lab2-block2_group5_report

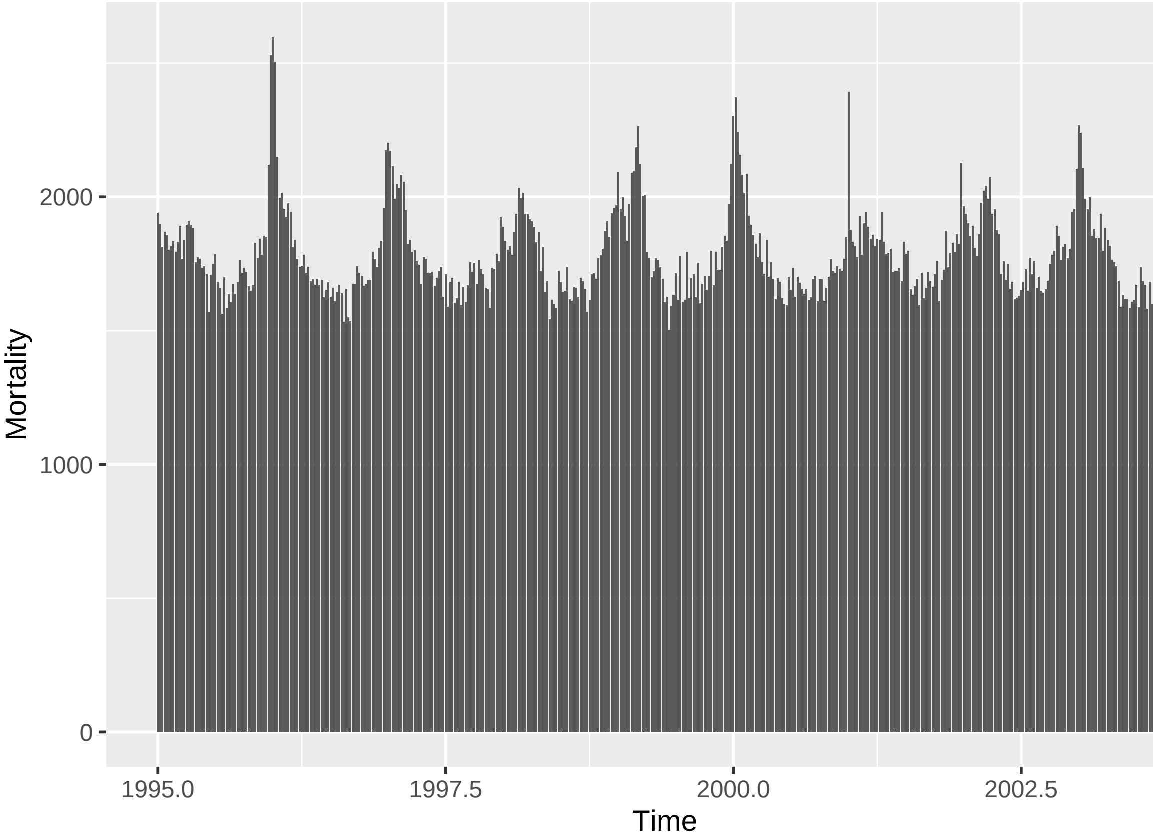
Bjorn_Hansen, Erik Anders, Ahmed Alhasan

