GLM with R

Code

D.-L. Couturier / R. Nicholls / M. Fernandes

Last modified: 09 Mar 2020

Section 1: Logistic regression

We will analyse the data collected by Jones (Unpublished BSc dissertation, University of Southampton, 1975). The aim of the study was to define if the probability of having Bronchitis is influenced by smoking and/or pollution.

The data are stored under data/Bronchitis.csv and contains information on 212 participants.

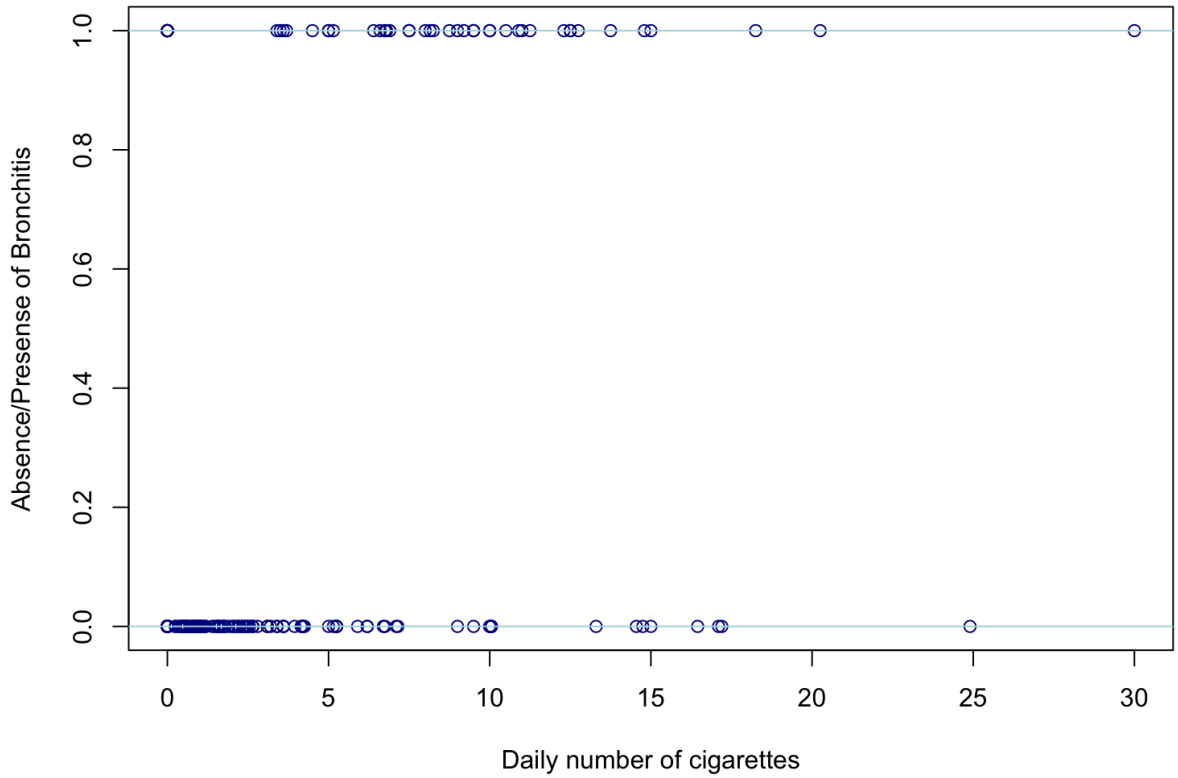
Section 1.1: importation and descriptive analysis

Lets starts by

- importing the data set *Bronchitis* with the function `read.csv()`
- displaying *bron* (a dichotomous variable which equals 1 for participants having bronchitis and 0 otherwise) as a function of *cigs*, the number of cigarettes smoked daily.

Hide

```
Bronchitis = read.csv("data/Bronchitis.csv",header=TRUE)
plot(Bronchitis$cigs,Bronchitis$bron,col="blue4",
       ylab = "Absence/Presense of Bronchitis", xlab = "Daily number of cigarette
        s")
abline(h=c(0,1),col="light blue")
```



Section 1.2: Model fit

Lets

- fit a logistic model by means the function `glm()` and by means of the function `gamlss()` of the library `gamlss` .
- display and analyse the results of the `glm` function : Use the function `summary()` to display the results of an R object of class `glm` .

