Lab1B2_EmilKlassonSvensson

Emil K Svensson

19 November 2016

Assignment 1

1.1 My Spline

```r
myspline <- function(Y, X, knots){

# Creates the H-matrix and names the columns.
H <- cbind(X,sapply(knots, FUN = function(k) pmax(X-k,0)) )
colnames(H) <- c("X",paste0("H",1:length(knots)))

#Creates the linear model
myLM <- lm(Y ~ H)

#Generates the predicted data
myPredictedData<-data.frame(cbind(Y,X, predict(myLM)))
colnames(myPredictedData) <- c("Y","X","Predictions")

# Plot whith ggplot
library(ggplot2)
p<-ggplot(data = myPredictedData) + geom_point(aes(x = X, y = Y)) +
  geom_point(aes(x = X, y = Predictions),color = "red")

plot(p)
return(myLM)
}
```

Pmax returns all values that are over the value specified in the vector supplied all other values the specified break point is returned. so for example in a vector with values -10:10 and 5 specified all numbers = and below 5 it will return 5 for the other values it will return their specific values.

So when the book specifies $h3(X) = (X - \xi_1)_+$ it basicly says for all values that are positive when the calculation $X - \xi_1$ is preformed should be kept as their original value all other are set to zero, therefor pmax(X - knots[1], 0) will return the correct values.
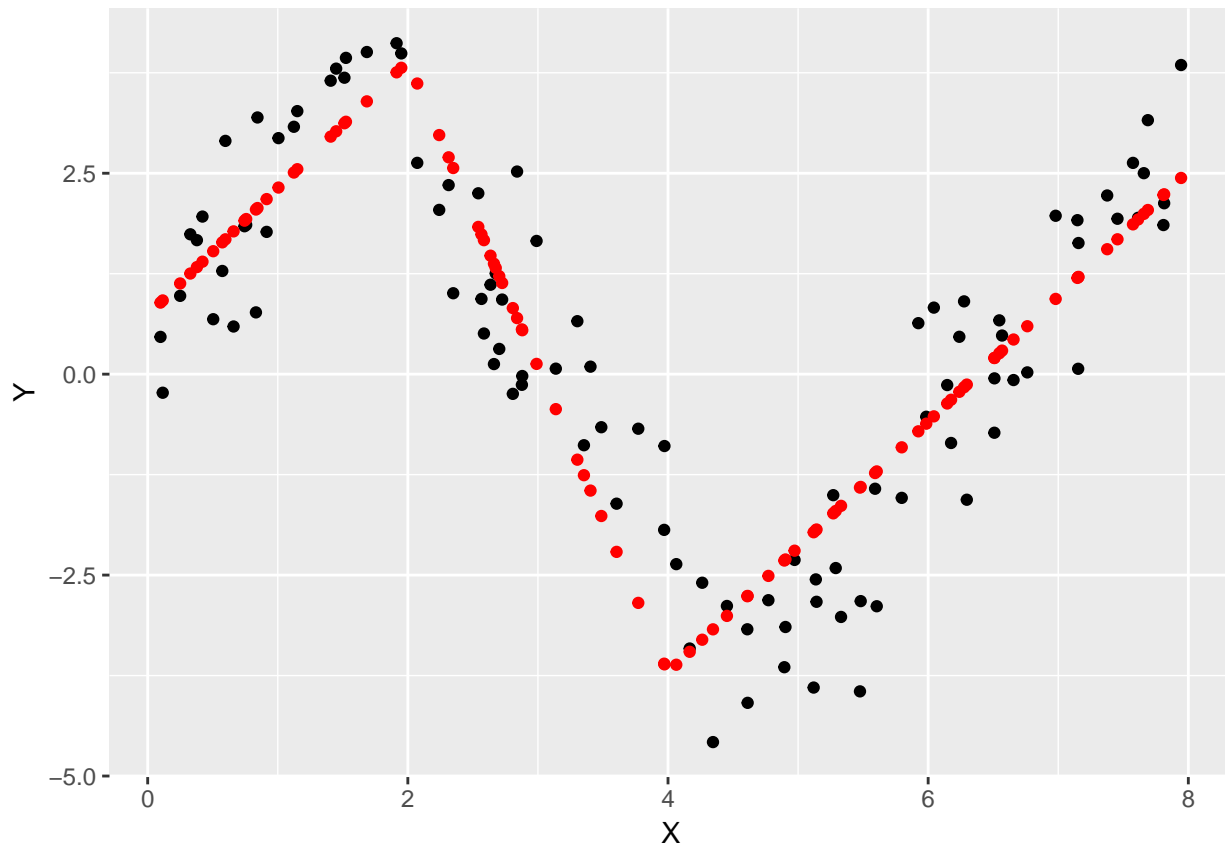
$H_1(X) = 1$ is added in the lm-function and is not neccesary to include in this case.

