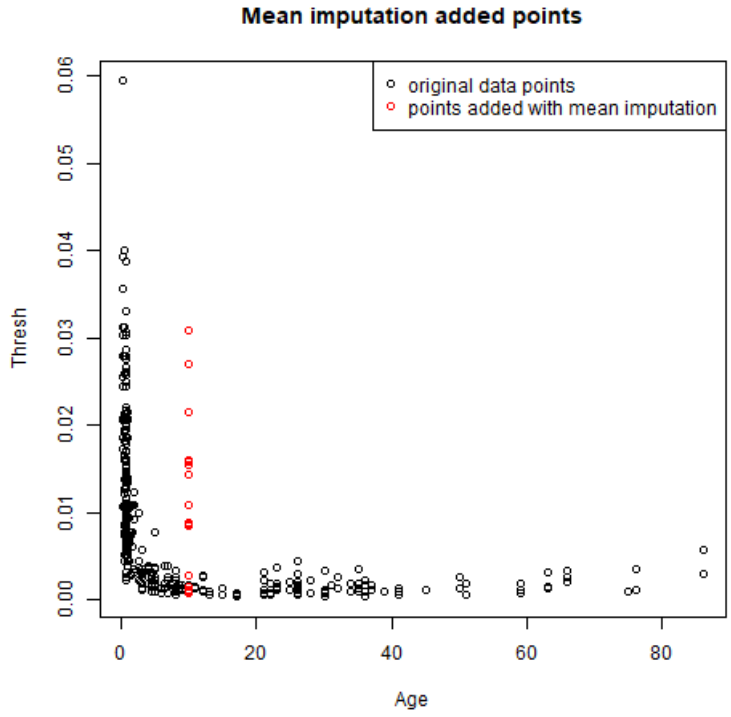

Figure 5: Observed thresholds against the age for original data (left) and for data obtained with mean computation (right)

We choose to produce a data set with the listwise deletion. It has been checked that the missing values seem not to come from the same group of the population. The data set produced is composed of $N = 366$ observations.

Question 3

The predictors *Gender* and *Axis* are discrete, then we need to introduce dummy regressors in our model. Treatment contrast is used and the dummy regressors are defined in Table 2.

Gender	G
Female	0
Male	1

Axis	$A^{(1)}$	$A^{(2)}$
Deutan	0	0
Protan	1	0
Tritan	0	1

Table 2: Dummy regressors for *Gender* and *Axis* predictors

The following simple linear model *mylm_0* with all the original predictors of the data is fitted in *R* :

$$Thresh_i = \beta_0 + \beta_1(Age)_i + \beta_2(IQ)_i + \beta_3 G_i + \beta_4 A_i^{(1)} + \beta_5 A_i^{(2)} + \epsilon_i \tag{1}$$

for $i = 1, ..., N$. We have the hypothesis $\epsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ for $j = 1, ..., N$ with $\sigma > 0$.
As stated before, the relationship between *Thresh* and *Age* is not linear. We obtained an adjusted R^2 equal to 0.2238 which indicates that the model performs poorly. As recommended in question 1, we fit another model *mylm_1* with the new variable $Age\_log = \log\{Age\}$:

$$Thresh_i = \beta_0 + \beta_1(Age)_i + \beta_2(IQ)_i + \beta_3 G_i + \beta_4 A_i^{(1)} + \beta_5 A_i^{(2)} + \beta_6 \log((Age)_i) + \epsilon_i \tag{2}$$

for $i = 1, ..., N$. We have the hypothesis $\epsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ for $j = 1, ..., N$ with $\sigma > 0$.
The performances seem to improve with an adjusted R^2 equal to 0.5676.

A transformation of the response *Thresh* may be considered to increase the fitting to the data. The box-cox transformation may be appropriate here. Before realising a box-cox transformation, it is first needed to detect potential outliers. The figure 6 shows that all observations have a cook distance lower than 0.5. Even though some observations have high leverages or residuals compared to the others, we choose to keep all the observations in this data set.

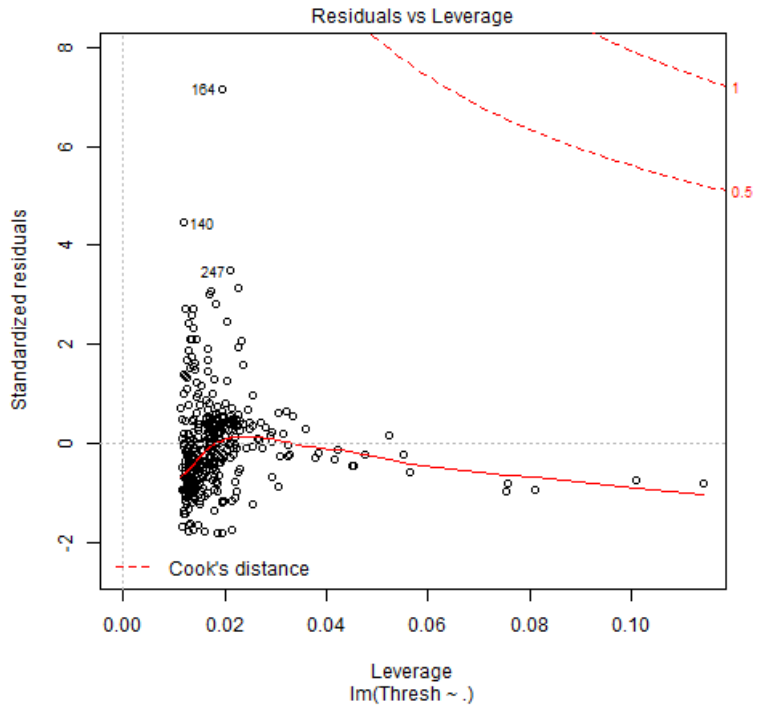

Figure 6: Cook distances

Furthermore, the ratio $\frac{\max_i(Thresh)_i}{min_i(Thresh)_i}$ can be calculated equal to 144.8 which seems high. The Box-Cox method is then appropriate to determine a transformation on the response under the form :

$$Thresh^{(\lambda)} = \begin{cases} \frac{Thresh^{\lambda}-1}{\lambda} \text{ if } \lambda \neq 0 \\ \log(Thresh) \text{ if } \lambda = 0 \end{cases} \tag{3}$$

The log-likelihood is maximized for $\lambda \simeq 0$ (figure 7). It suggests to use the transformation $Thresh^{(0)} = \log(Thresh)$.