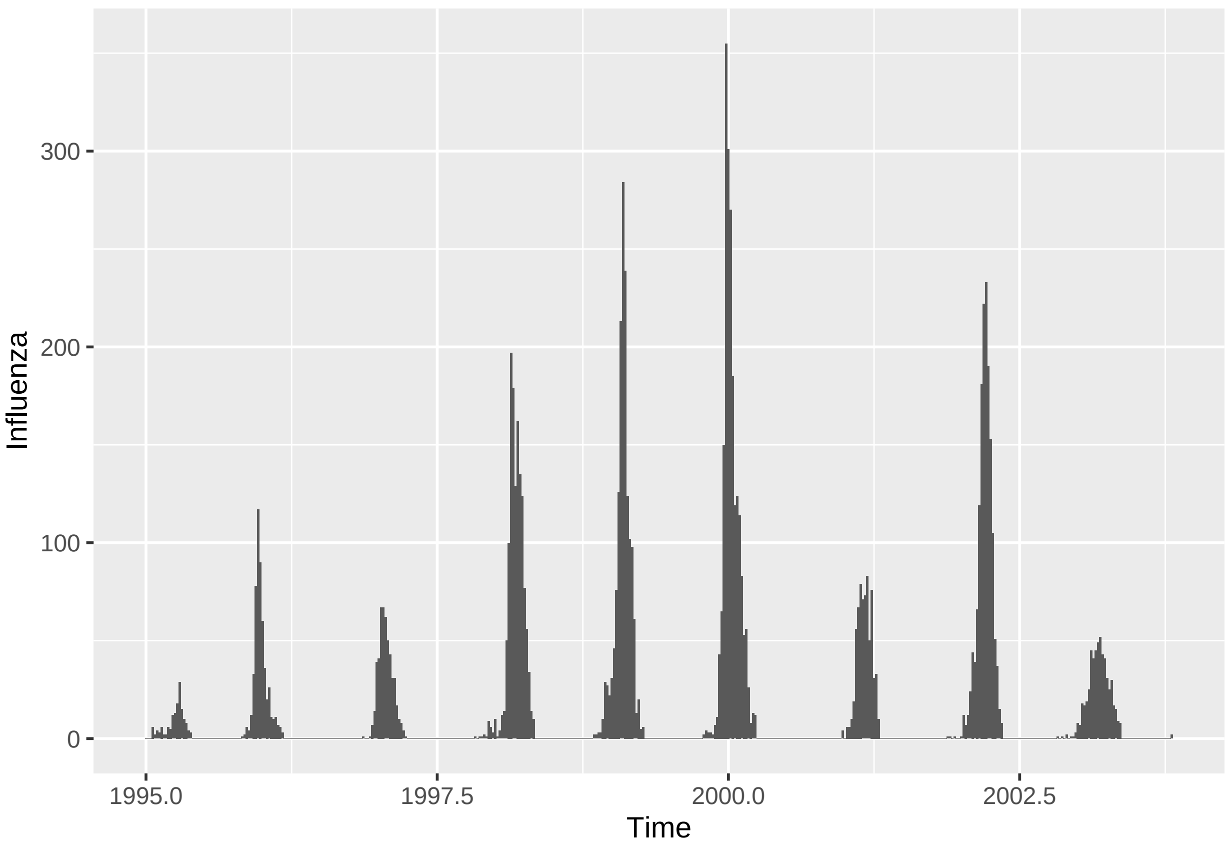
12/16/2019

Assignment 1. Using GAM and GLM to examine the mortality rates

1

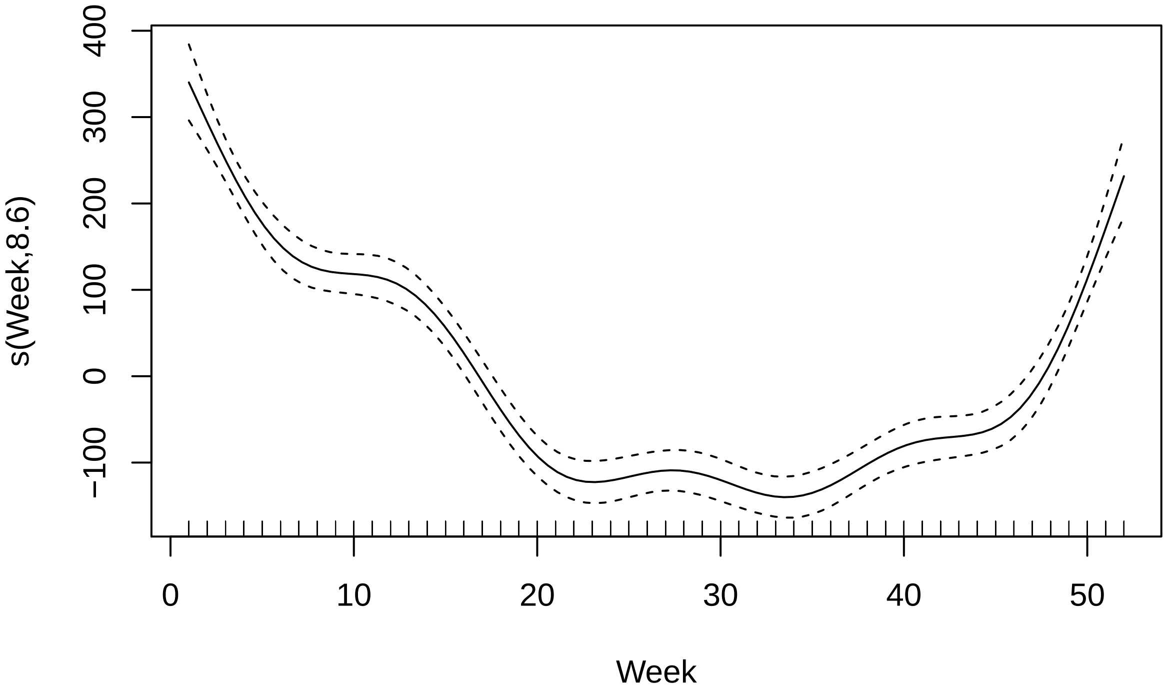
As can be seen from the plots below there does not seem to be a clear relationship between influenza and mortality by viewing the plots. There are perhaps peaks of influenza when there are relative peaks of mortality during the year of 2000.