$H_2 = X$

$H_3 : length(knots)h3(X) = (X - \xi_i)$

1.2 Using myspline

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

The plot shows the result for the implemented spline function. The spline seems to fit the data nicely, although the second knot at 4 could be moved closer to 5. The functions are what appears to be seamless and continious in the knots. In other words, in both knots there doesn't seem to be any deviations between the different spline functions. This serves as an good indication that the function is continous in the first derviate in these knots.

```
summary(a)
```

```
##
## Call:
## lm(formula = Y ~ H)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.53539 -0.73346 -0.02351  0.67269  2.71647
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7345     0.3566   2.060   0.0421 *
## HX            1.5785     0.2657   5.942 4.51e-08 ***
## HH1          -5.3816     0.4339 -12.403  < 2e-16 ***
## HH2           5.3637     0.2934  18.281  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.003 on 96 degrees of freedom
## Multiple R-squared:  0.8014, Adjusted R-squared:  0.7952
## F-statistic: 129.1 on 3 and 96 DF,  p-value: < 2.2e-16
```

On a 5 % significance level all coefficients are significant and all separate coefficients are relevant to explaining Y. The Adjusted R-squared was calculated to 79.5 % and this along with the rest of the meassures this indicates that this is a good model for predicting Y.

1.3 Using smooth.spline()

```
smoothSpline <- smooth.spline(y = cube$y, x = cube$x)

SSpline<-data.frame(cbind(cube$y,cube$x, fitted(smoothSpline)))
colnames(SSpline) <- c("Y","X","Predictions")

ggplot(data = SSpline) + geom_point(aes(x = X, y = Y)) + geom_line(aes(x = X, y = Predictions),color =
```
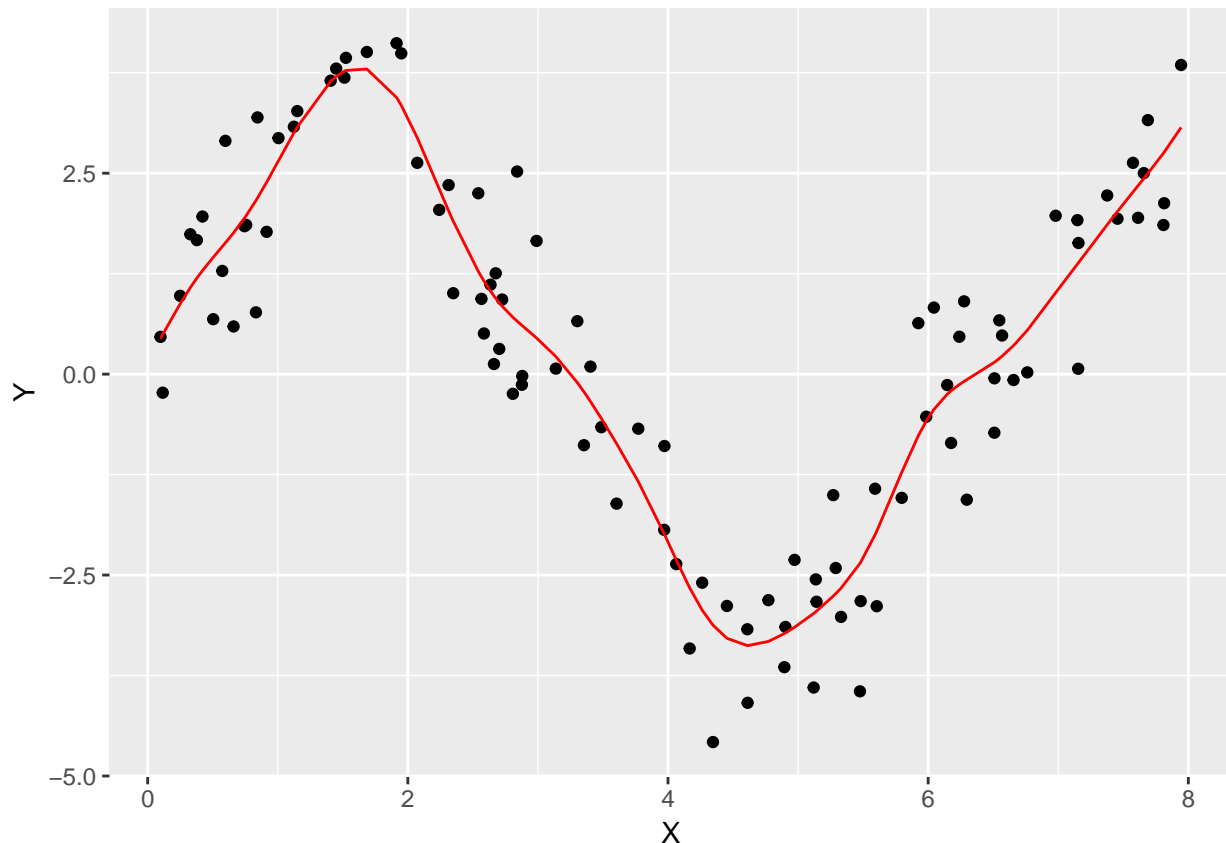