Hide

```
fit.glm = glm(bron~cigs,data=Bronchitis,family=binomial)

library(gamlss)
fit.gamlss = gamlss(bron~cigs,data=Bronchitis,family=BI)
```

```
## GAMLSS-RS iteration 1: Global Deviance = 181.7072
## GAMLSS-RS iteration 2: Global Deviance = 181.7072
```

Hide

```
summary(fit.glm)
```

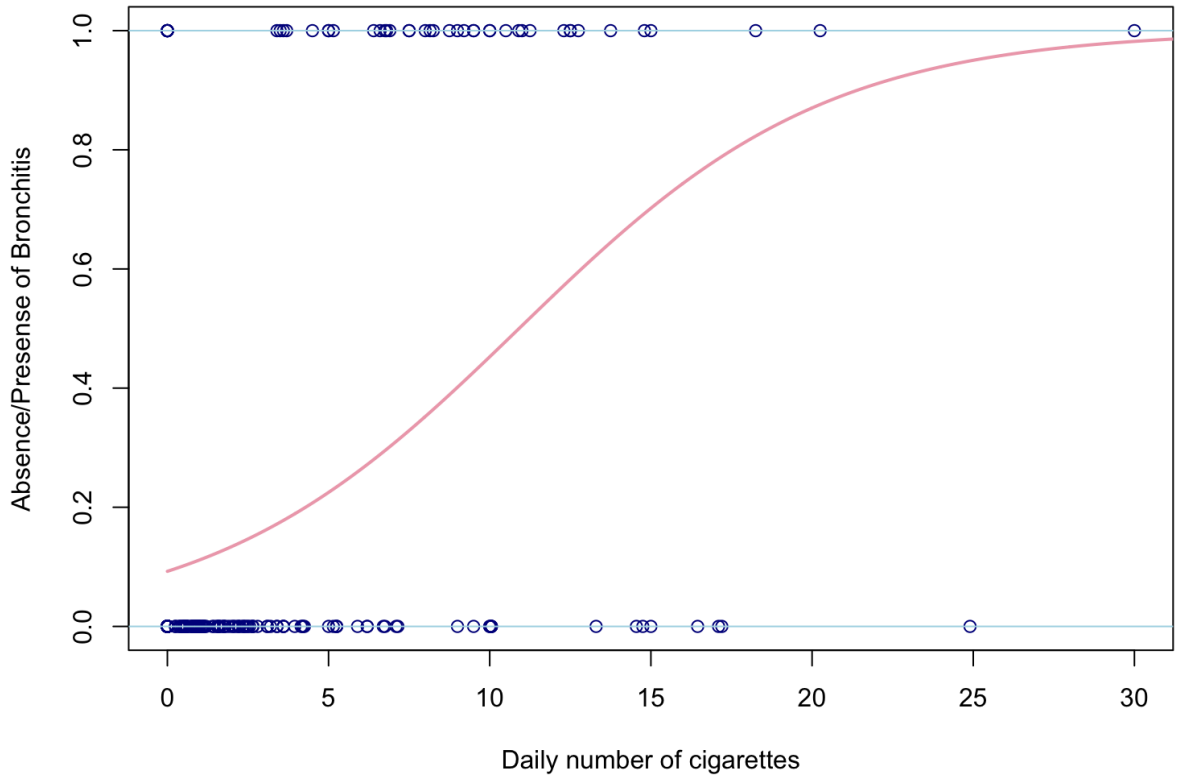
```
##
## Call:
## glm(formula = bron ~ cigs, family = binomial, data = Bronchitis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4418  -0.5472  -0.4653  -0.4405   2.1822
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.2840     0.2731  -8.365  < 2e-16 ***
## cigs          0.2094     0.0376   5.567 2.59e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 221.78  on 211  degrees of freedom
## Residual deviance: 181.71  on 210  degrees of freedom
## AIC: 185.71
##
## Number of Fisher Scoring iterations: 4
```

Let's now define the estimated probability of having bronchitis for any number of daily smoked cigarette and display the corresponding logistic curve on a plot:

Hide

```
plot(Bronchitis$cigs,Bronchitis$bron,col="blue4",
       ylab = "Absence/Presense of Bronchitis", xlab = "Daily number of cigarette
        s")
abline(h=c(0,1),col="light blue")

axe.x = seq(0,40,length=1000)
f.x = exp(fit.glm$coef[1]+axe.x*fit.glm$coef[2])/(1+exp(fit.glm$coef[1]+axe.x*fit.g
        lm$coef[2]))
lines(axe.x,f.x,col="pink2",lwd=2)
```



Section 1.3: Model selection

As for linear models, model selection may be done by means of the function `anova()` used on the glm object of interest.