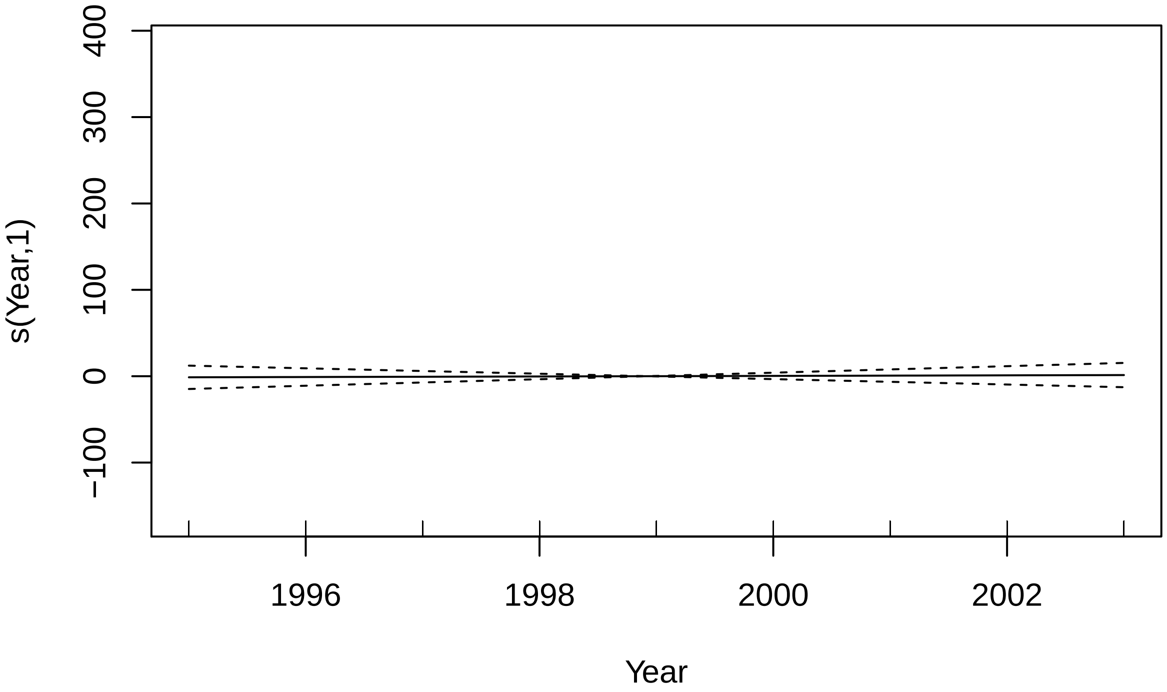




2

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ Year + s(Week, k = length(unique(data$Week)))
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -680.598   3367.760  -0.202    0.840
## Year           1.233      1.685   0.732    0.465
##
## Approximate significance of smooth terms:
##           edf Ref.df     F p-value
## s(Week) 14.32  17.87 53.86  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 52/53
## R-sq.(adj) =  0.677   Deviance explained = 68.8%
## GCV = 8708.6  Scale est. = 8398.9    n = 459
```

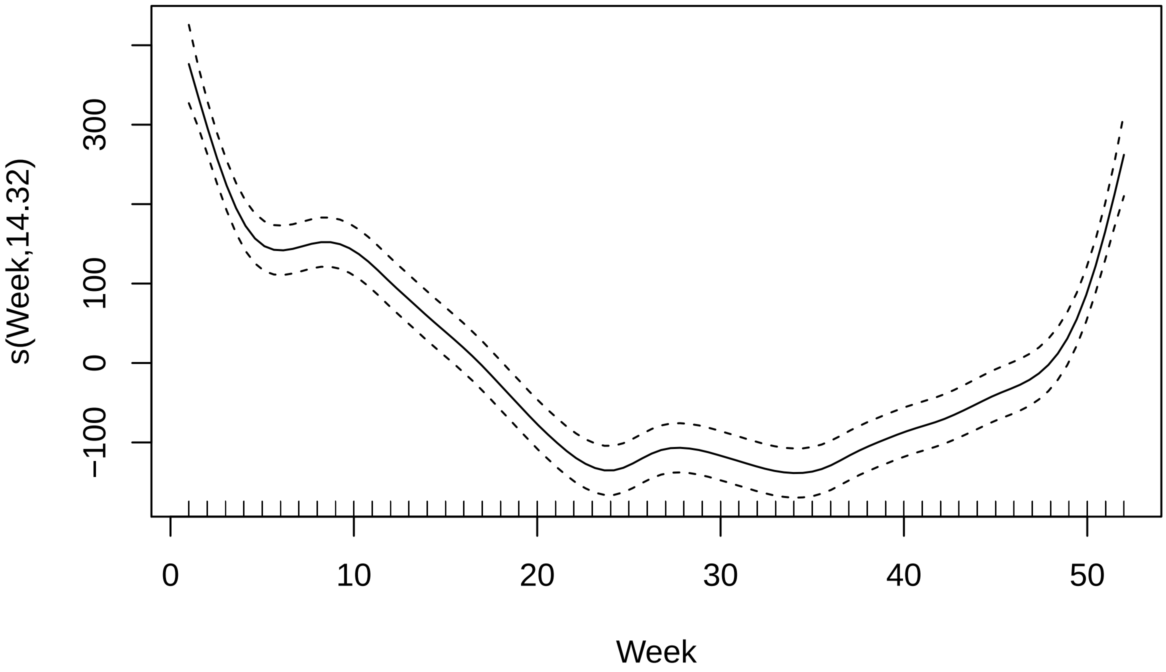

s(Week,8.6)
400
300
200
100
0
−100
0
10
20
30
40
50
Week