Compared to the plot in section 1.2 this spline have a similar pattern as the linear spline but us a bit wiggly and fits the data in this example better because it doesn't miss that many predictions for values around x-values around 5. In this case i would choose this model over the linear spline although I risking overfitting. This because of the linear spline the poorly choosen knot at x = 4 for the linear spline which should be around x = 5 instead.
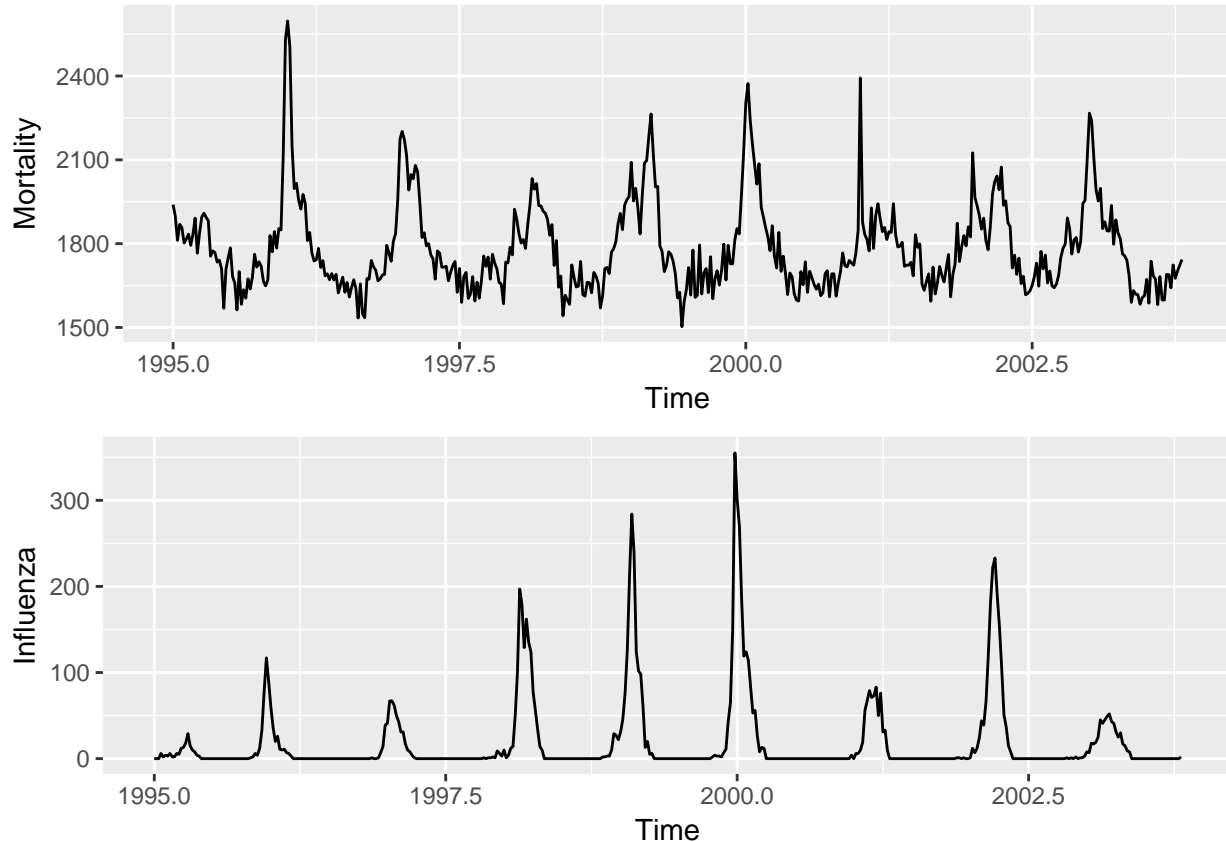
Assignment 2

2.1

```
Infu<-read.csv2("Influenza.csv")
attach(Infu)
library(gridExtra)

p<- ggplot(data = Infu, aes(x = Time))
aM<- p + geom_line(aes(y = Mortality))
aI<- p + geom_line(aes(y = Influenza))
#p + geom_line(aes(y = Mortality)) + geom_line(aes(y = Influenza))
plot(arrangeGrob(aM,aI))
```

The Mortality-rate and the number of comfirmed Influenza cases seem to have some kind of correlation although the size of the Influenza spikes doesn't explain the size of the Mortality in a good manner.

