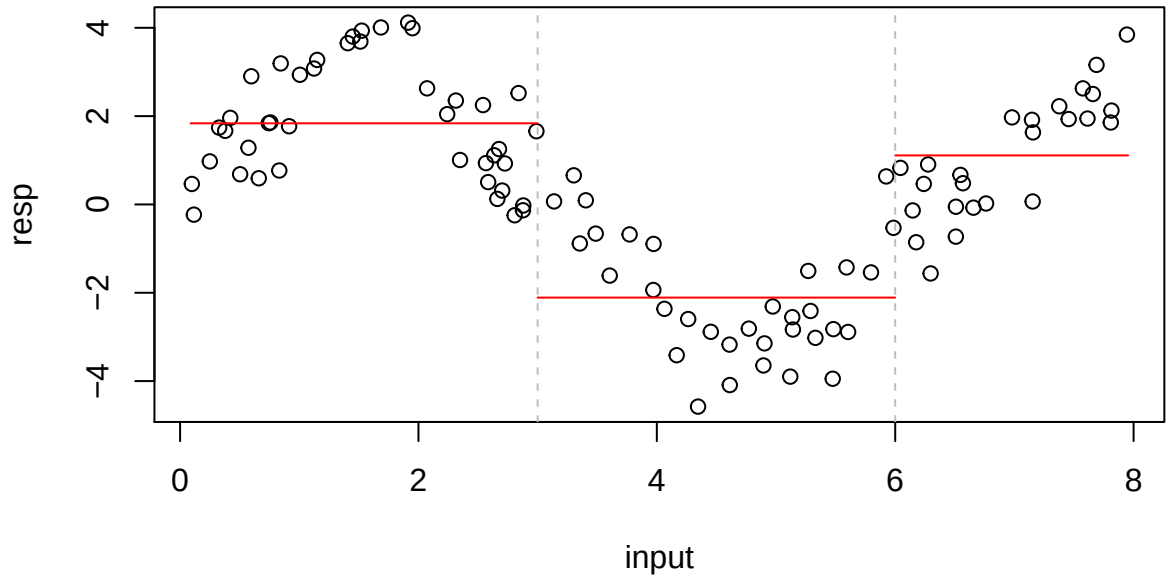
Computer Lab 7

Thomas Zhang

2015-11-30

Assignment 1

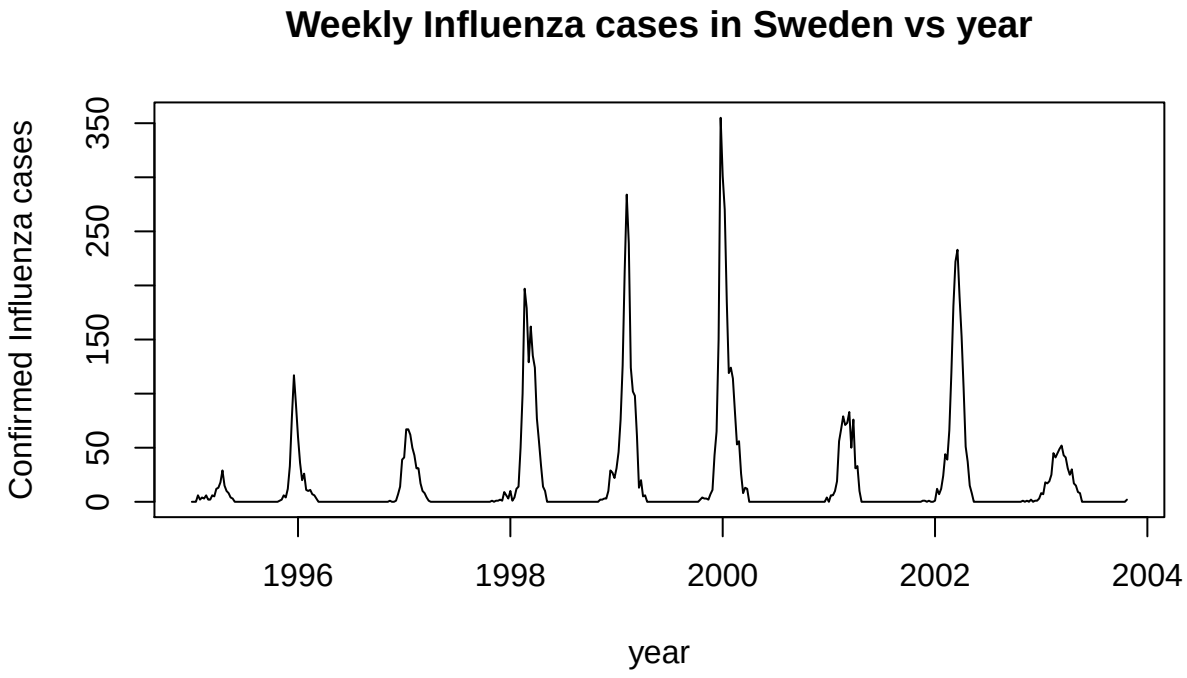
We fit a piece-wise constant function to our data with knots at $x = 3$ and $x = 6$.