Hide

```
anova(fit.glm,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: bron
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                  211     221.78
## cigs  1    40.07       210     181.71 2.45e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Section 1.3: Model check

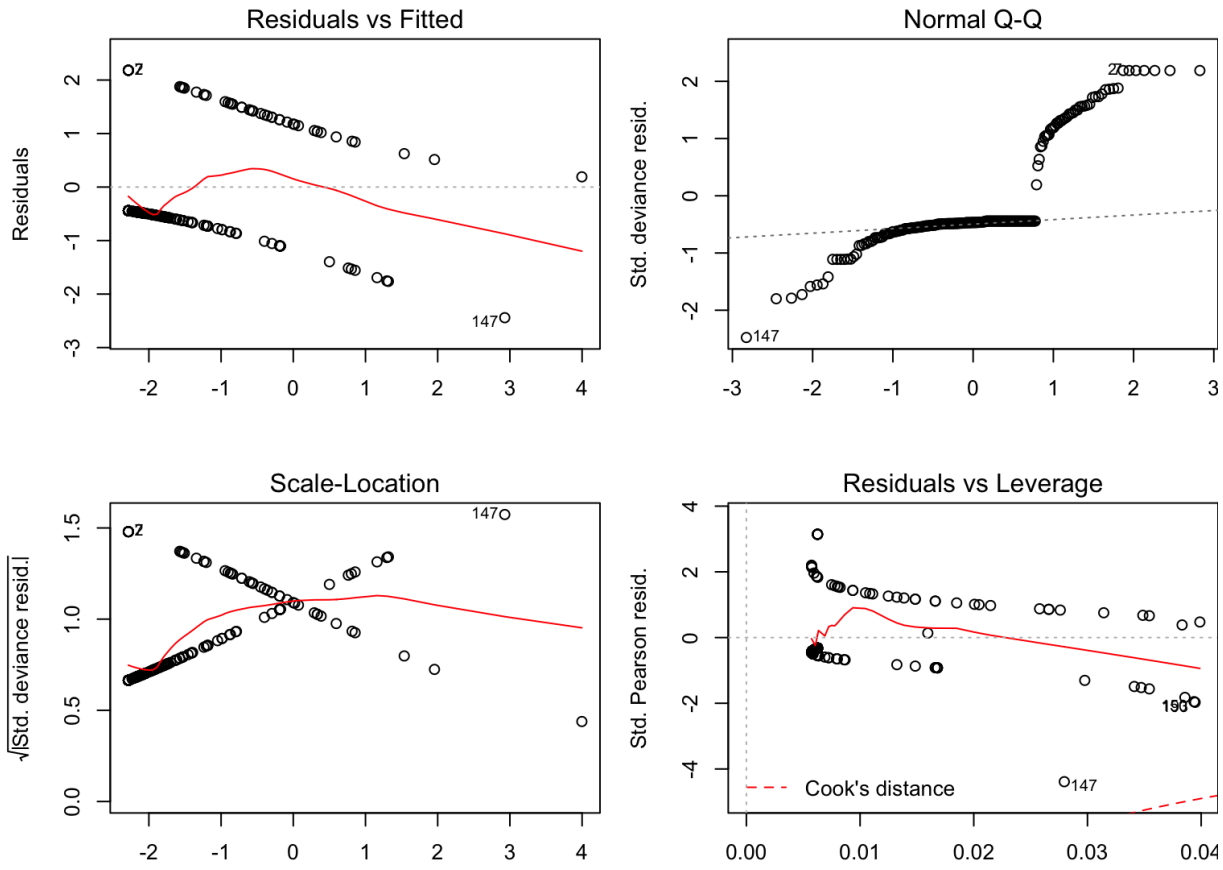
Lets assess is the model fit seems satisfactory by means

- of the analysis of deviance residuals (function `plot()` on an object of class `glm`,

- of the analysis of randomised normalised quantile residuals (function `plot()` on an object of class `gamlss`,

Hide

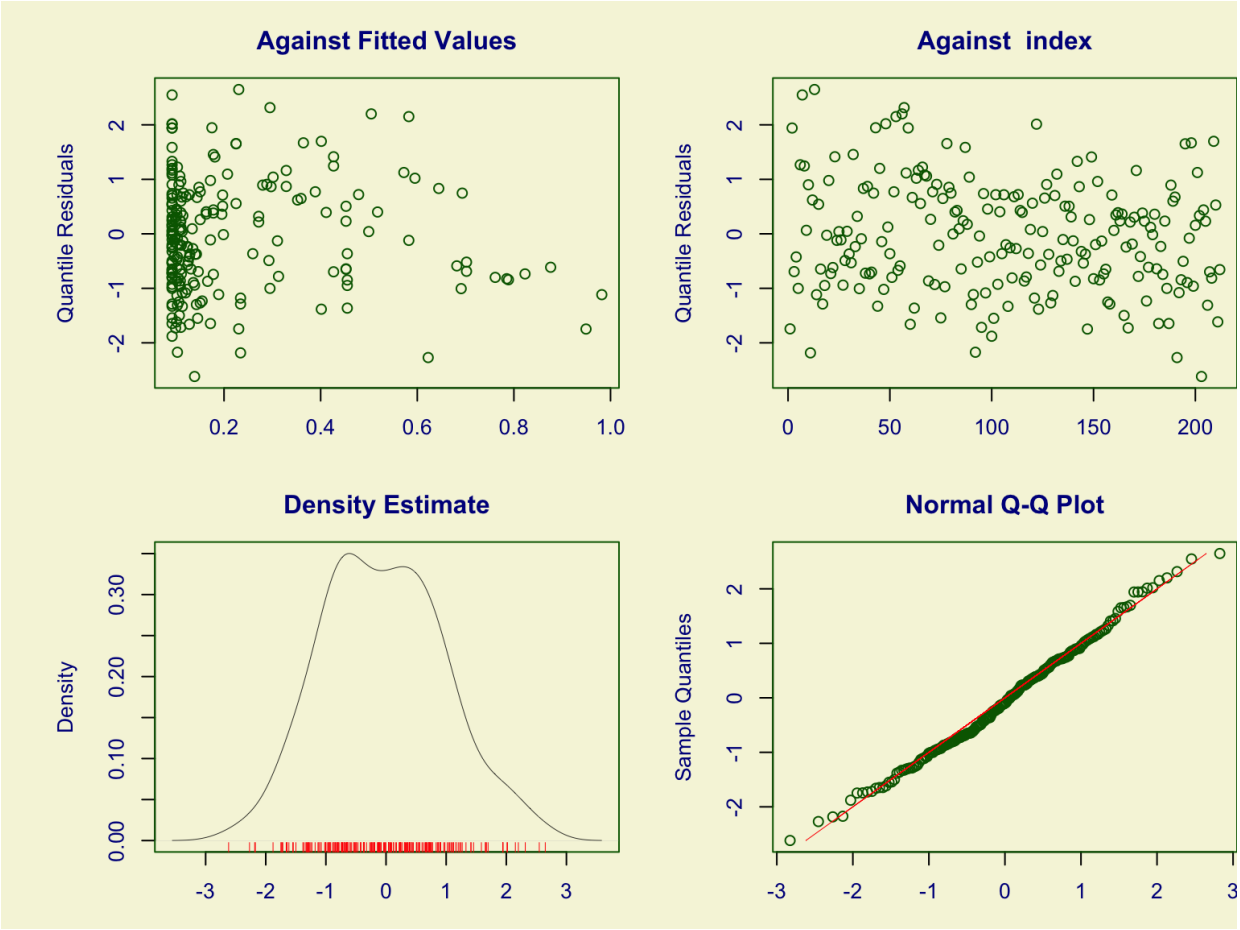
```
# deviance
par(mfrow=c(2,2),mar=c(3,5,3,0))
plot(fit.glm)
```



Hide