2.2

```
library(mgcv)
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.8-16. For overview type 'help("mgcv-package")'.
```

```
addM <- gam(Mortality ~ Year + s(Week),data = Infu )
```
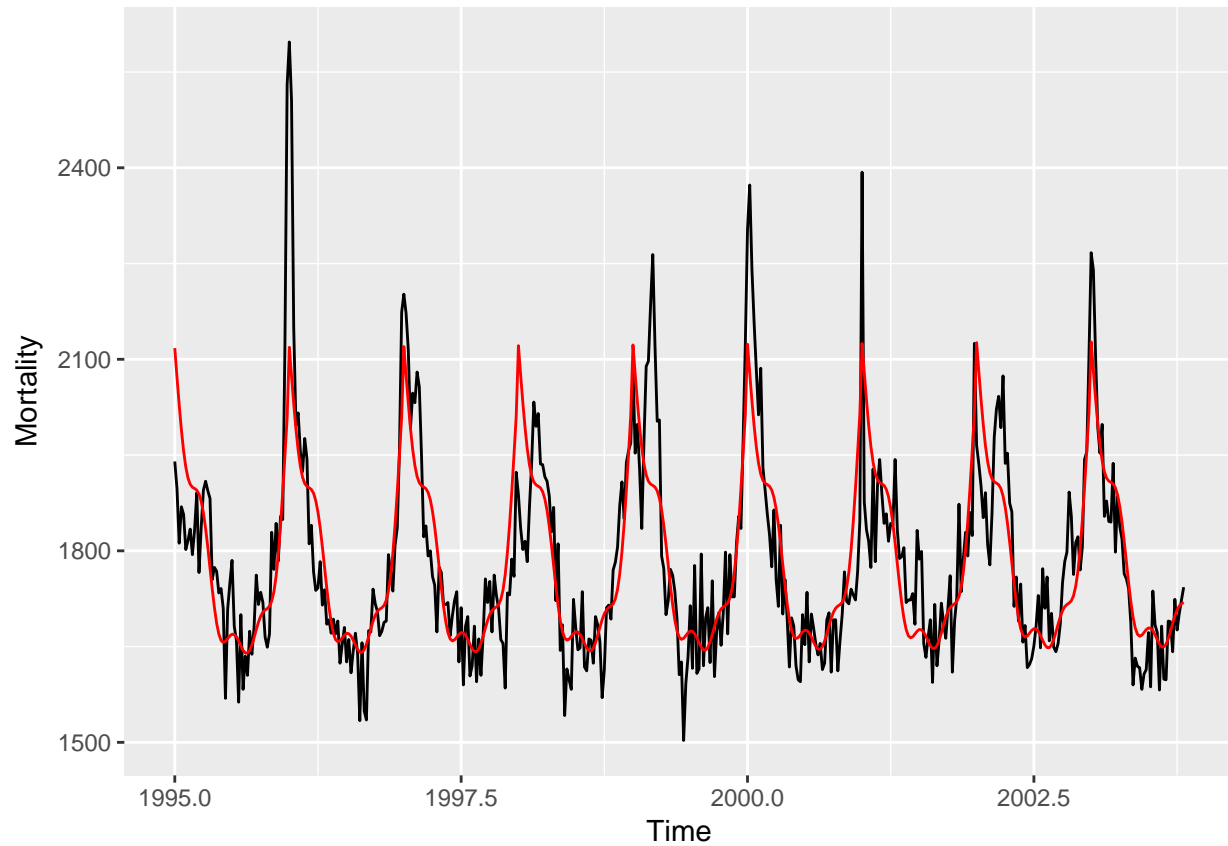
The $E(Mortality|Year, Week) = \alpha + f_1(Year) + f_2(Week)$

Where $f_1(Year) = \beta_0 + \beta_1 Year_1 + \epsilon_1$

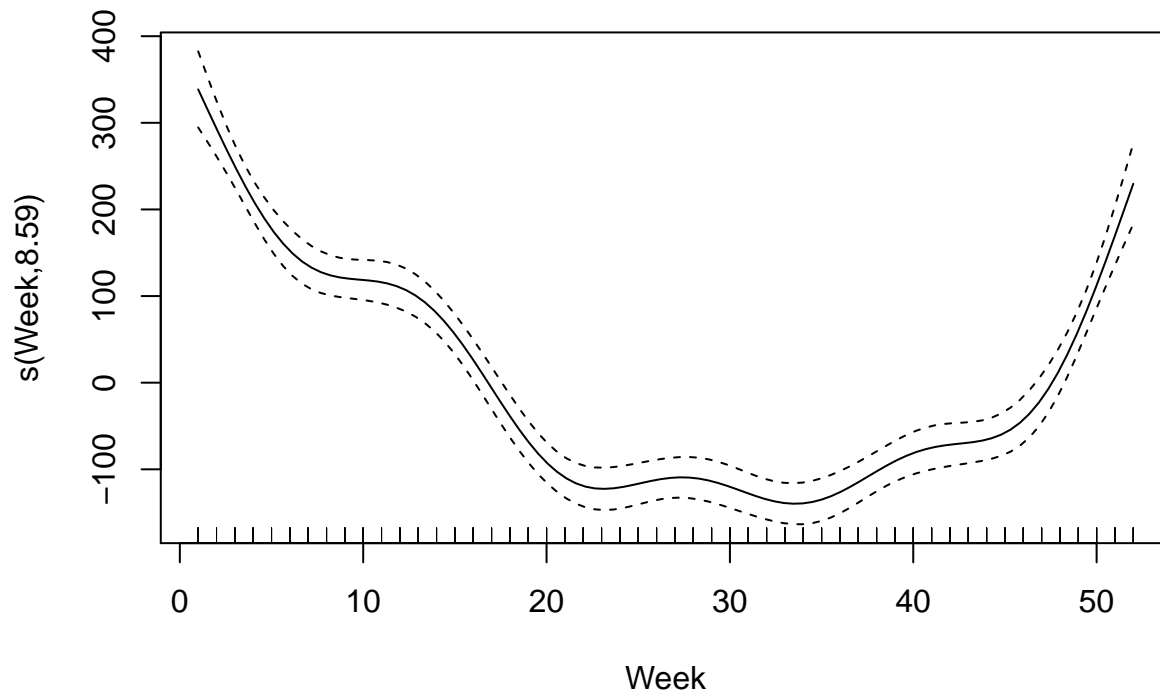and where $f_2(Week) = \sum_{m=1}^{M} \beta_m h_m(Week)$ and m are the number of knots.