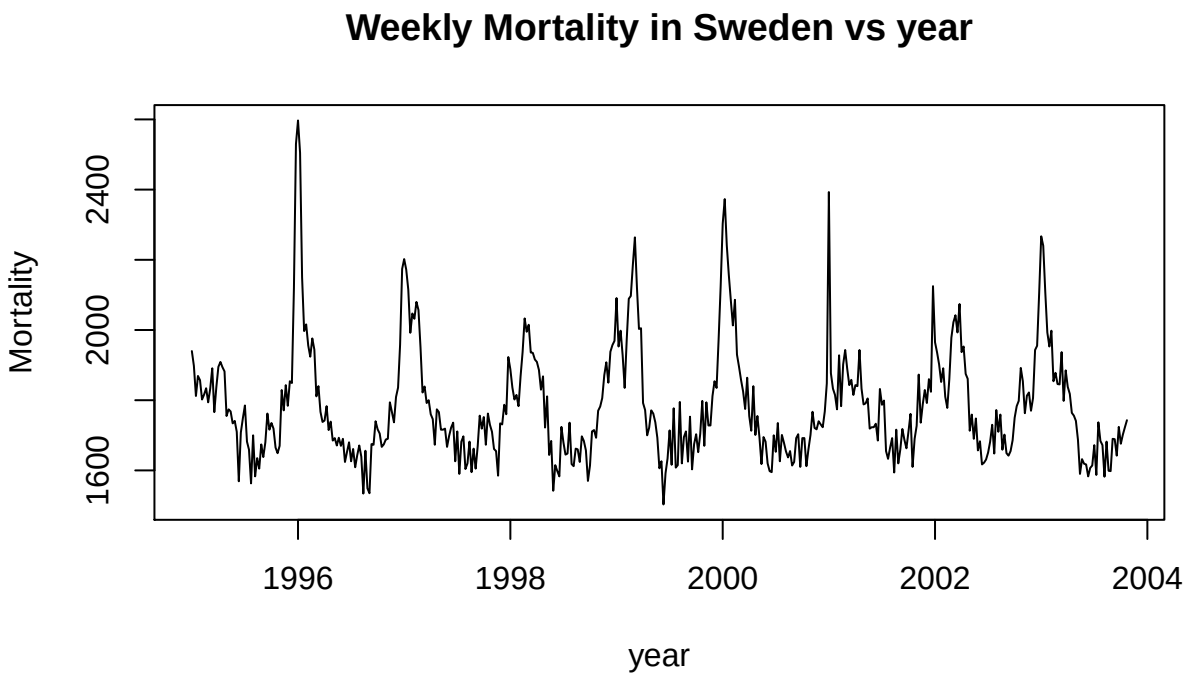


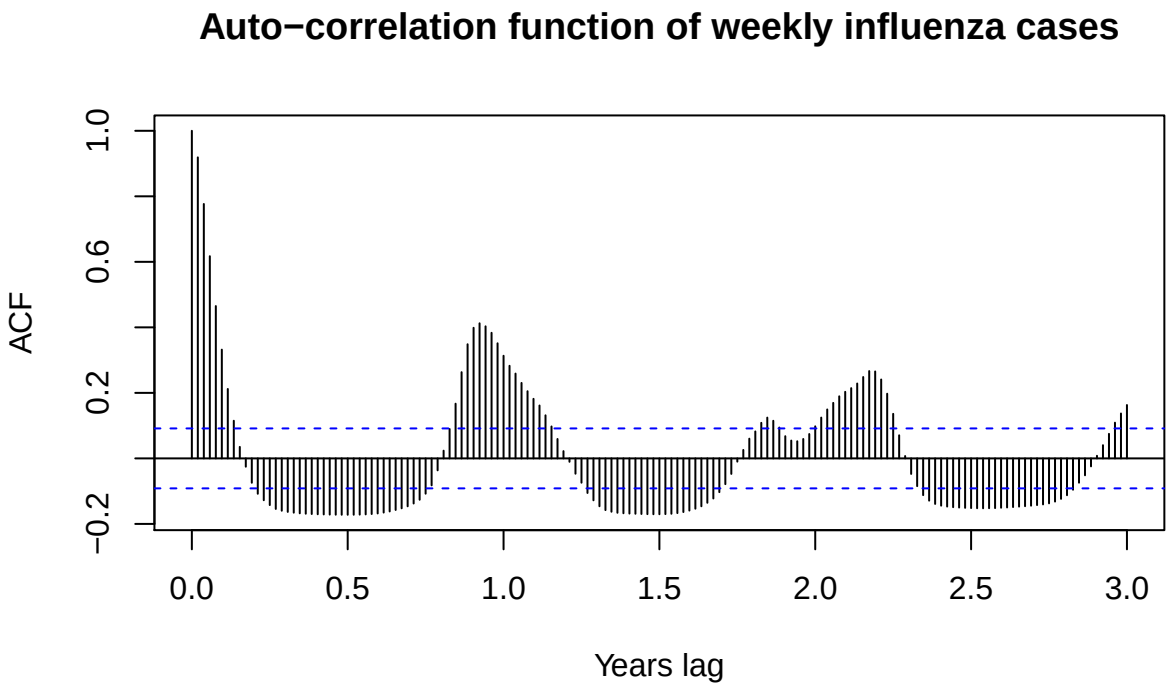
The red lines lie at $y = 1.838$, $y = -2.109$ and $y = 1.111$ respectively.