```
# randomised normalised quantile residuals
plot(gamlss(bron~cigs,data=Bronchitis,family=BI))
```

```
## GAMLSS-RS iteration 1: Global Deviance = 181.7072
## GAMLSS-RS iteration 2: Global Deviance = 181.7072
```




```
## ****************************************************************
##   Summary of the Randomised Quantile Residuals
##                            mean   =  -0.04544699
##                        variance   =  1.033809
##                coef. of skewness  =  0.1878859
##                coef. of kurtosis  =  2.676177
## Filliben correlation coefficient  =  0.9970683
## ****************************************************************
```

Section 1.4: Fun

Hide

```
# long format:
long = data.frame(mi = rep(c("MI","No MI"),c(104+189,11037+11034)),
                  treatment = rep(c("Aspirin","Placebo","Aspirin","Placebo"),c(104,
        189,11037,11034)))
# short format: 2 by 2 table
table2by2 = table(long$treatment,long$mi)
print(table2by2)
```

```
##
##              MI No MI
##   Aspirin   104 11037
##   Placebo   189 11034
```

Hide

```
#
chisq.test(table2by2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table2by2
## X-squared = 23.782, df = 1, p-value = 1.079e-06
```

Hide

```
prop.test(table2by2[,"MI"],apply(table2by2,1,sum))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  table2by2[, "MI"] out of apply(table2by2, 1, sum)
## X-squared = 23.782, df = 1, p-value = 1.079e-06
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.010570828 -0.004440228
## sample estimates:
##      prop 1      prop 2
## 0.009334889 0.016840417
```

Hide