2.3

```
aM + geom_line(aes(y = fitted(addM)),col = "red")
```


The fit seems to be decent and it follows the general pattern of the Mortality but it has problems following along with the higher spikes and at times where it peaks two times during a year.

```
plot(addM)
```

The spline-function plot shows how the pattern for the spline component model predicts each week. The mortality seems to peak around the first week of every year. Between week 20 and 40 there seem to be a lower mortality rate according to the spline model component. Overall it seems like the mortality rate increases during the autumn and winter months and decreases during the summer period.
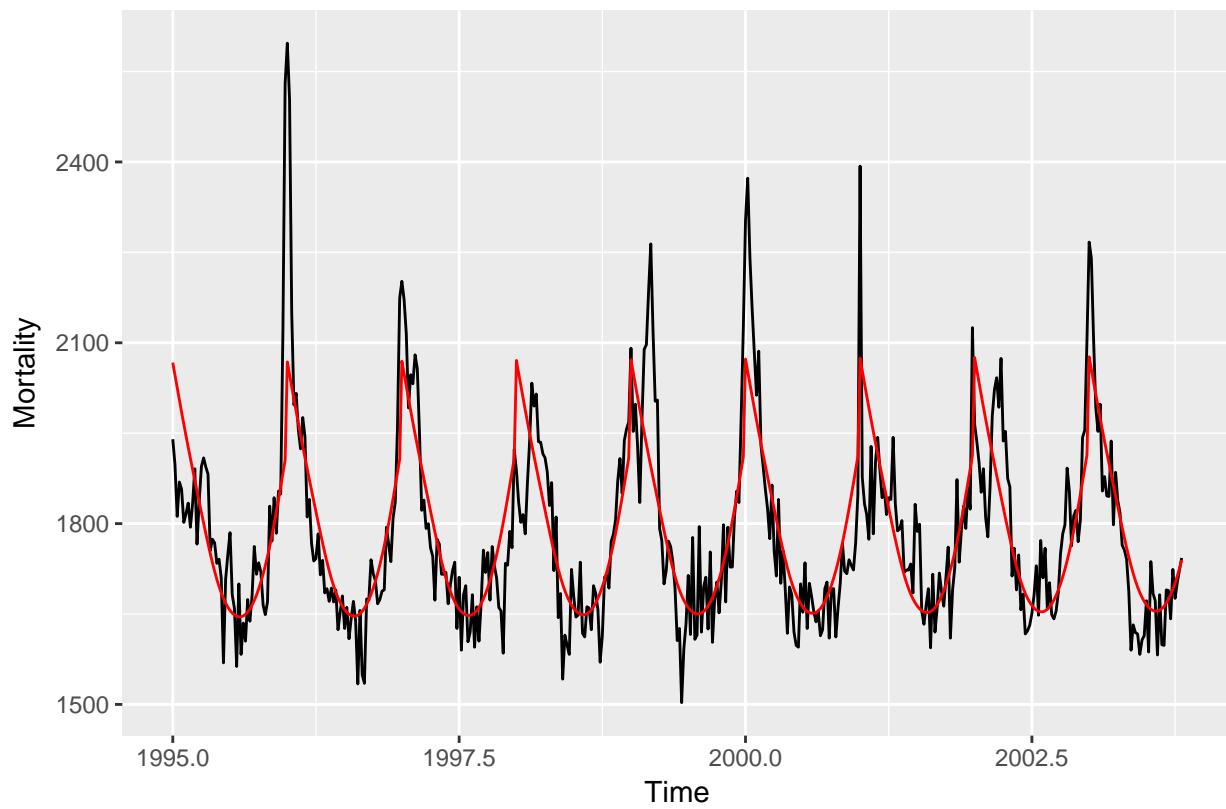
2.4

```
for (spval in c(-100,1,50,10000)){
  plot(aM + geom_line(aes(y = fitted(gam(Mortality ~ Year + s(Week, sp = spval),data = Infu))),col = "r
    labs(title = paste("Fits when lambda is:",spval) ))
}
```
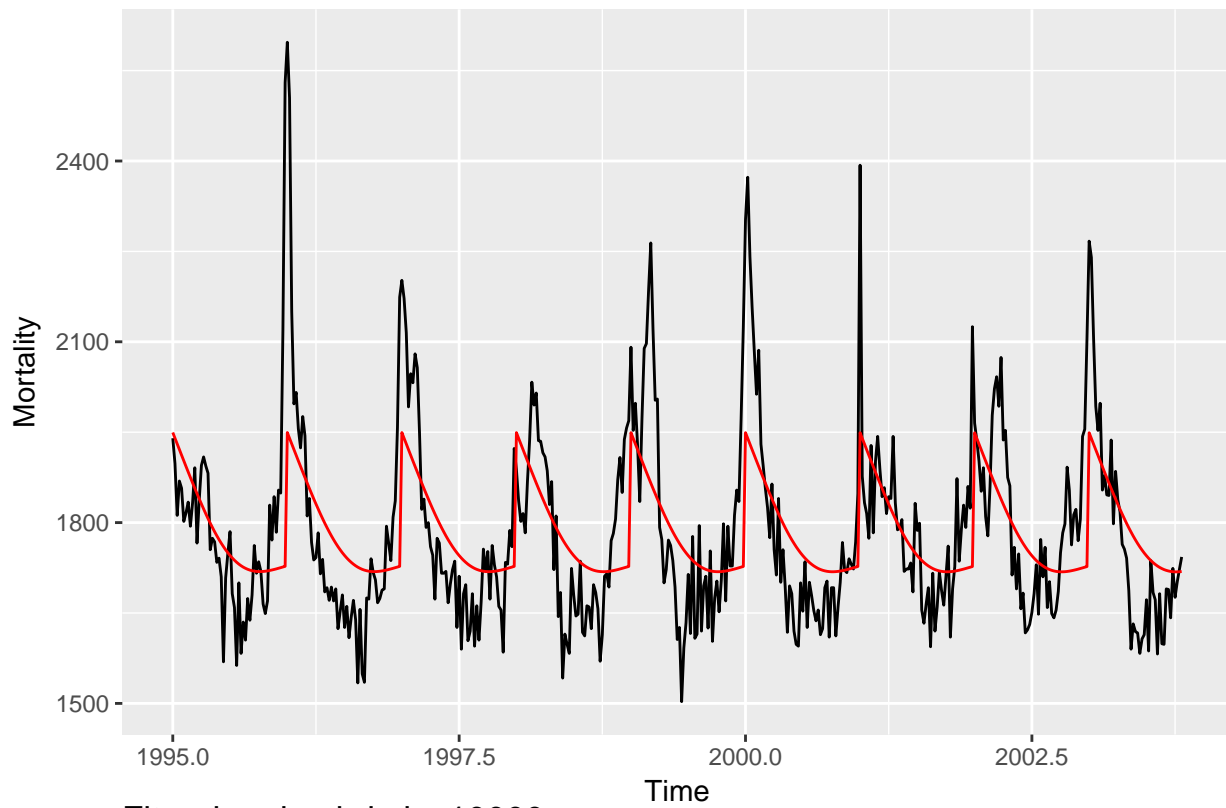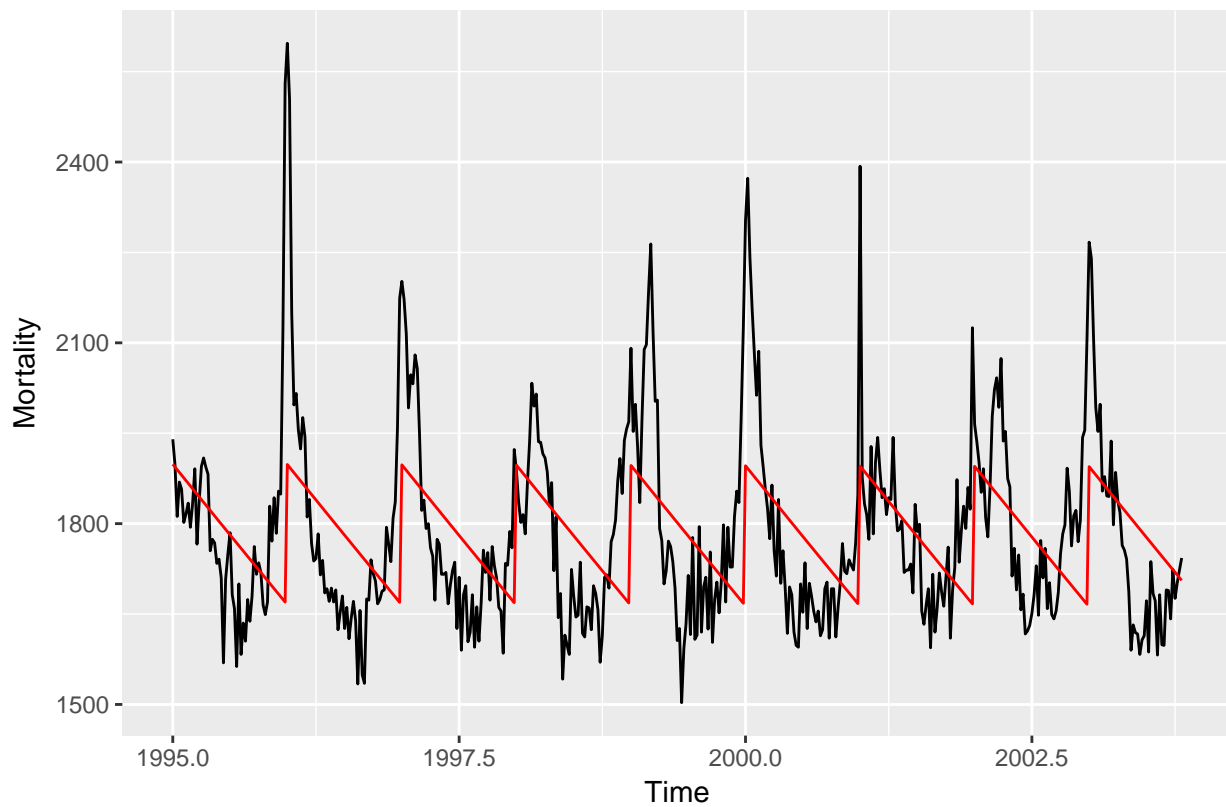
Fits when lambda is: −100
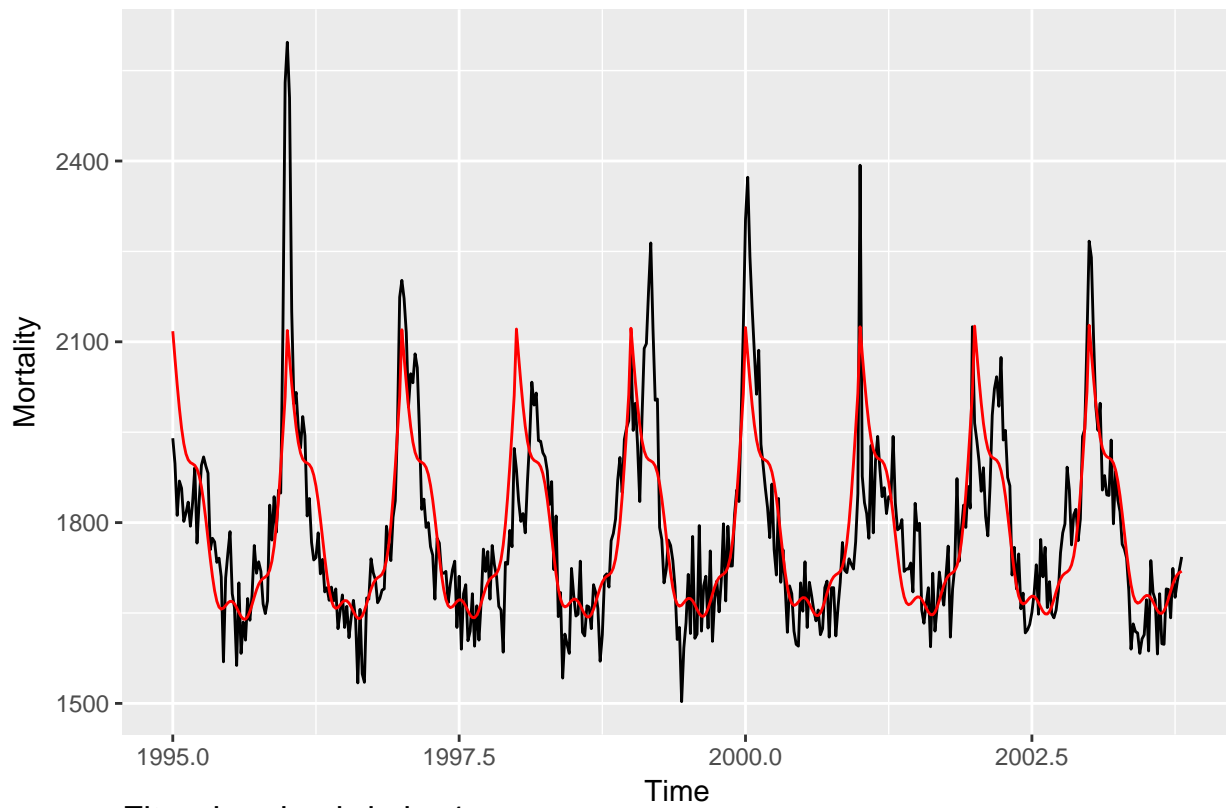
Fits when lambda is: 1