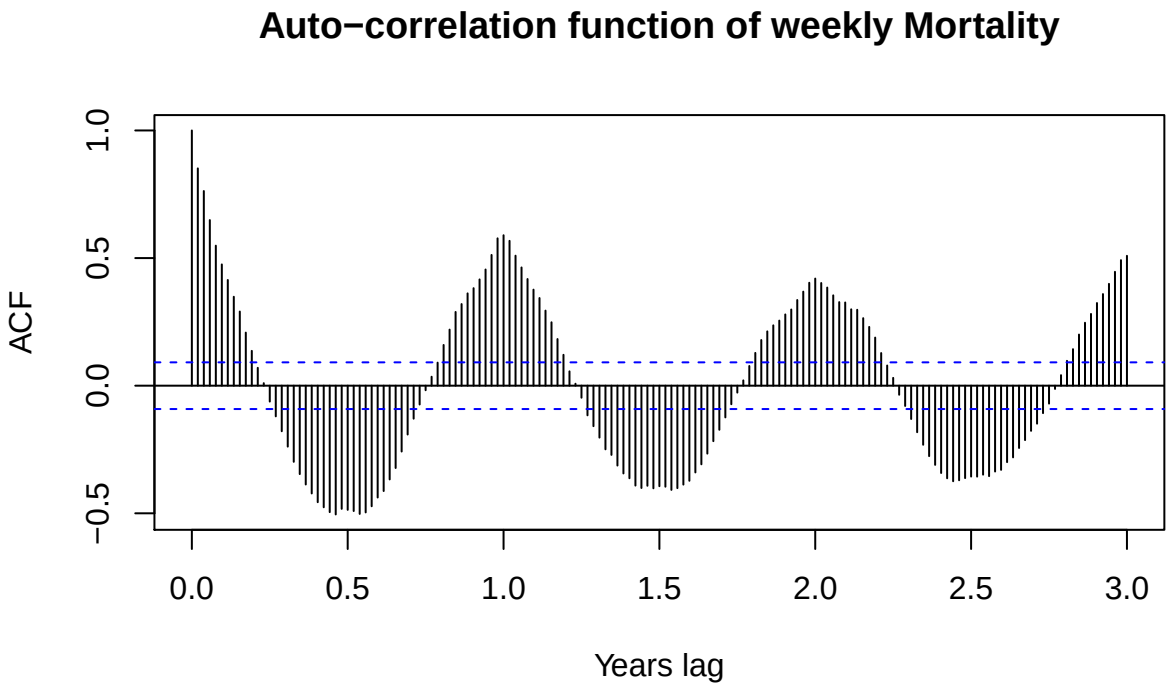
Assignment 2

We turn out attention to a data set containing data over weekly Mortality numbers in Sweden from 1995 to 2005 We also have data over the weekly clinically confirmed influenza cases in the data set. Let us plot this data and see if there is and autocorrelation within them.


Weekly Influenza cases in Sweden vs year
Confirmed Influenza cases
0
50
150
250
350
1996
1998