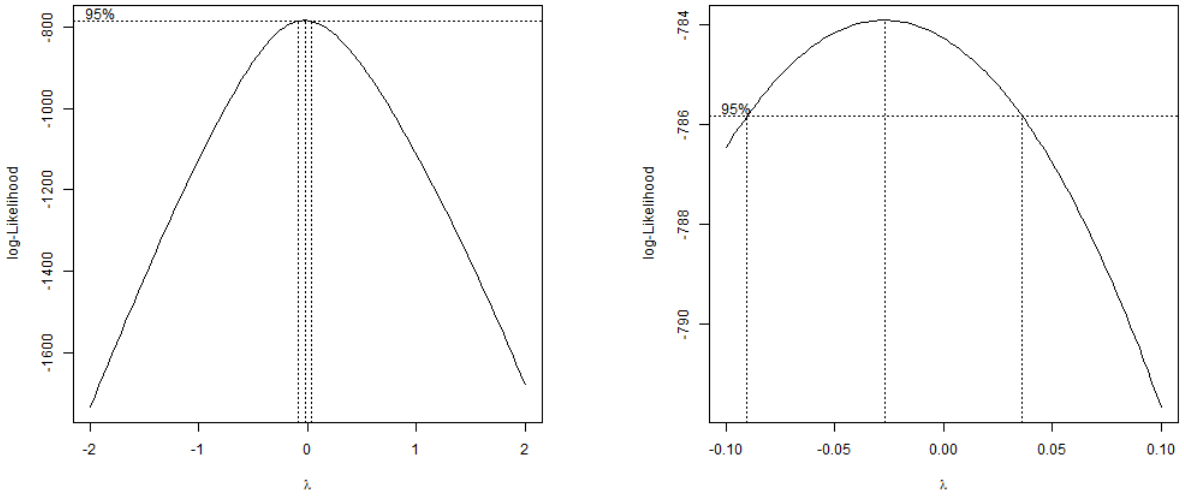

Figure 7: Log-likelihood against λ

Then, the following linear model $mylm\_2$ is fitted :

$$\log(Thresh_i) = \beta_0 + \beta_1(Age)_i + \beta_2(IQ)_i + \beta_3 G_i + \beta_4 A_i^{(1)} + \beta_5 A_i^{(2)} + \beta_6 \log\left((Age)_i\right) + \epsilon_i \tag{4}$$

for $i = 1, ..., N$. We have the hypothesis $\epsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ for $j = 1, ..., N$ with $\sigma > 0$.
The summary of the fitted model is given in Table 3.

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.3884	0.0214	18.16	$< 2\times10^{-16}$
Age	-0.0007	0.0002	-4.54	$< 2\times10^{-16}$
AxisProtan	0.0030	0.0050	0.60	0.5462
AxisTritan	0.0313	0.0048	6.45	3.41×10^{-15}
GenderMale	0.0032	0.0040	0.80	0.4230
IQ	-0.0000	0.0002	-0.01	0.9946
Age_inv	0.0697	0.0030	22.89	$< 2\times10^{-16}$

Table 3: Summary of $mylm\_2$

It is observed that predictors *Gender* and *Age* have high p-values for Student test. We conduct the following Fisher test to test the nullity of the associated parameters :

$$\begin{cases} H_0 : \beta_2 = \beta_3 = 0 \\ H_1 : \beta_2 \neq 0 \text{ or } \beta_3 \neq 0 \end{cases} \tag{5}$$

The obtained p-value is equal to 0.8511. For both test sizes 5% and 1% we fail to reject H_0. Therefore, the coefficients β_2 and β_3 wil be set equal to zero. It leads to fit the following linear model $mylm\_3$:

$$\log(Thresh_i) = \beta_0 + \beta_1(Age)_i + \beta_4 A_i^{(1)} + \beta_5 A_i^{(2)} + \beta_6 \log\left((Age)_i\right) + \epsilon_i \tag{6}$$

for $i = 1, ..., N$. We have the hypothesis $\epsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ for $j = 1, ..., N$ with $\sigma > 0$. The p-values for the student test calculated for the predictors are all significant (Table 4) except for the dummy regressor $A^{(1)}$. Nevertheless, we choose to conserve this predictor to keep the three levels of axis and conserve easy interpretability for this variable. Removing a predictor would decrease substantially the performances of the model, so the final normal linear model selected is $mylm\_3$. A diagnostic of this model is conducted. The adjusted R^2 is equal to 0.839 and suggests that the linear model $mylm\_3$ performs well. Furthermore, the predicted response follows the pattern of the original data (Figure 8).