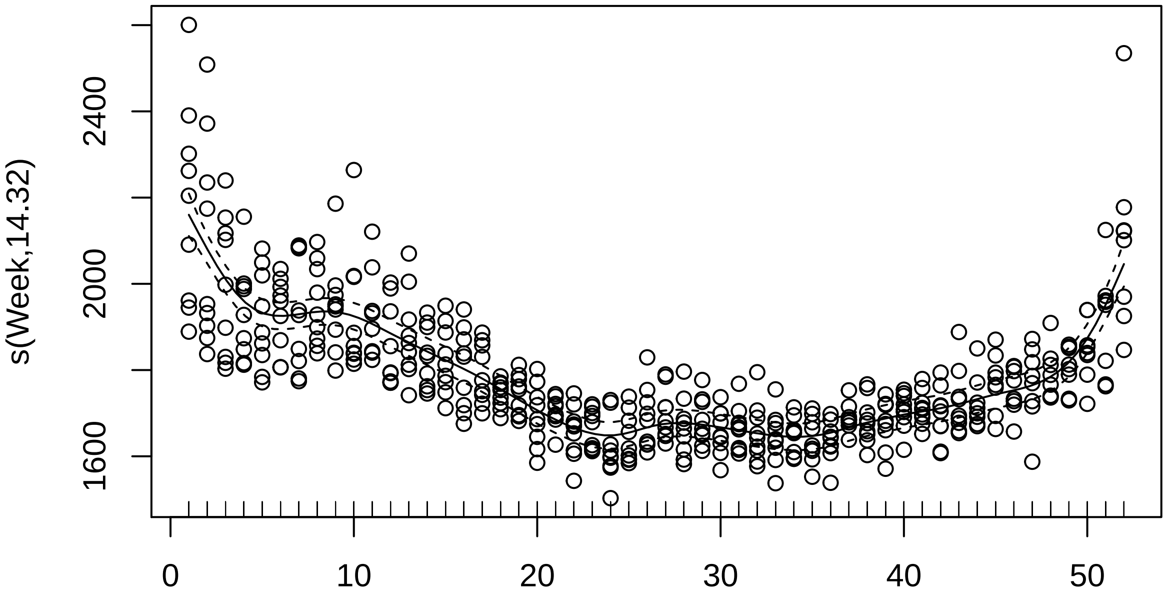


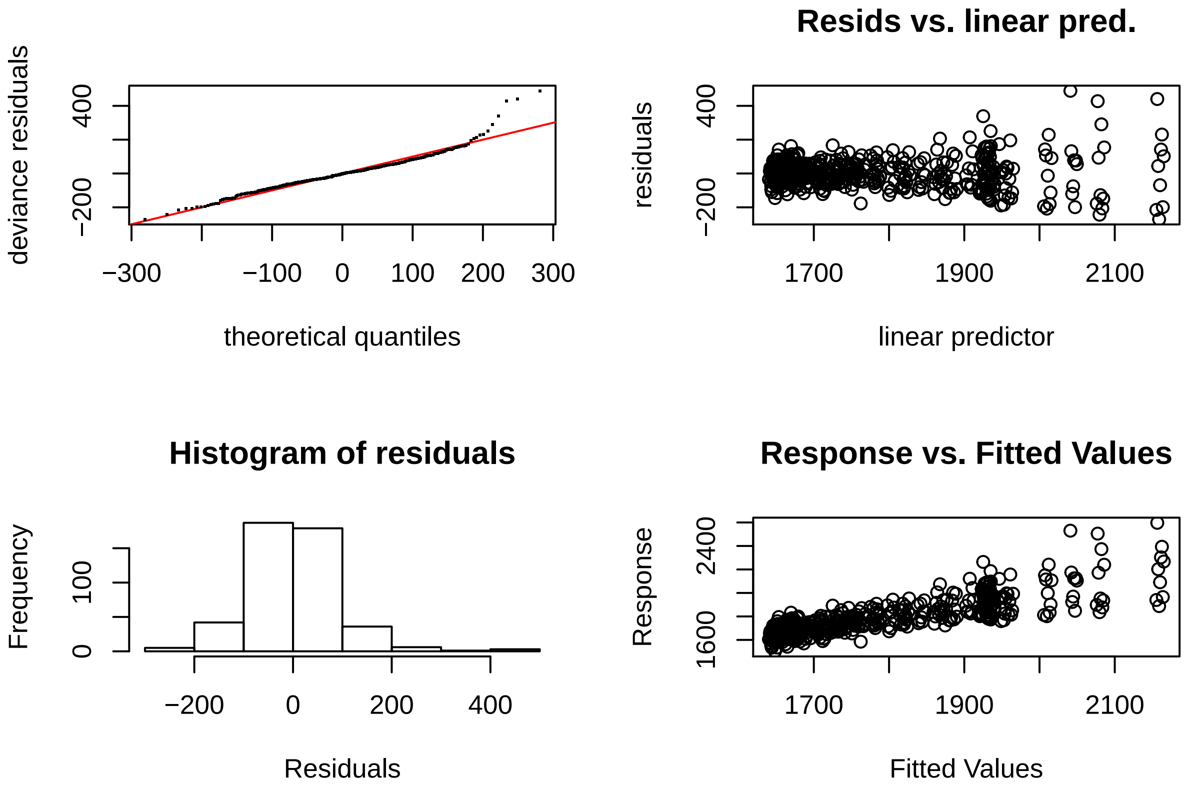
3

As is seen in the plots below, the histogram of the residuals seem to be normaly distributed and the model seems fit the data well. The week of the year is the most significant feature. The year is not of importance as mortality does not change much between years. It can also be seen from the previous summary of the model that the year has a large p-value.

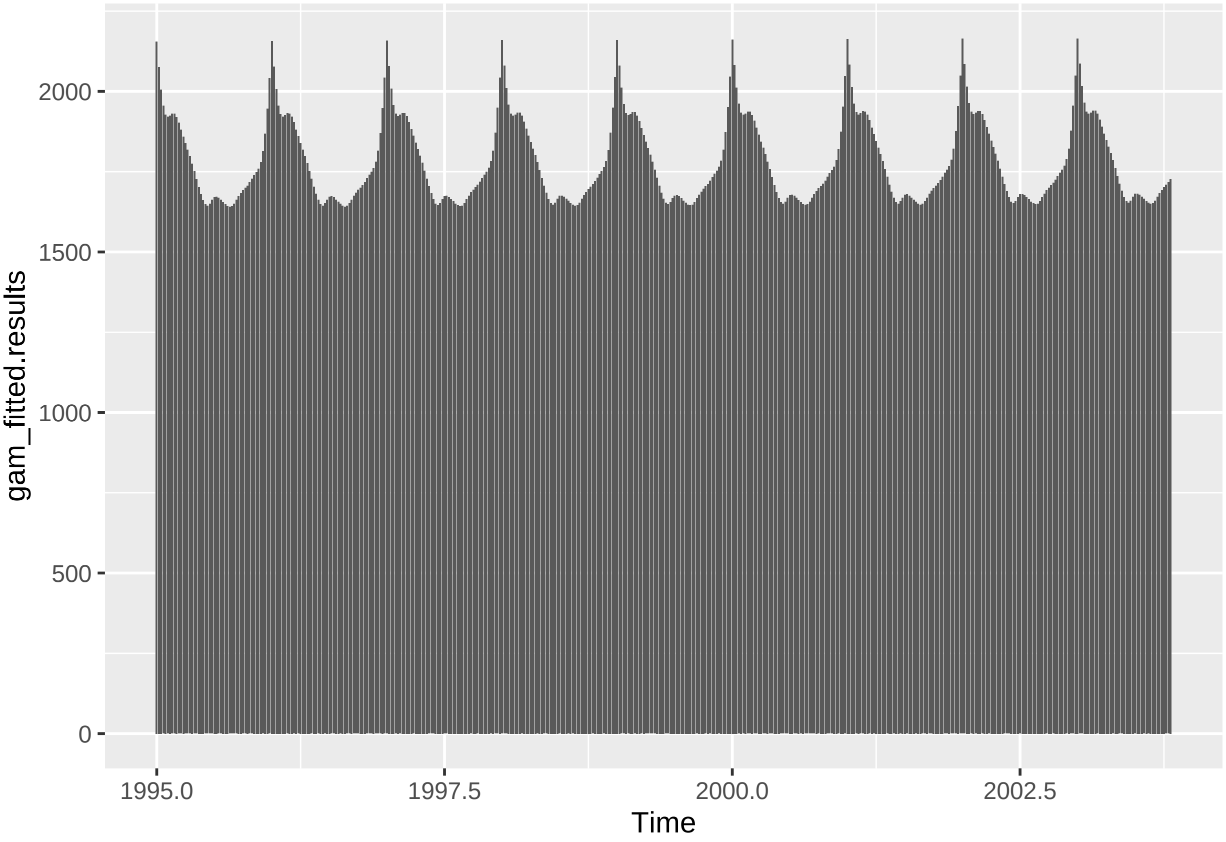
```
##      s(Week)
## 0.0001131932
```


s(Week,14.32)
300
100
0
−100
0
10
20
30
40
50
Week


s(Week,14.32)
1600
2000
2400
0
10
20
30
40
50

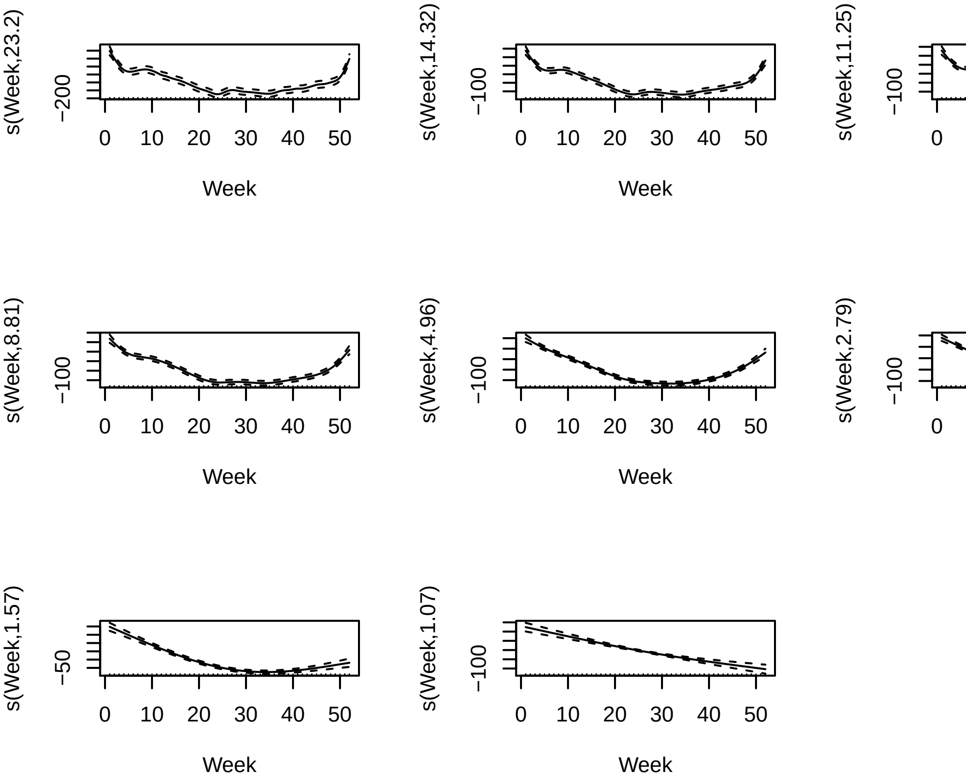