2000
2002
2004
year


Weekly Mortality in Sweden vs year
Mortality
1600
2000
2400
1996
1998
2000
2002
2004
year

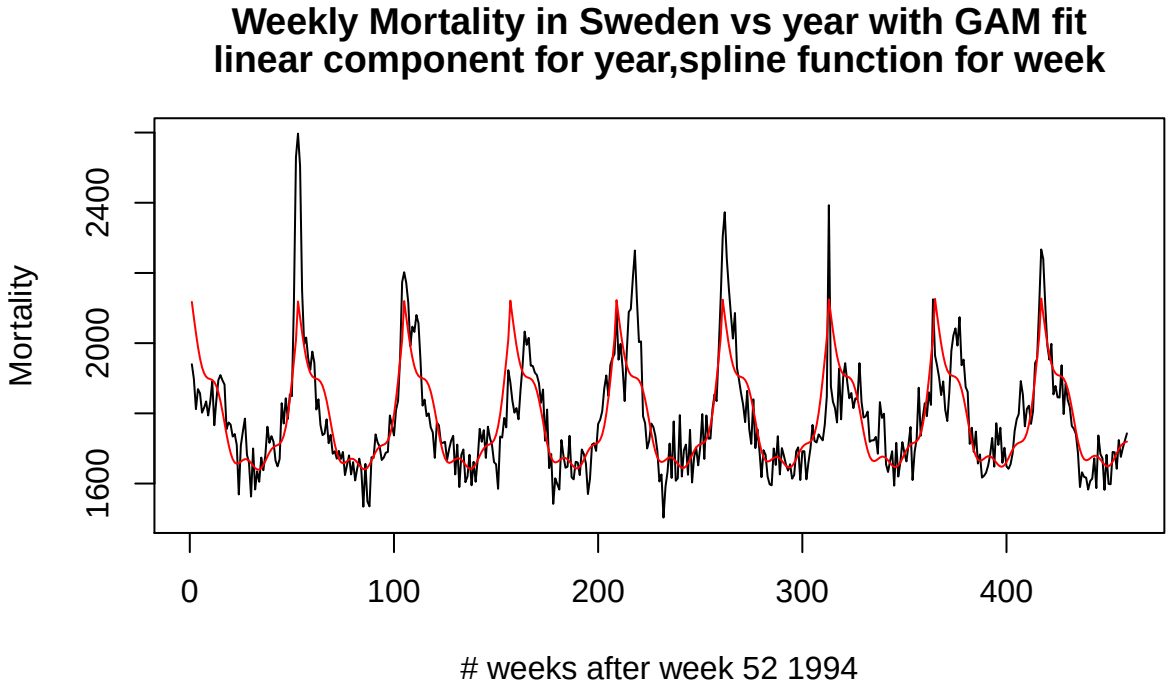




We clearly see that both Mortality and Influenza cases are highly autocorrelated, both with nearby data points and with a seasonal trend. We try to create a GAM in R where Mortality depends on a linear term in Year and a spline term in Week, that is we say that