	Estimate	Std. Error	t value	Pr(¿—t—)
(Intercept)	0.3899	0.0049	78.92	$< 2 \times 10^{-16}$
Age	-0.0007	0.0001	-4.54	$< 2 \times 10^{-16}$
AxisProtan	0.0029	0.0050	0.58	0.5640
AxisTritan	0.0313	0.0048	6.48	2.34×10^{-15}
Age_inv	0.0698	0.0030	22.95	$< 2 \times 10^{-16}$

Table 4: Summary of *mylm_3*

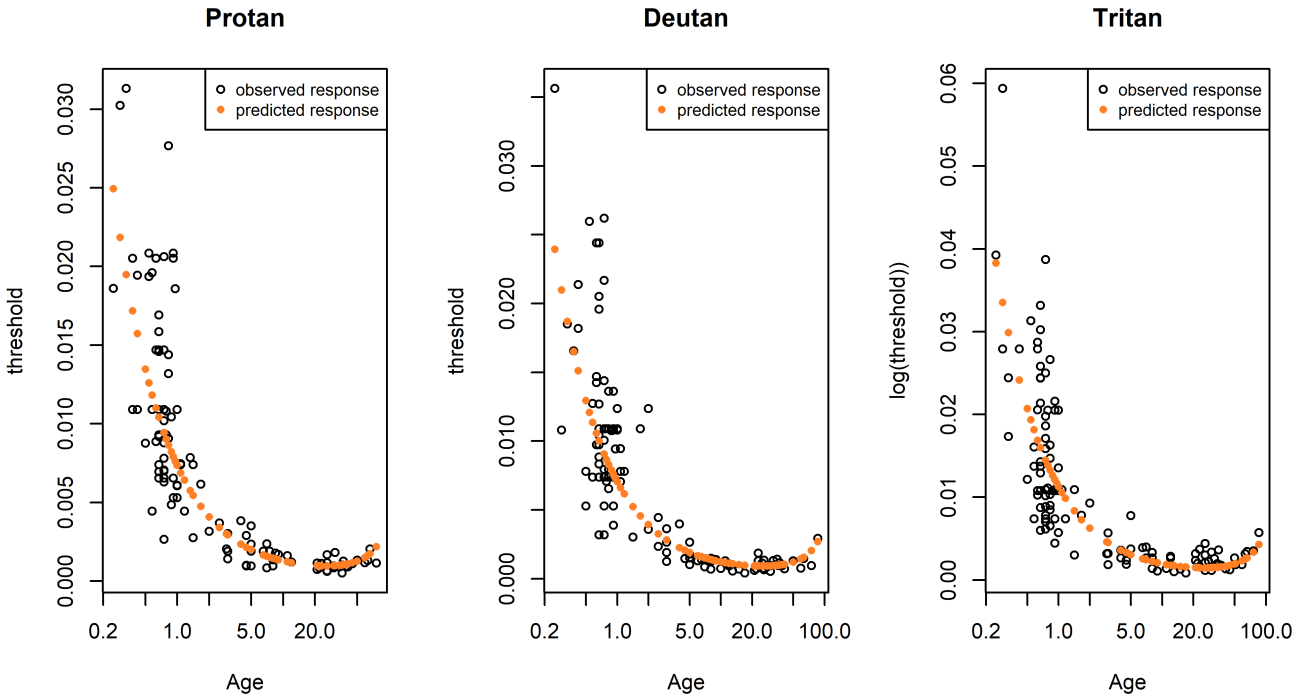

Figure 8: Fitted response for the normal linear model *mylm_3*

A plot of the residuals against the predicted response does not contradict the homoscedasticity of the residuals (Figure 9). However, it is observed some heterogeneity of the residuals as a function of the fitted values. On the quantile-quantile plot, the points closely follow a straight line. Therefore, the normality assumption for the residuals is not violated.

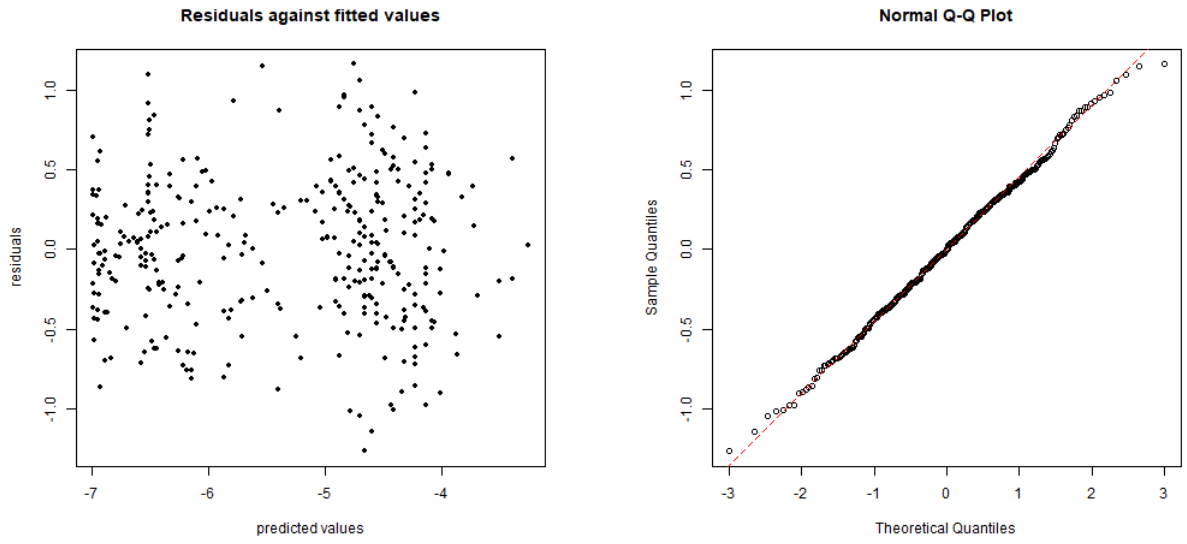

Figure 9: Diagnostics plots for the residuals of *mylm_3*

Question 4

It is first proved that the gamma distribution $\Gamma(\alpha, \beta)$, $\alpha > 0$, $\beta > 0$, is a two parameters exponential family. Recall the expression of the density f for $x > 0$:

$$f(x; \alpha, \beta) = x^{\alpha-1} \frac{\beta^\alpha e^{-\beta x}}{\Gamma(\alpha)} \tag{7}$$

where Γ is the gamma function. Introducing the parameters $\theta = -\beta/\alpha$ and $\phi = \alpha$, it is calculated :

$$\begin{aligned} f(x; \alpha, \beta) &= \exp\{(\alpha - 1)\log(x) + \alpha \log(\beta) - \beta x - \log(\Gamma(\alpha))\} \\ &= \exp\{(\phi - 1)\log(x) + \phi \log(-\theta\phi) + \theta\phi x - \log(\Gamma(\phi))\} \\ &= \exp\{\phi[x\theta - (-\log(-\theta))] + \phi \log(\phi) + (\phi - 1)\log(x) - \log(\Gamma(\phi))\} \end{aligned}$$