```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 9 iterations by steepest
## descent step failure.
## The RMS GCV score gradient at convergence was 0.00106719 .
## The Hessian was positive definite.
## Model rank =  52 / 53
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##            k'  edf k-index p-value
## s(Week) 51.0 14.3    1.09    0.97
```



4

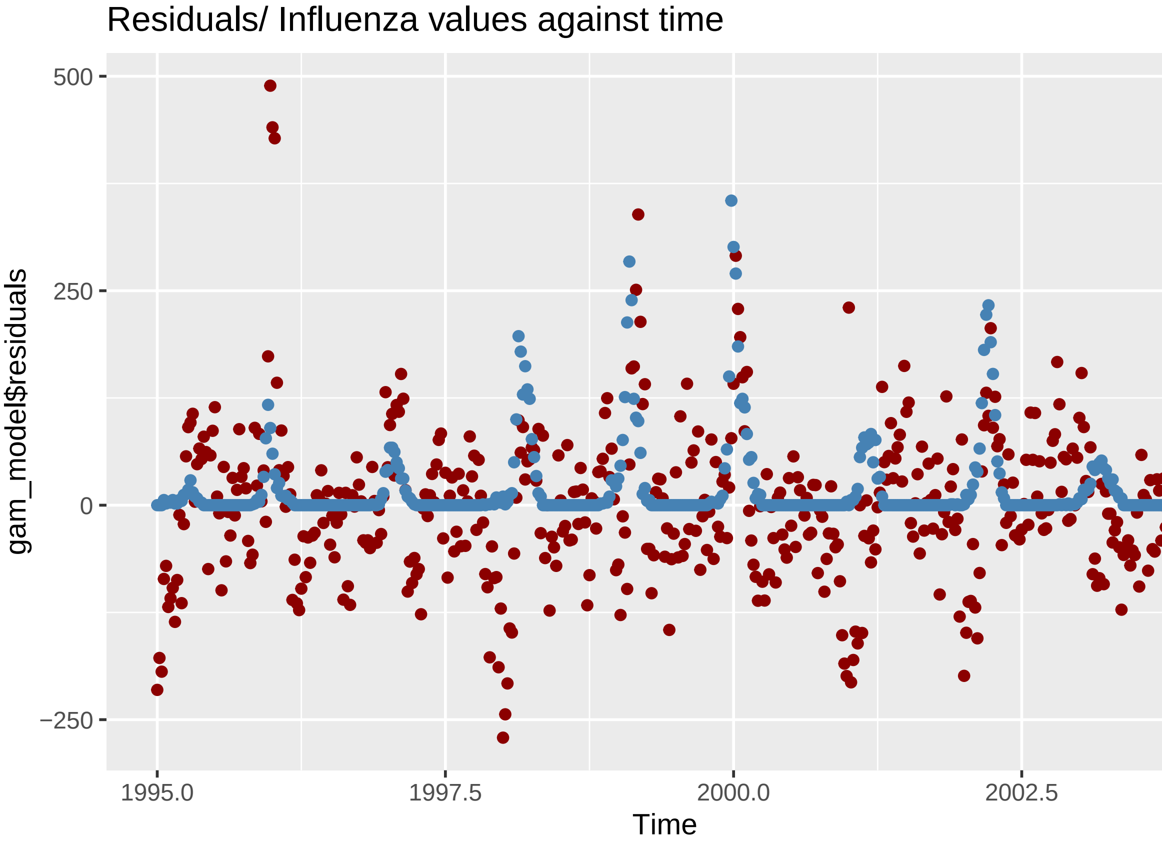
As can be seen from the output below, the lower penatly factor makes the model fit the data more loosly whereas a larger penalty factor gives the model a exact fit. The fit of the model does however not change much



after a penalty factor of 13 is introduced.

5