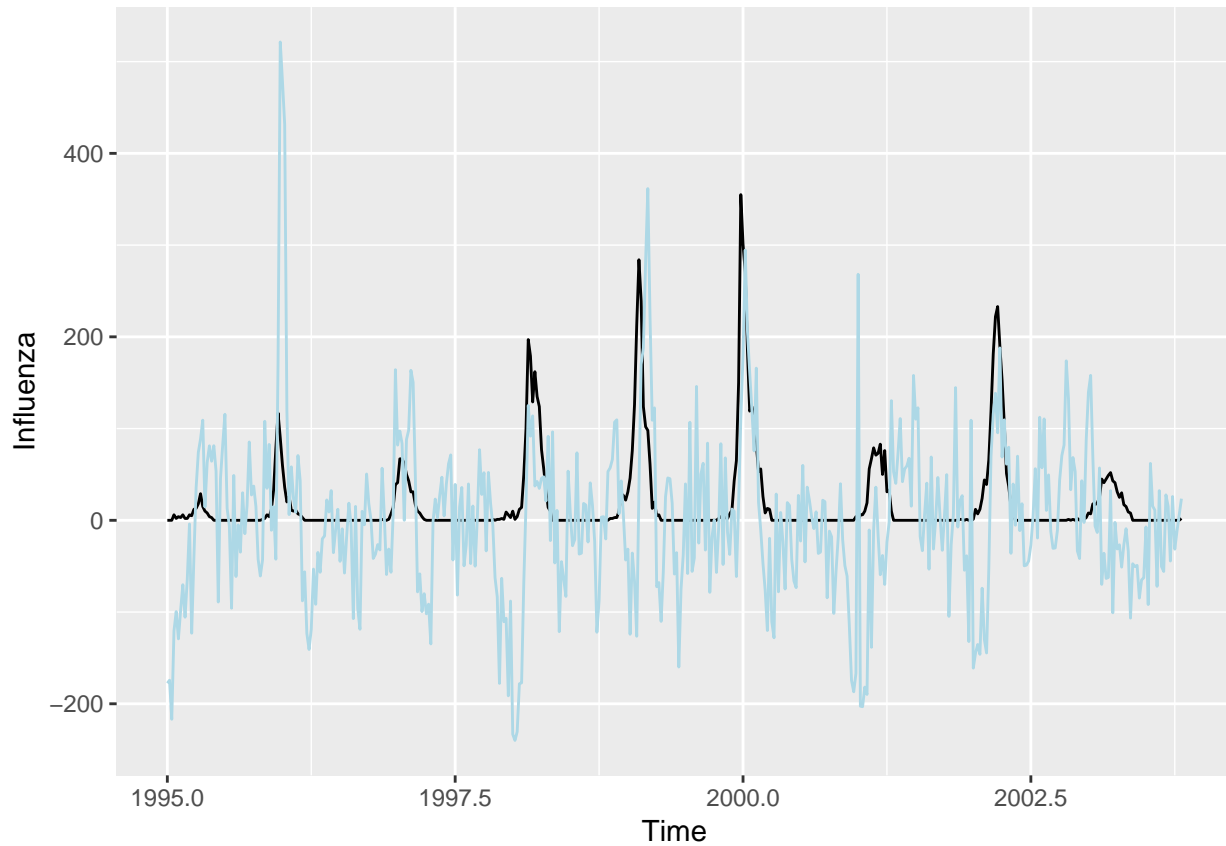## Fits when lambda is: 50

Fits when lambda is: 10000

When lambda is raised the penalty factor increases and in these four plots we can se how the adhesiveness of

the fits diminishes when the penalty factor is raised.