Therefore, $f(x, \theta, \phi)$ can be written in the form :

$$f(x, \theta, \phi) = \exp\left[\frac{x\theta - b(\theta)}{a(\phi)} + c(x, \phi)\right]$$

where it is identified :

- $\theta = -\beta/\alpha$
- $\phi = \alpha$
- $a(\phi) = 1/\phi$
- $b(\theta) = -\log(-\theta)$
- $c(x, \phi) = \phi \log(\phi) + (\phi - 1)\log(x) - \log(\Gamma(\phi))$

It is deduced that $\Gamma(\alpha, \beta)$ is a two parameters exponential family. A Gamma Generalized Linear Models is then defined by :

- The response variable is the random variable $\mathbf{Y}$ that takes its values in $(0, +\infty)^N$. The components of $\mathbf{Y} = (Y_1, .., Y_N)^T$ are independent and are distributed $Y_i \sim \text{Gamma}(\alpha_i, \beta_i)$ with $\alpha_i > 0$ and $\beta_i > 0$ for $i = 1, .., N$. It has be proved that the gamma distribution with parameters (α, β) is a two parameters exponential family. The expectation of $\mathbf{Y}$ verifies $\mathbf{E}(\mathbf{Y}) = \boldsymbol{\mu}$ such as $\forall i \in [1..N]$, $\mu_i = \frac{\alpha_i}{\beta_i}$.
- The predictors are denoted $\mathbf{X_1}, .., \mathbf{X_p}$ and the parameters of the model are $\boldsymbol{\beta} = (\beta_1, .., \beta_p)$. The design matrix is defined by $\mathbf{X} = [\mathbf{1}\, \mathbf{X_1} \, ... \, \mathbf{X_p}]$ (where $\mathbf{1}$ is the vector of $\mathbb{R}^N$ with all components equal to 1). The linear predictor takes the form $\boldsymbol{\eta} = X\boldsymbol{\beta}$.
- The link function is an invertible function g which provides a functional relationship between the linear predictor and the expectation of the response in the random component : $\eta_i = g(\mu_i)$ for $i = 1, .., n$.

As the support of a Gamma distribution is $(0, +\infty)$, a Gamma Generalized Linear Model is appropriate to fit real-valued response on a range from 0 to $+\infty$. This is the case in our problem as the response $Thresh$ is known to be strictly positive in the range of the data. Moreover, the Gamma Generalized Linear Model is suitable if there is reason to think that there is a linear dependency between the mean and the variance of the observed response. Indeed, recall that for a gamma distribution $Y \sim \Gamma(\alpha, \beta)$ we have $\mathbb{E}(Y) = \frac{\alpha}{\beta}$ and $Var(Y) = \frac{\alpha}{\beta^2}$. Then, $Var(Y) = \frac{1}{\beta}\mathbb{E}(Y)$ and the variance is linearly dependent of the mean. We plot the variance against the mean for $N = 10^4$ bootstrap samples of the observed thresholds. A linear dependency is well observed, and the Pearson correlation is calculated equal to $\rho = 0.68$. This suggests effectively that, in our problem, there is a linear dependency between the mean and the variance of the observed response. therefore the use of a Gamma Generalized Linear Model is appropriate.

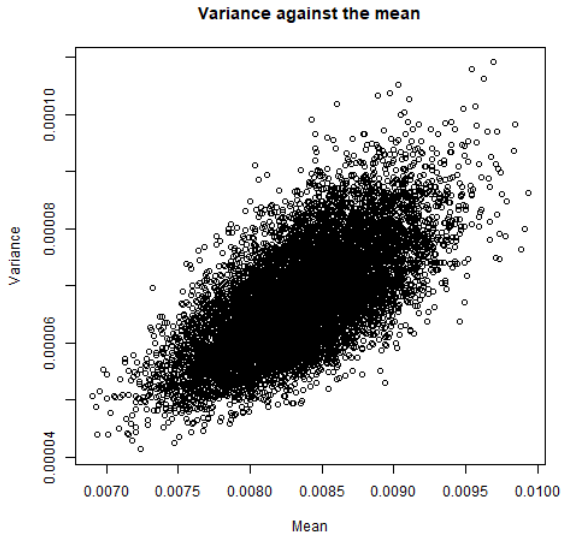

Figure 10: Linkage between mean and variance of bootstrapped thresholds

A first Generalized Linear Model glm_1 is fitted with response $Thresh$ and predictors $Axis$, Age and the reciprocal of Age denoted Age_inv. The link function g chosen is the identity and an intercept term is included in the model. Keeping the notations for dummy regressors defined in table 2, the linear predictor of glm_1 can be written under the form :

$$\eta_i = \beta_0 + \beta_1(Age)_i + \beta_2(1/Age_i) + \beta_3 A_i^{(1)} + \beta_4 A_i^{(2)} \tag{8}$$

for $i = 1, ..., N$

A second Generalized Linear Model glm_2 is fitted with response $Thresh$ and the predictors are the interaction term $Axis \times Age$ between Axis and Age on one hand, and the interaction term $Axis \times Age\_inv$ between $Axis$ and Age_inv on the other hand. The link function g chosen is the identity and an intercept term is included in the model. The linear predictor of glm_2 can be written under the form :

$$\begin{aligned}\eta_i = \beta_0 &+ (Protan)_i\,[\beta_1(Age)_i + \beta_2(1/Age_i)] \\ &+ (Deutan)_i\,[\beta_3(Age)_i + \beta_4(1/Age_i)] \\ &+ (Tritan)_i\,[\beta_5(Age)_i + \beta_6(1/Age_i)]\end{aligned}$$

for $i = 1, ..., N$, where $(Protan)_i = 1$ (resp. $(Deutan)_i = 1$, $(Tritan)_i = 1$) if $Axis_i$ ='Protan' (resp. if $Axis_i$ ='Deutan', if $Axis_i$ ='Tritan') and 0 otherwise.