Viewing the plot bleow it would seem viable to say that the temporal pattern in the residuals correlate to the outbreaks of influenza.

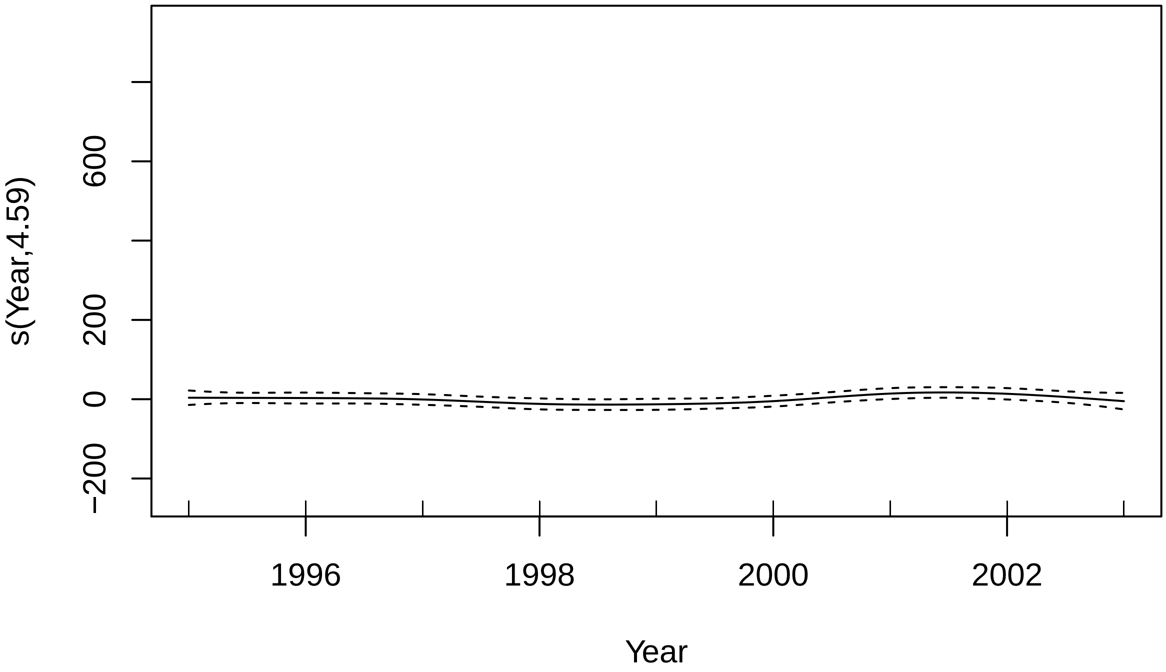


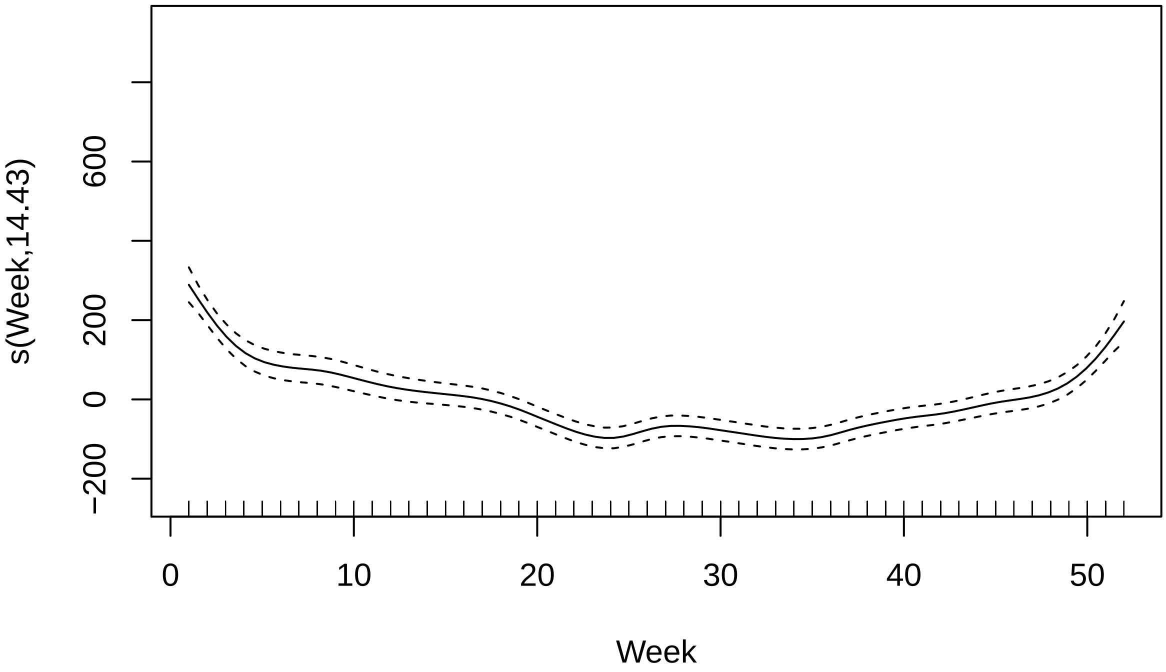
6

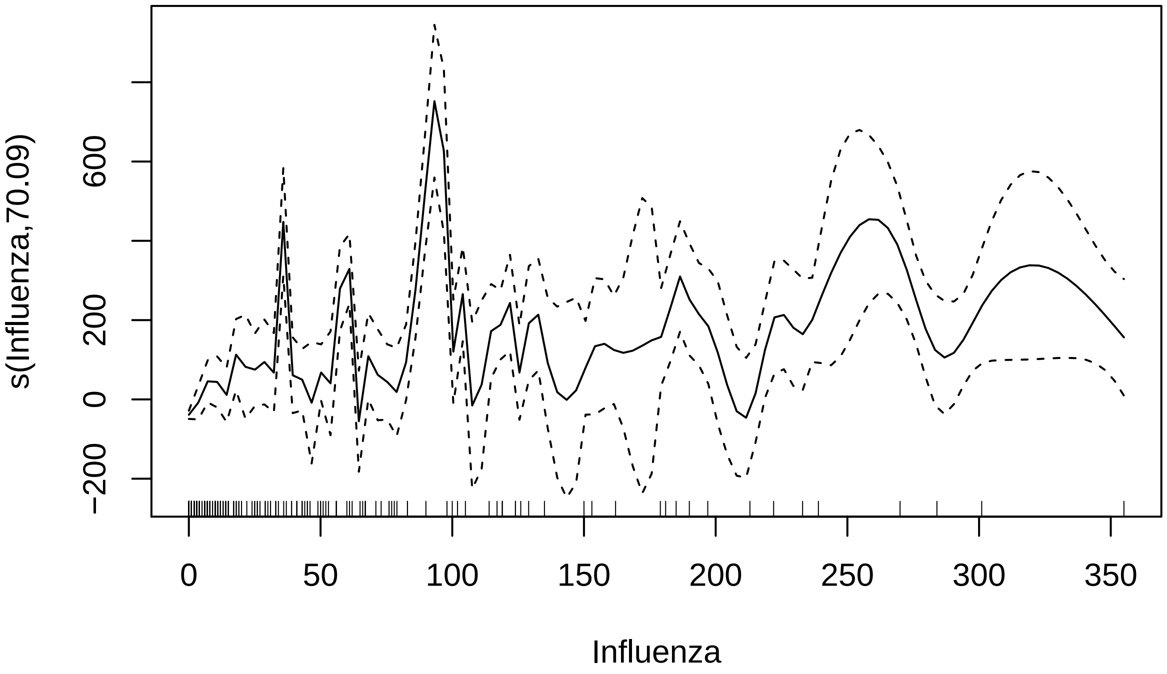
Plot shows that the fitted values are good aproximations of the mortality values. Using the summary function it is seen from the p- values that influenza and week are significant contributers of the mortality rate. The year is however not a significant contributer.

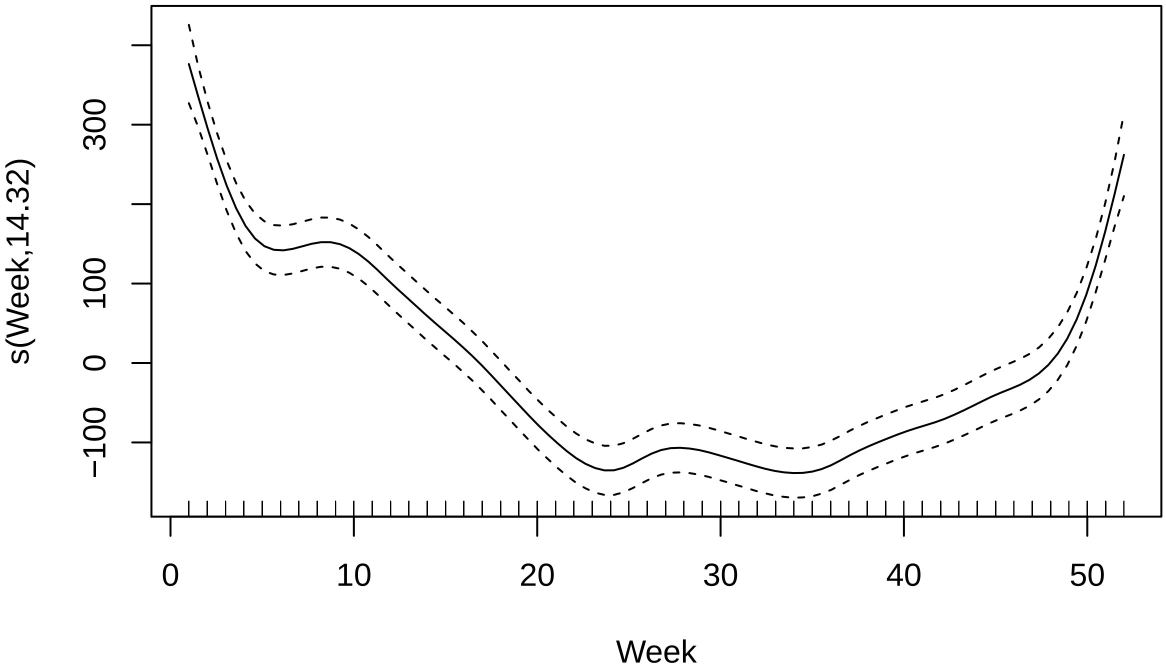
The first plot below is from the original (previous) model and the second from the spline. The spline model is a better model than the previous GAM model. The last output plot shows that the fitted values match the data very well.