```
summary(glm(mi~treatment,data=long,family="binomial"))
```

```
##
## Call:
## glm(formula = mi ~ treatment, family = "binomial", data = long)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0574   0.1370   0.1370   0.1843   0.1843
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       4.66462    0.09852  47.348  < 2e-16 ***
## treatmentPlacebo -0.59763    0.12283  -4.865 1.14e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3122.5  on 22363  degrees of freedom
## Residual deviance: 3097.8  on 22362  degrees of freedom
## AIC: 3101.8
##
## Number of Fisher Scoring iterations: 7
```

Section 2: Poisson regression

The dataset *students.csv* shows the number of high school students diagnosed with an infectious disease for each day from the initial disease outbreak.

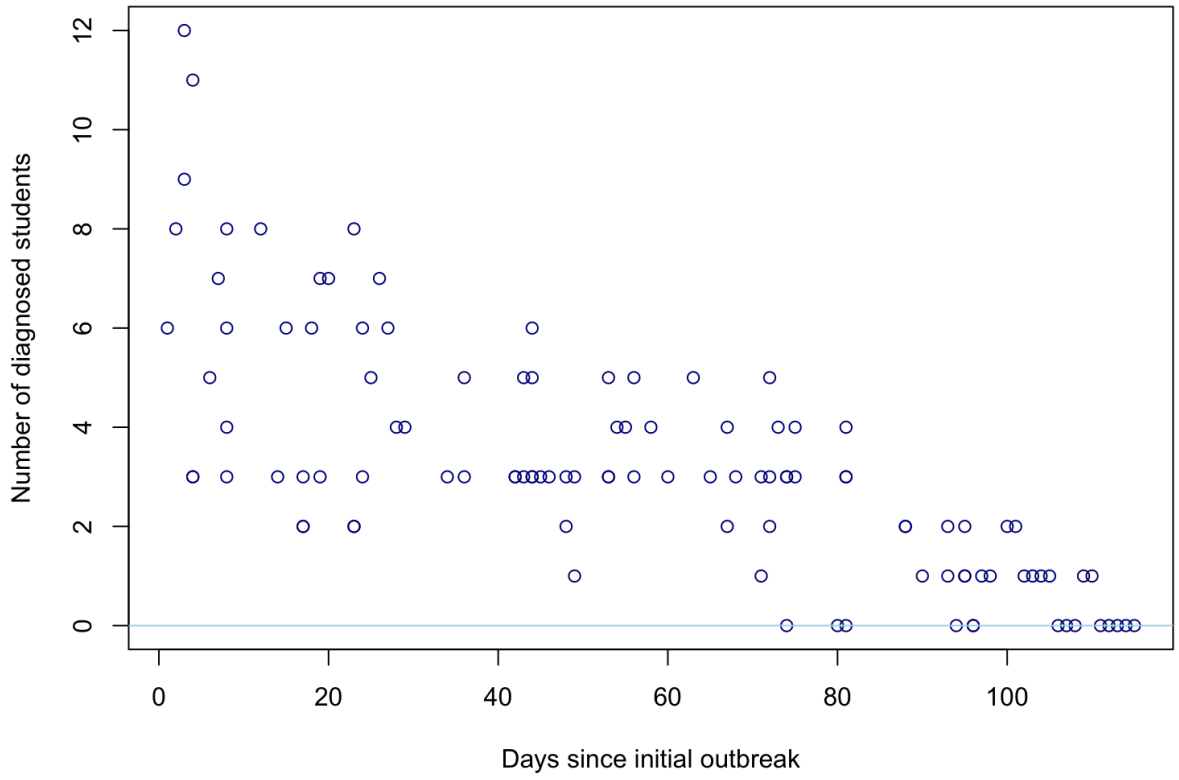
Section 2.2: Importation

Lets

- import the dataset by means of the function `read.csv()`
- display the daily number of students diagnosed with the disease (variable `cases`) as a function of the days since the outbreak (variable `day`).

Hide

```
students = read.csv("data/students.csv",header=TRUE)
plot(students$day,students$cases,col="blue4",
       ylab = "Number of diagnosed students", xlab = "Days since initial outbreak"
        )
abline(h=c(0),col="light blue")
```



Section 2.2: Model fit

Lets

- fit a poisson model by means the function `glm()` and by means of the function `gamlss()` of the library `gamlss`.
- display and analyse the results of the `glm` function : Use the function `summary()` to display the results of an R object of class `glm`.

Hide