Both the models fit well the data (figures 11 and 12). To compare the two generalized linear model, one can use the Akaike's Information Criteria (AIC) and the deviance residuals. The best model will be the one that minimises AIC and the deviance residuals. However, it is observed that the values for these two quantities are very close for the two models (table 5). One one hand, one can select the model glm_1 based on the fact that it has less parameters than glm_2. On the other hand, glm_2 gives the possibility to have different coefficients for each axis for both the predictors Age and $1/Age$. We choose to select this last option to get a better interpretability of our fitted coefficients.

	Deviance Residuals	AIC
glm_1	69.98	-3435.52
glm_2	69.19	-3435.81

Table 5: Deviance residuals and AIC calculated for glm_1 and glm_2

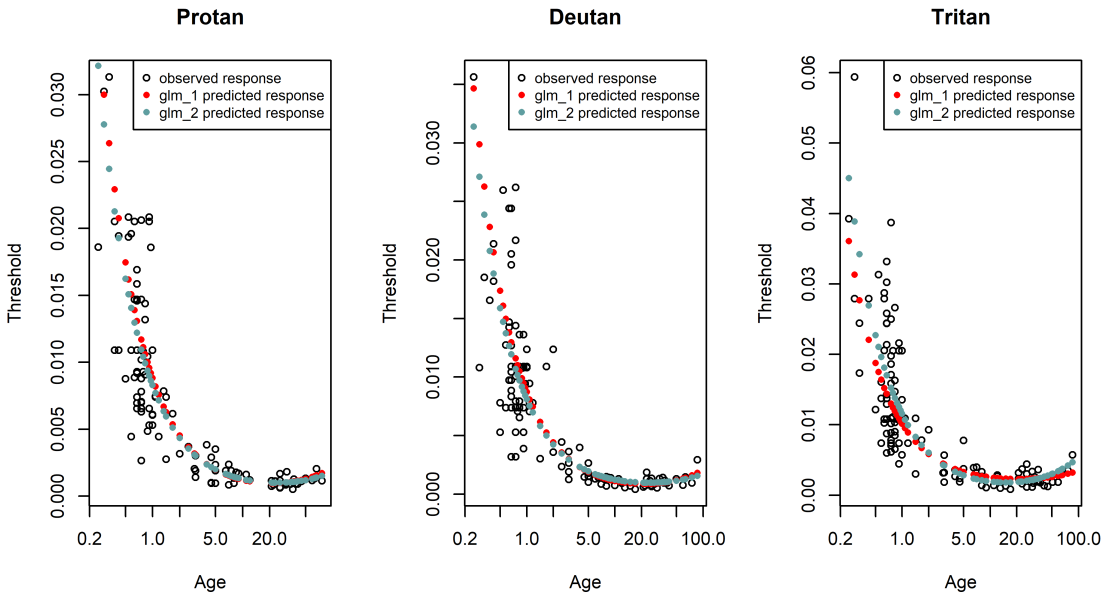

Figure 11: Fitted response for the generalized linear model glm_1

Question 5

We aim to compare the models $mylm\_3$ and glm_2 obtained respectively in question 3 and question 4. In terms of performance, the two models fit well the data (figure 12). The residuals sum of square for the logarithm of the threshold can be computed for the two models. The obtained values are $RSS_{mylm\_3} = 72.71$ and $RSS_{glm\_2} = 74.25$. The normal linear model obtains slighlty better predictions for the observed data than the generalized linear model. However, in terms of interpretability, a Gamma Generalized Linear Model with identity link is more suitable because the response are the untransformed thresholds. The estimated parameters can then be given a much more meaningful interpretation than for a simple linear model that predicts the logarithm of the threshold. Furthermore, the predictions and the confidence intervals obtained with $mylm\_3$ are less convenient because they are computed for the logarithm of the threshold. A transformation will be needed to use them in practice. On the contrary, glm_2 gives directly the prediction and properties for the threshold. To conclude, it is recommended to use the Generalized Linear Model glm_2 instead of $mylm\_3$