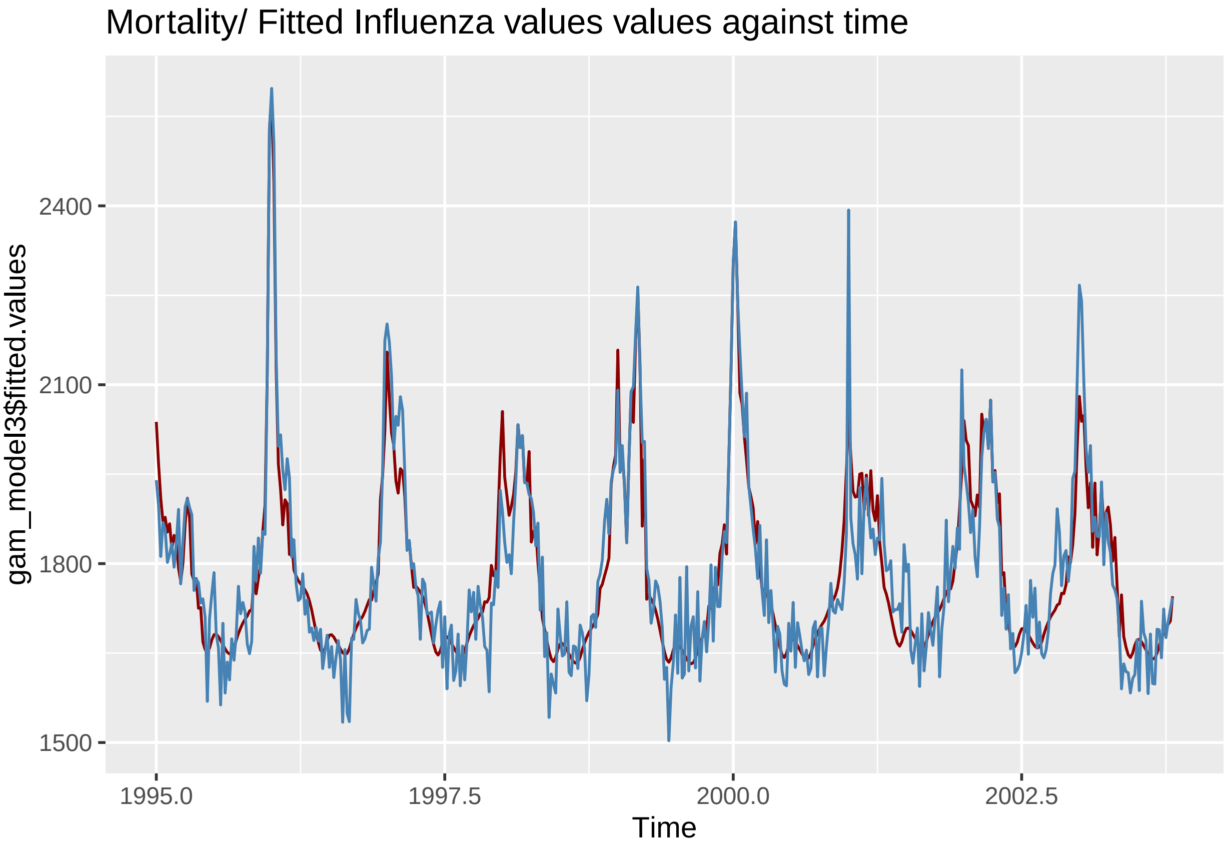
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ s(Year, k = length(unique(data$Year))) + s(Week,
##     k = length(unique(data$Week))) + s(Influenza, k = length(unique(data$Influenza)))
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1783.765      3.198   557.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                  edf Ref.df      F p-value
## s(Year)        4.587  5.592  1.500   0.178
## s(Week)       14.431 17.990 18.763  <2e-16 ***
## s(Influenza) 70.094 72.998  5.622  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 134/144
## R-sq.(adj) =  0.819   Deviance explained = 85.4%
## GCV = 5840.5  Scale est. = 4693.7    n = 459
```