```
fit.glm = glm(cases~day,data=students,family=poisson)

library(gamlss)
fit.gamlss = gamlss(cases~day,data=students,,family=PO)
```

```
## GAMLSS-RS iteration 1: Global Deviance = 389.1082
## GAMLSS-RS iteration 2: Global Deviance = 389.1082
```

Hide

```
summary(fit.glm)
```

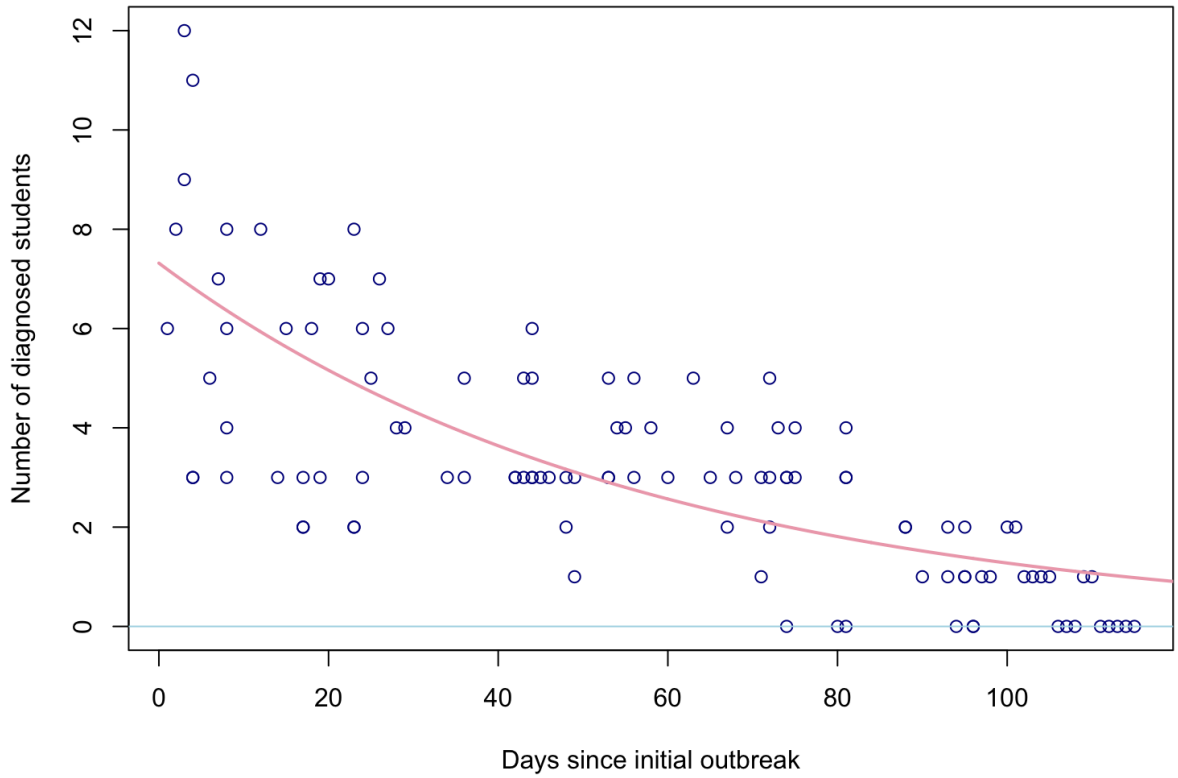
```
##
## Call:
## glm(formula = cases ~ day, family = poisson, data = students)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.00482  -0.85719  -0.09331   0.63969   1.73696
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.990235   0.083935   23.71   <2e-16 ***
## day         -0.017463   0.001727  -10.11   <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: 215.36  on 108  degrees of freedom
## Residual deviance: 101.17  on 107  degrees of freedom
## AIC: 393.11
##
## Number of Fisher Scoring iterations: 5
```

Let's now define the estimated probability of having bronchitis for any number of daily smoked cigarette and display the corresponding logistic curve on a plot:

Hide

```
plot(students$day,students$cases,col="blue4",
        ylab = "Number of diagnosed students", xlab = "Days since initial outbreak"
         )
abline(h=c(0),col="light blue")

axe.x = seq(0,120,length=1000)
f.x = exp(fit.glm$coef[1]+axe.x*fit.glm$coef[2])
lines(axe.x,f.x,col="pink2",lwd=2)
```



Section 2.3: Model selection

As for linear models, model selection may be done by means of the function `anova()` used on the glm object of interest.

Hide

```
anova(fit.glm,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: cases
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                   108     215.36
## day   1   114.18       107     101.17 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Section 2.3: Model check