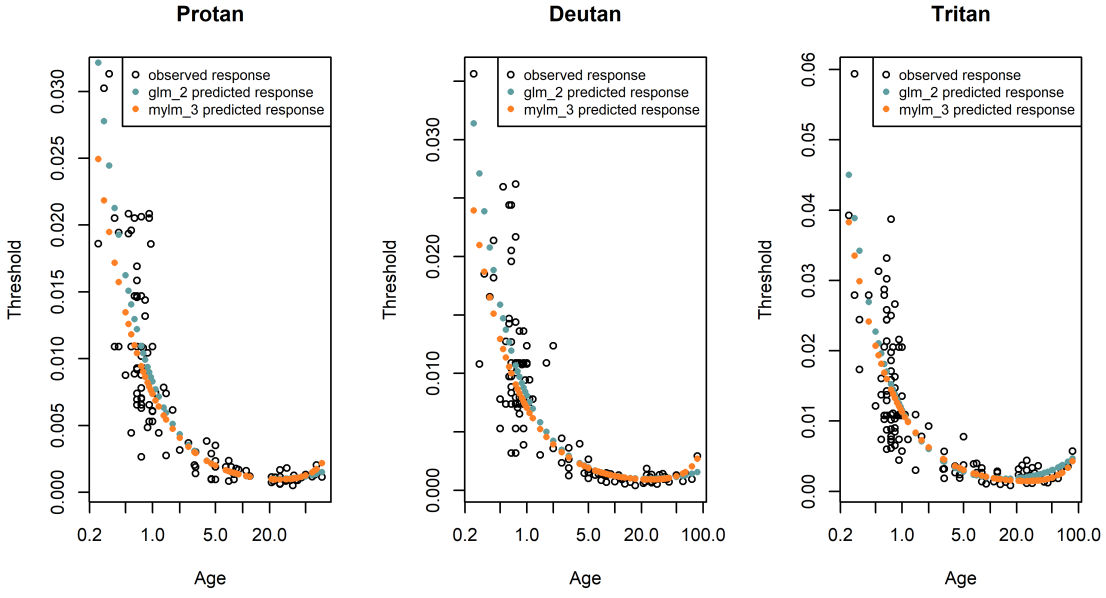

Figure 12: Fitted response for the generalized linear model glm_2

Appendices

A R code

```
###   MATH97125 - Applied Statistics, Coursework final, question 1  ###

### code of Robin Mathelier ###


setwd("C:/Users/robin/Dropbox/Applications/Overleaf/Coursework Final Applied") #
personnaliser
getwd()
.libPaths("C:/Users/robin/Documents/R/win-library/3.6")
rm(list=objects())

library(xtable)
library(latex2exp)
library(expm)
library(coda)
library(ggplot2)
library(MASS)
library('pracma')
library("lme4")
library(graphics)
library(faraway)
library(rcompanion)
library("RColorBrewer")


set.seed(42)

data = read.csv('CID1945214.csv')
head(data)
dim(data)

## Question 1 ##

summary(data)
xtable(summary(data, digits=3))

png(filename='pairplot_data_full.png')
pairs(data, upper.panel = NULL, pch = 16,cex=0.8,main='Pairplot of the variables')
dev.off()

par(mfrow=c(1,1))
png(filename='thresh_age.png')
plot(data$Thresh~data$Age,
     main='Thresh against Age',
     xlab='Age',
     ylab='Thresh')
dev.off()

png(filename='thresh_age_inv.png')
y = log(data$Age)
plot(data$Thresh~y,
```

```
      main='Thresh against log(Age)',
      xlab='log(Age)',
      ylab='Thresh')
dev.off()

par(mfrow=c(1,1))
png(filename = 'thresh_axis.png')
boxplot(data$Thresh~data$Axis,
        main='Thresh against Axis',
        xlab = 'Axis',
        ylab = 'Thresh')
dev.off()

png(filename = 'thresh_gender.png')
boxplot(data$Thresh~data$Gender,
        main='Thresh against Gender',
        xlab = 'Axis',
        ylab = 'Thresh')
dev.off()

png(filename='age_for_gender.png')
data_male = subset(data,data$Gender=='Male')
plot(data_male$Age,log(data_male$Thresh),
      main='Threshold against for different gender',
      xlab='Age',
      ylab='log(threshold))',pch=16,log='x',
      cex=0.7,col= 'cadetblue')
data_female = subset(data,data$Gender=='Female')
lines(data_female$Age,log(data_female$Thresh),
      main='Threshold against for different axis',
      xlab='Age',
      ylab='log(threshold))',pch=16,
      cex=0.7,col= 'brown2',type='p')
legend('bottomleft',
        c('Male','Female'),
        col=c('cadetblue','brown2'),
        pch=16,
        cex=0.9)
dev.off()

png(filename='Age_for_axis.png')
data_protan = subset(data,data$Axis=='Protan')
plot(data_protan$Age,log(data_protan$Thresh),
      main='Threshold against for different axis',
      xlab='Age',
      ylab='log(threshold))',pch=16,log='x',
      cex=0.7,col= 'cadetblue')
data_deutan = subset(data,data$Axis=='Deutan')
lines(data_deutan$Age,log(data_deutan$Thresh),
       main='Deutan',
       xlab='Age',
       ylab='log(threshold))',
       cex=0.7,pch=16,type='p',col='brown3')
data_tritan = subset(data,data$Axis=='Tritan')
lines(data_tritan$Age,log(data_tritan$Thresh),
       main='Tritan',
```

```
       xlab='Age',
       ylab='log(threshold))',
       cex=0.7,pch=16,type='p',col='chocolate1')
legend('bottomleft',
       c('Protan','Deutan','Tritan'),
       col=c('cadetblue','brown3','chocolate1'),
       pch=16,
       cex=0.8)
dev.off()

## Question 2 ##

# M1 Listwise deletion

sum(is.na(data))
list_lign_na = sapply(1:dim(data)[1],function(i) sum(is.na(data)[i,])>0)
sum(list_lign_na)
sum(list_lign_na)/dim(data)[1]

data_listwise_del = data[list_lign_na==F,]
summary(data_listwise_del)

# M2 Mean imputation

data_mean_imputation = data
col_cont = c('Age','Thresh','IQ')
for(col_name in col_cont){
  rep_val = mean(data_mean_imputation[,col_name],
                 na.rm=T)
  data_mean_imputation[,col_name][is.na(data_mean_imputation[,col_name])] = rep_val
}

data_mean_imputation$Gender[is.na(data_mean_imputation$Gender)] = 'Male'
data_mean_imputation$Axis[is.na(data_mean_imputation$Axis)] = 'Deutan'

summary(data_mean_imputation)

# comparison

summary(data)
summary(data_listwise_del)
summary(data_mean_imputation)

pairs(data, upper.panel = NULL, pch = 16,cex=0.8)
pairs(data_listwise_del,upper.panel = NULL, pch = 16,col='red',cex=0.8)
pairs(data_mean_imputation,upper.panel = NULL, pch = 16,col='green',cex=0.8)

png(filename = 'mean_imput.png')
par(mfrow=c(1,1))
plot(data$Age,data$Thresh,cex=1,
     main='Mean_imputation_added_points',
     xlab='Age',
     ylab='Thresh'
     )
lines(data_mean_imputation$Age[is.na(data$Age)],