s(Week,14.43)
−200
0
200
600
0
10
20
30
40
50
Week


s(Influenza,70.09)
−200
0
200
600
0
50
100
150
200
250
300
350
Influenza


s(Week,14.32)
300
100
0
−100
0
10
20
30
40
50
Week



Assignment 2. High-dimensional methods

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(readxl)
library(ggplot2)
setwd("C:/Users/Bjorn/Documents/LIU/machine_learning/labs")
data = read_excel("Influenza.xlsx")
ggplot(data=data, aes(x=Time, y=Mortality))+
  geom_bar(stat = "identity")
ggplot(data=data, aes(x=Time, y=Influenza))+
  geom_bar(stat = "identity")
library(mgcv)
gam_model = gam(Mortality~Year+s(Week, k=length(unique(data$Week))), data=data,
                family = gaussian(link = "identity"), method="GCV.Cp")
gam_model2 = gam(Mortality~Year+s(Week)+s(Year, k=length(unique(data$Year))),
                 data=data, family = gaussian(link = "identity"))
summary(gam_model)
plot(gam_model2)
gam_model$sp
# s=interp(data$Year,data$Week, fitted(gam_model))
# plot_ly(x=~s$x , y=~s$y, z=~s$z, type="surface")
plot(gam_model)
plot(gam_model, shift=mean(data$Mortality), residuals=T, pch=1, xlab="") #plot with data points include
gam.check(gam_model) #Gives some interesting inforamtion about the model.

gam_fitted.results = predict(gam_model, newdata=data)

ggplot(data=data, aes(x=Time, y=gam_fitted.results))+
  geom_bar(stat = "identity")
par(mfrow=c(3,3))
k=c(0.000011,0.0001131932,0.0003,0.0008,0.008,0.08,0.8,10)
for(i in k){
model = gam(Mortality~Year+s(Week, k=length(unique(data$Week))), data=data,
                        family = gaussian(link = "identity"), sp=i)
mod = model[i]
plot(model)
}
ggplot(data=data, aes(x=Time))+
  geom_point(aes(y=gam_model$residuals), color="darkred")+
  geom_point(aes(y=Influenza), color="steelblue")+
  ggtitle("Residuals/ Influenza values against time")
gam_model3 = gam(Mortality~ s(Year, k=length(unique(data$Year)))+s(Week, k=length(unique(data$Week)))+
                   s(Influenza, k=length(unique(data$Influenza))), data=data, family = gaussian(link =
summary(gam_model3)
plot(gam_model3) # plot new model.
plot(gam_model) # previous model.
ggplot(data=data, aes(x=Time))+
  geom_line(aes(y=gam_model3$fitted.values), color="darkred")+
  geom_line(aes(y=Mortality), color="steelblue")+
  ggtitle("Mortality/ Fitted Influenza values values against time")
```