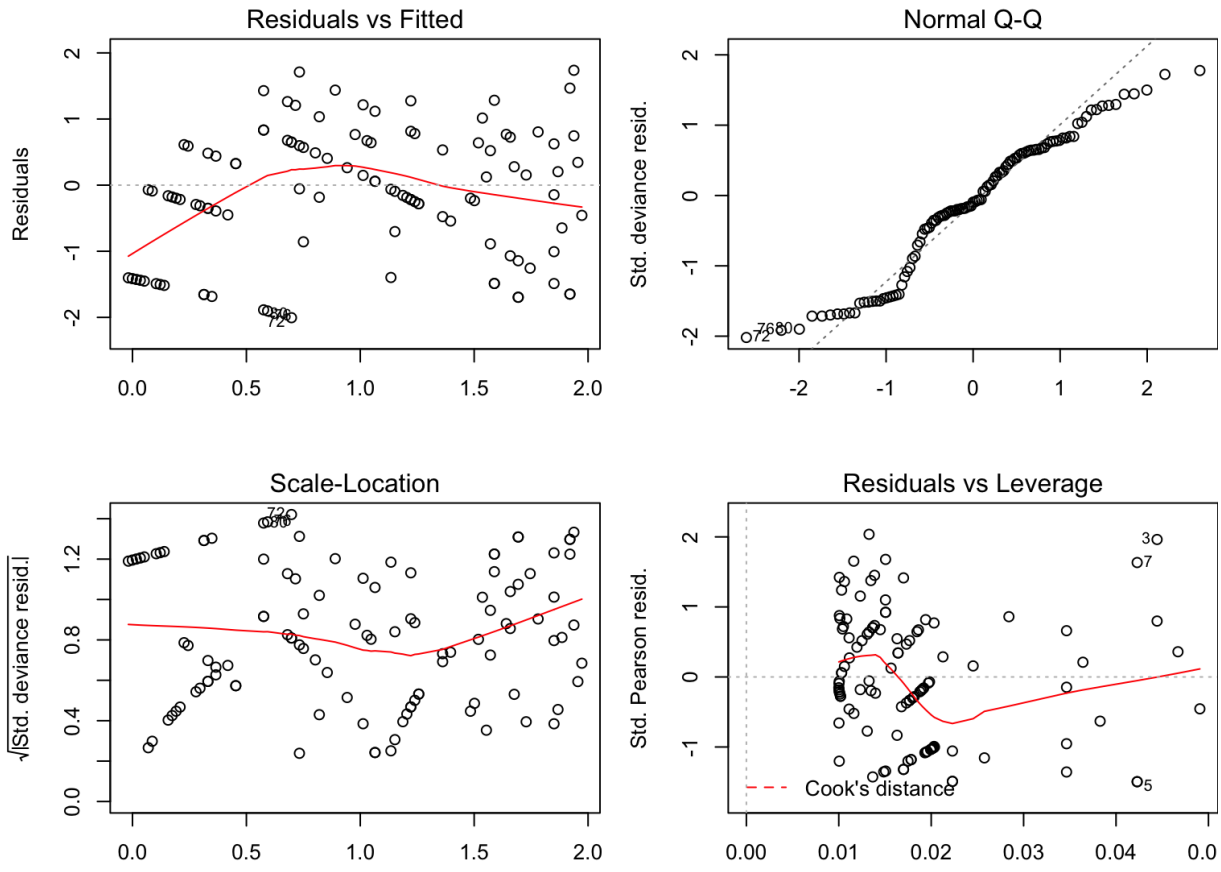
Lets assess is the model fit seems satisfactory by means

- of the analysis of deviance residuals (function `plot()` on an object of class `glm`,

- of the analysis of randomised normalised quantile residuals (function `plot()` on an object of class `gamlss`,