     data_mean_imputation$Thresh[is.na(data$Age)],col='red',
```

```
      type='p',cex=1)
legend('topright',
       c('original data points','points added with mean imputation'),
       cex=1,pch=1,col=c(1,2))
dev.off()

df = data_listwise_del
xtable(summary(df, digits=3))

## Question 3 ##

# df0

df0 = df

mylm_0 <- lm(Thresh~.,data=df0)
summary(mylm_0,digit=3)
contrasts(df0$Axis)
contrasts(df0$Gender)

# df1

df1 = df
df1$Age_log = log(df1$Age)

mylm_1 <- lm(Thresh~.,data=df1)
summary(mylm_1,digit=3)

png(filename = 'cook.png')
plot(mylm_1,which=5) # distance cook << 0.5, no outliers
dev.off()

max(df1$Thresh)/min(df1$Thresh) # big => boxplot useful

png(filename = 'boxcox.png')
boxcox(mylm_1, plotit = TRUE)
dev.off()

png(filename = 'boxcox_zoom.png')
boxcox(mylm_1, plotit = TRUE, lambda = seq(-0.1,0.1, by = 0.1))
dev.off()

df2 = df1
df2$Thresh_log = log(df2$Thresh)
mylm_2 = lm(Thresh_log~.,data=df2[,-2])
contrasts(df2$Gender)
summary(mylm_2)
xtable(summary(mylm_2))

df3 = df2[,c("Age",'Axis','Age_log','Thresh_log')]
mylm_3 = lm(Thresh_log~.,data=df3)
summary(mylm_3,digit=4)
xtable(summary(mylm_3,digit=4))
levels(df3$Axis) # Toutes les variables doivent tre gard es sinon perte R^2

anova(mylm_2,mylm_3,test="F")
```

```
xtable(anova(mylm_2,mylm_3,test="F"),digit=4) # on garde mylm_3

png(filename = 'std_res_mylm_3.png')
par(mfrow=c(1,1))
x=predict(mylm_3)
y=mylm_3$res
plot(exp(x),y,main='Residuals against predicted threshold',
     xlab="Predicted threshold",ylab = 'Residuals',pch=16,cex=0.8)
dev.off()

png(filename = 'qqplots_mylm_3.png')
qqnorm(residuals(mylm_3))
qqline(residuals(mylm_3),col='red',lty=2)
dev.off()


png(filename = 'fitted_nlm.png',width = 17.5,height=10,units='cm',res=600)

par(mfrow=c(1,3))

df3_protan = subset(df3,df3$Axis=='Protan')
t_pred_protan = predict(mylm_3,df3_protan)
plot(df3_protan$Age,exp(df3_protan$Thresh),
     main='Protan',
     xlab='Age',
     ylab='threshold',
     cex=0.8,log=c('x','y'))
legend('topright',col=c(1,'chocolate1'),
       legend=c('observed response','predicted response'),
       ,cex=0.8,pch=c(1,16))
lines(df3_protan$Age,exp(t_pred_protan),col='chocolate1',type='p',pch=16,cex=0.8)

df3_deutan = subset(df3,df3$Axis=='Deutan')
t_pred_deutan = predict(mylm_3,df3_deutan)
plot(df3_deutan$Age,exp(df3_deutan$Thresh),
     main='Deutan',
     xlab='Age',
     ylab='threshold',
     cex=0.8,log=c('x','y'))
legend('topright',col=c(1,'chocolate1'),
       legend=c('observed response','predicted response'),
       ,cex=0.8,pch=c(1,16))
lines(df3_deutan$Age,exp(t_pred_deutan),col='chocolate1',type='p',pch=16,cex=0.8)


df3_tritan = subset(df3,df3$Axis=='Tritan')
t_pred_tritan = predict(mylm_3,df3_tritan)
plot(df3_tritan$Age,exp(df3_tritan$Thresh),
     main='Tritan',
     xlab='Age',
     ylab='log(threshold))',
     cex=0.8,log=c('x','y'))
legend('topright',col=c(1,'chocolate1'),
       legend=c('observed response','predicted response'),
       ,cex=0.8,pch=c(1,16))
lines(df3_tritan$Age,exp(t_pred_tritan),col='chocolate1',type='p',pch=16,cex=0.8)
```

```
dev.off()

## Question 4 ##

rep=data$Thresh
hist(rep)

# glm_1

glm_1 = glm(Thresh~Axis+Age+I(Age^-1),
            family = Gamma(link ="identity"),
            data=df)

summary(glm_1)
glm_1$coefficients

glm_2 = glm(Thresh ~ Age:Axis + Axis:I(Age^-1),
            family = Gamma(link ="identity"),
            data=df)

summary(glm_2)
glm_2$coefficients

png(filename = 'fitted_glm_1.png',width = 17.5,height=10,units='cm',res=600)

par(mfrow=c(1,3))

df_protan = subset(df,df$Axis=='Protan')
t_pred_protan = predict(glm_1,df_protan)
plot(df_protan$Age,df_protan$Thresh,
     main='Protan',
     xlab='Age',
     ylab='Threshold',
     cex=0.8,log=c('x','y'))
legend('topright',col=c(1,2,'cadetblue'),
       legend=c('observed response',
                'glm_1 predicted response',
                'glm_2 predicted response'),
       ,cex=0.8,pch=c(1,16,16))
lines(df_protan$Age,t_pred_protan,col='red',type='p',pch=16,cex=0.8)
t_pred_protan = predict(glm_2,df_protan)
lines(df_protan$Age,t_pred_protan,col='cadetblue',type='p',pch=16,cex=0.8)

df_deutan = subset(df,df$Axis=='Deutan')
t_pred_deutan = predict(glm_1,df_deutan)
plot(df_deutan$Age,df_deutan$Thresh,
     main='Deutan',