2.5

```r
aI + geom_line(aes(y=resid(addM),x = Time),col = "lightblue")
```


There seem to be some large residuals coincides with the peaks of the comfirmed Influenza cases. Over the whole time series the residuals doesn't seem to have some trend deviating from the y-intercept $y = 0$.

2.6

```r
add26<-gam(Mortality ~ s(Week) + s(Year, k = 9) + s(Influenza) ,data = Infu)
aM + geom_line(aes(y =fitted(add26)), col ="red")
```
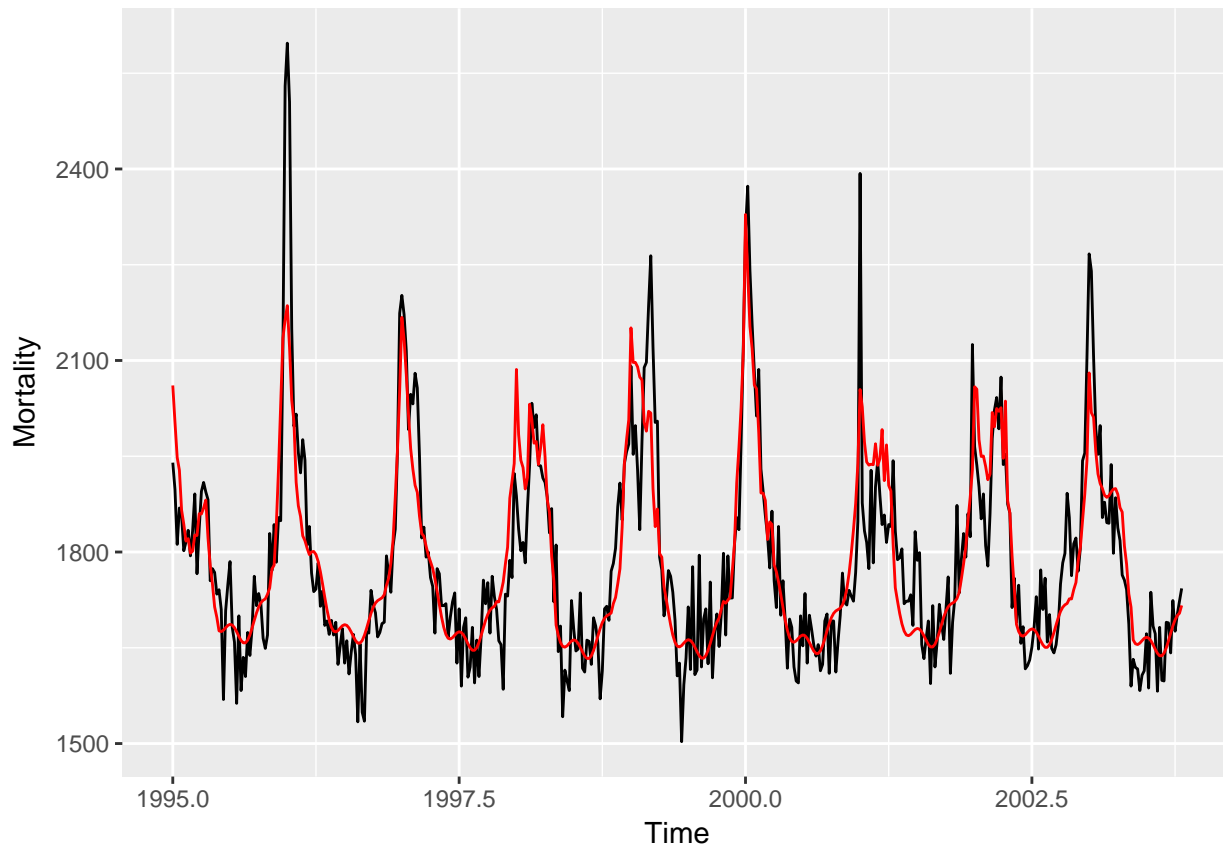
Compared to the plot in 2.3 this model seem to fit the data a bit better as it follows the spikes better.

```
summary(add26)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ s(Week) + s(Year, k = 9) + s(Influenza)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1783.765      3.813   467.8   <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(Week)      8.624  8.959 40.846  <2e-16 ***
## s(Year)      4.037  4.970  1.369    0.24
## s(Influenza) 8.175  8.802 16.703  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.743   Deviance explained = 75.5%
## GCV = 7006.3  Scale est. = 6673      n = 459
```

10

Accordin to the F-test for the smooth terms the spline function for influenza seems to contribute to the GAM-model i a significant manner as the p-value is below the standard 0.05 level.