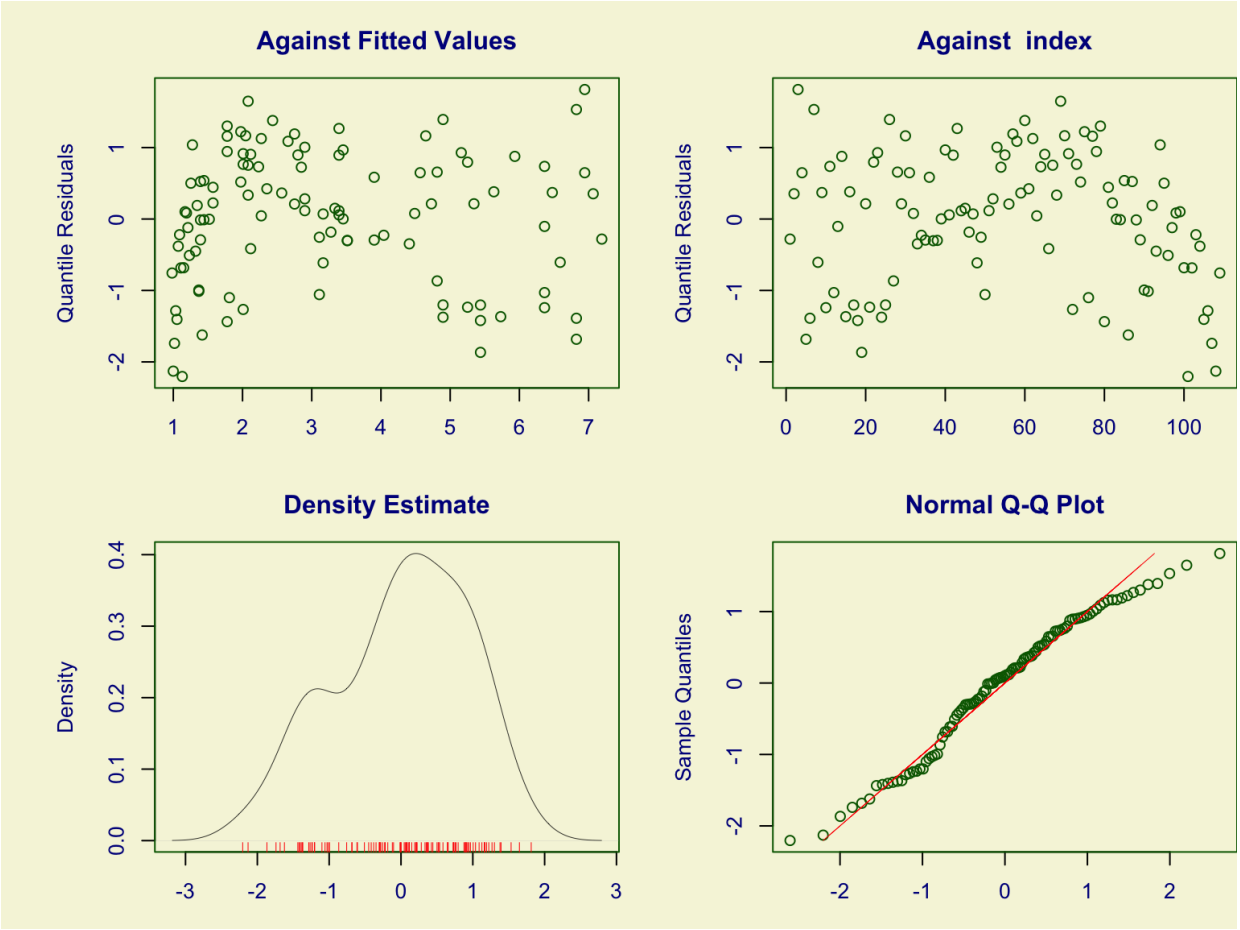
Hide

```
# deviance
par(mfrow=c(2,2),mar=c(3,5,3,0))
plot(fit.glm)
```



Hide

```
# randomised normalised quantile residuals
plot(fit.gamlss)
```




```
## ****************************************************************
##   Summary of the Randomised Quantile Residuals
##                            mean   =  0.005443867
##                        variance   =  0.8657058
##                coef. of skewness  =  -0.3582784
##                coef. of kurtosis  =  2.278417
## Filliben correlation coefficient  =  0.9871338
## ****************************************************************
```

Section 6: Practicals

(i) *Bronchitis.csv*

Analyse further the Bronchitis data of Jones (1975) by

- first investigating if the probability of having bronchitis also depends on *pollution* (variable `poll`),
- second investigating if there is an interaction between the variables `cigs` and `poll`.

(ii) *myocardialinfarction.csv*

The file *myocardialinfarction.csv* indicates if a participant had a myocardial infarction attack (variable `infarction`) as well the participant's treatment (variable `treatment`).

Does *Aspirin* decrease the probability to have a myocardial infarction attack ?

(ii) *crabs.csv*

This data set is derived from Agresti (2007, Table 3.2, pp.76-77). It gives 6 variables for each of 173 female horseshoe crabs:

- Explanatory variables that are thought to affect this included the female crab's color (C), spine condition (S), weightweight (Wt)
- C: the crab's colour,
- S: the crab's spine condition,
- Wt: the crab's weight,
- W: the crab's carapace width,
- Sa: the response outcome, i.e., the number of satellites.

Check if the width of female's back can explain the number of satellites attached by fitting a Poisson regression model with width.