     xlab='Age',
     ylab='Threshold',
     cex=0.8,log=c('x','y'))
lines(df_deutan$Age,t_pred_deutan,col='red',type='p',pch=16,cex=0.8)
legend('topright',col=c(1,2,'cadetblue'),
       legend=c('observed response',
                'glm_1 predicted response',
                'glm_2 predicted response'),
```

```
        ,cex=0.8,pch=c(1,16,16))
t_pred_deutan = predict(glm_2,df_deutan)
lines(df_deutan$Age,t_pred_deutan,col='cadetblue',type='p',pch=16,cex=0.8)


df_tritan = subset(df,df$Axis=='Tritan')
t_pred_tritan = predict(glm_1,df_tritan)
plot(df_tritan$Age,df_tritan$Thresh,
     main='Tritan',
     xlab='Age',
     ylab='Threshold',
     cex=0.8,log=c('x','y'))
lines(df_tritan$Age,t_pred_tritan,col='red',type='p',pch=16,cex=0.8)
t_pred_tritan = predict(glm_2,df_tritan)
lines(df_tritan$Age,t_pred_tritan,col='cadetblue',type='p',pch=16,cex=0.8)
legend('topright',col=c(1,2,'cadetblue'),
       legend=c('observed response',
       'glm_1 predicted response',
       'glm_2 predicted response'),
       ,cex=0.8,pch=c(1,16,16))

dev.off()

dev_res = rbind(sum(residuals(glm_1,type="deviance")^2),
                sum(residuals(glm_2,type="deviance")^2))
AICs = rbind(AIC(glm_1),
             AIC(glm_2))
tab = data.frame(cbind(dev_res,AICs))
xtable(tab)

compareGLM(glm_1,glm_2)
?compareGLM
glm_1$residuals^2

sum((log(predict(glm_1,df))-log(df$Thresh))^2)
sum((log(predict(glm_2,df))-log(df$Thresh))^2)

f = function(x) sd(data$Thresh[data$Thresh<x],na.rm=T)
g = Vectorize(f)
f_mean = function(x) mean(data$Thresh[data$Thresh<x],na.rm=T)
g_mean = Vectorize(g)
abs=seq(0.005,0.06,0.001)
abs_mean = g_mean((abs))
ord_sd=g(abs)

png(filename = 'var_vs_mean.png')
n=length(df$Thresh)
x=c()
y=c()
for(i in 1:10000){
abs_boot=sample(1:n,replace=T)
thresh_boot=df$Thresh[abs_boot]
x = c(x,mean(thresh_boot))
y = c(y,var(thresh_boot))}
par(mfrow=c(1,1))
plot(x,y,
```

```
      main = 'Variance against the mean',
      xlab='Mean',
      ylab='Variance')
dev.off()

cor(x,y)

png(filename = 'scatter_plot_thresh.png')
plot(data$Thresh,
      main='scatter plot of the thresholds',
      xlab='i',
      ylab=expression(Threshold[i]))
dev.off()

## Question 5

png(filename = 'fitted_glm_2.png',width = 17.5,height=10,units='cm',res=600)

par(mfrow=c(1,3))

df_protan = subset(df,df$Axis=='Protan')
t_pred_protan = predict(glm_2,df_protan)
plot(df_protan$Age,df_protan$Thresh,
      main='Protan',
      xlab='Age',
      ylab='Threshold',
      cex=0.8,log=c('x','y'))
legend('topright',col=c(1,'cadetblue','chocolate1'),
        legend=c('observed response',
                 'glm_2 predicted response',
                 'mylm_3 predicted response'),
        ,cex=0.8,pch=c(1,16,16))
lines(df_protan$Age,t_pred_protan,col='cadetblue',type='p',pch=16,cex=0.8)
t_pred_protan = predict(mylm_3,df3_protan)
lines(df3_protan$Age,exp(t_pred_protan),col='chocolate1',type='p',pch=16,cex=0.8)

df_deutan = subset(df,df$Axis=='Deutan')
t_pred_deutan = predict(glm_2,df_deutan)
plot(df_deutan$Age,df_deutan$Thresh,
      main='Deutan',
      xlab='Age',
      ylab='Threshold',
      cex=0.8,log=c('x','y'))
legend('topright',col=c(1,'cadetblue','chocolate1'),
        legend=c('observed response',
                 'glm_2 predicted response',
                 'mylm_3 predicted response'),
        ,cex=0.8,pch=c(1,16,16))
lines(df_deutan$Age,t_pred_deutan,col='cadetblue',type='p',pch=16,cex=0.8)
t_pred_deutan = predict(mylm_3,df3_deutan)
lines(df3_deutan$Age,exp(t_pred_deutan),col='chocolate1',type='p',pch=16,cex=0.8)

df_tritan = subset(df,df$Axis=='Tritan')
t_pred_tritan = predict(glm_2,df_tritan)
plot(df_tritan$Age,df_tritan$Thresh,
      main='Tritan',
```

```
       xlab='Age',
       ylab='Threshold',
       cex=0.8,log=c('x','y'))
legend('topright',col=c(1,'cadetblue','chocolate1'),
         legend=c('observed response',
                  'glm_2 predicted response',
                  'mylm_3 predicted response'),
         ,cex=0.8,pch=c(1,16,16))
lines(df_tritan$Age,t_pred_tritan,col='cadetblue',type='p',pch=16,cex=0.8)
t_pred_tritan = predict(mylm_3,df3_tritan)
lines(df3_tritan$Age,exp(t_pred_tritan),col='chocolate1',type='p',pch=16,cex=0.8)


dev.off()


sum(mylm_3$res^2)
sum((log(predict(glm_1,df))-log(df$Thresh))^2)
```