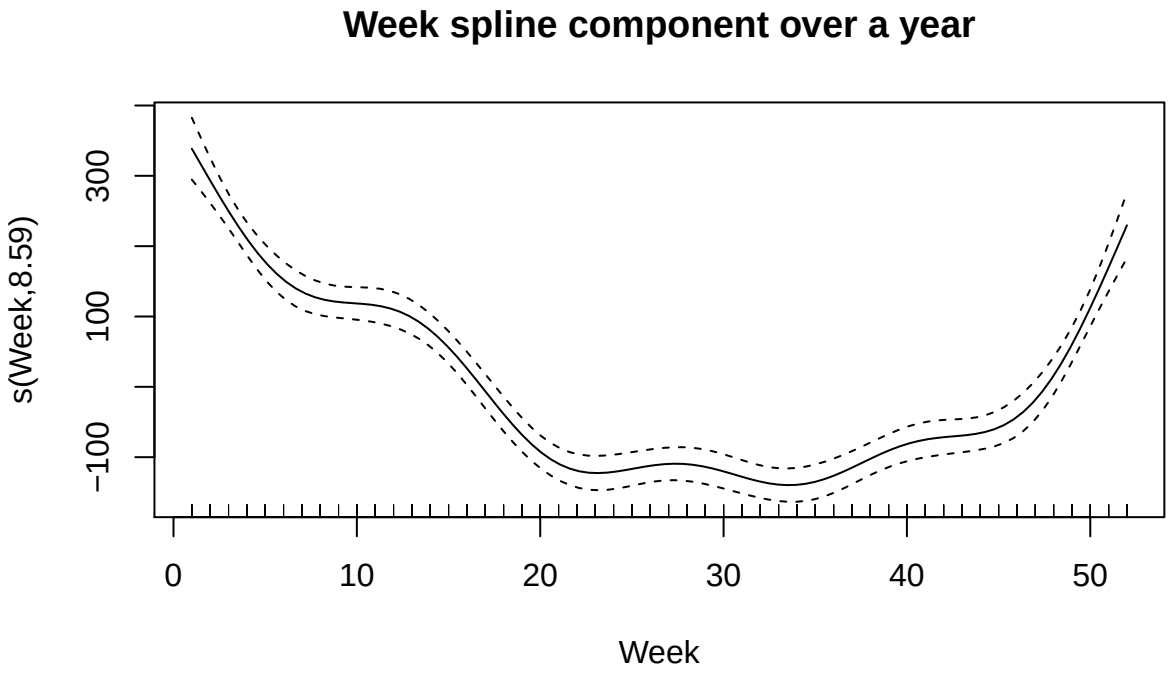
$$Mortality = \beta_{year} X_{year} + s(X_{week}) + \varepsilon$$

where ε is normally distributed noise and $s()$ is a spline function based on week data points.



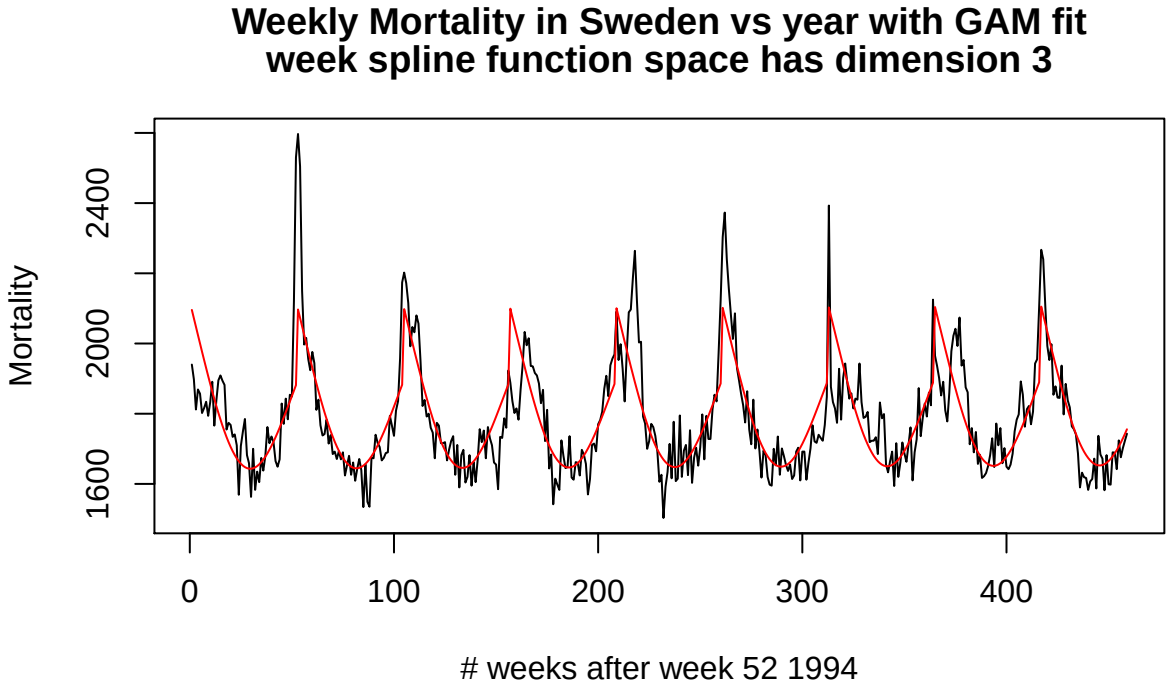

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ Year + s(Week)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -652.058   3448.379  -0.189     0.85
## Year           1.219      1.725   0.706     0.48
##
## Approximate significance of smooth terms:
##           edf Ref.df     F p-value
## s(Week) 8.587  8.951 100.6  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.661   Deviance explained = 66.8%
## GCV = 9014.6  Scale est. = 8806.7    n = 459
```

The fitted value of this GAM looks, ok. It looks as if the Year coefficient is not significant, while the smooth spline term is highly significant. Let us plot the smooth spline function over all weeks.

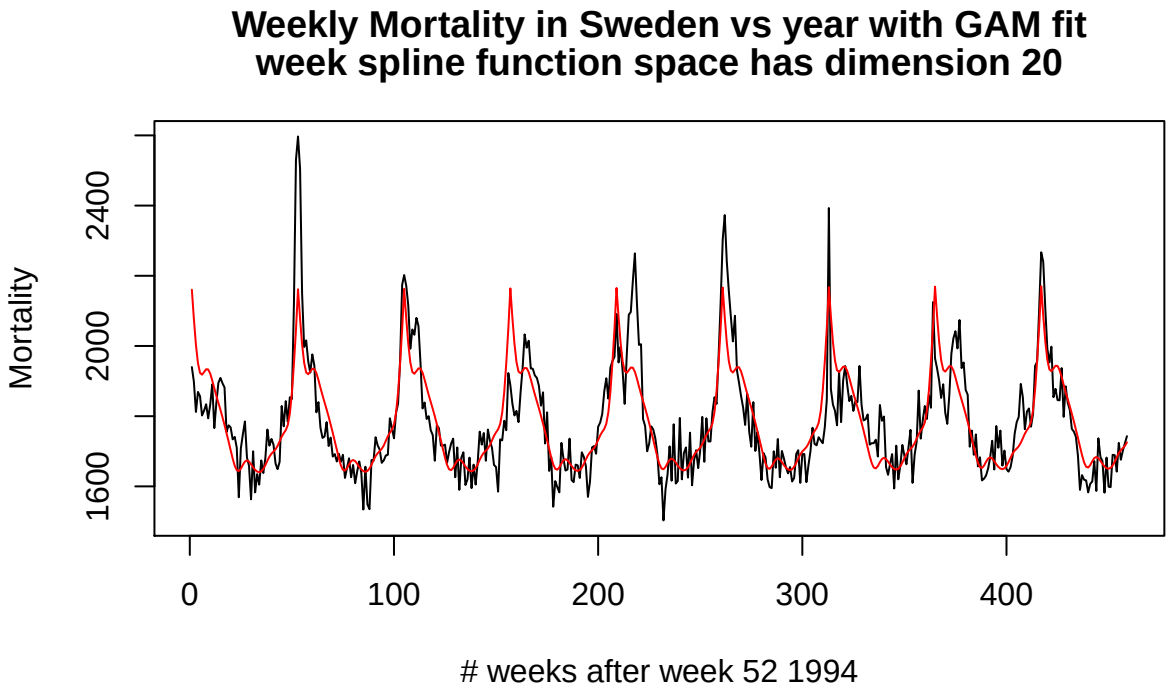


We clearly see that the mortality is higher during the winter weeks compared to mortality during the summer. In seems this accounts for the seasonal pattern in the GAM.

Let us try to vary the number of dimensions of the basis of the spline term and see how that affects the deviance explained by the model and the fitted model itself.



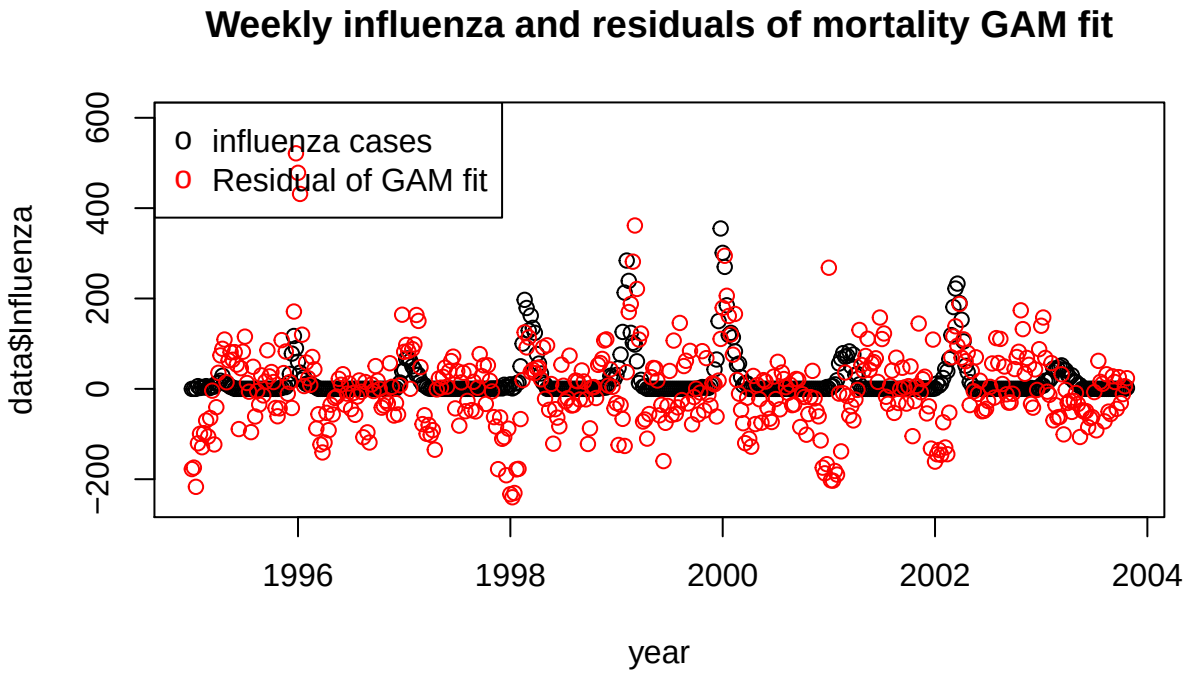
```
## [1] "Deviance explained by fit, Week spline basis dimension 3: 61.6 %"
```




```
## [1] "Deviance explained by fit, Week spline basis dimension 20: 68.8 %"
```

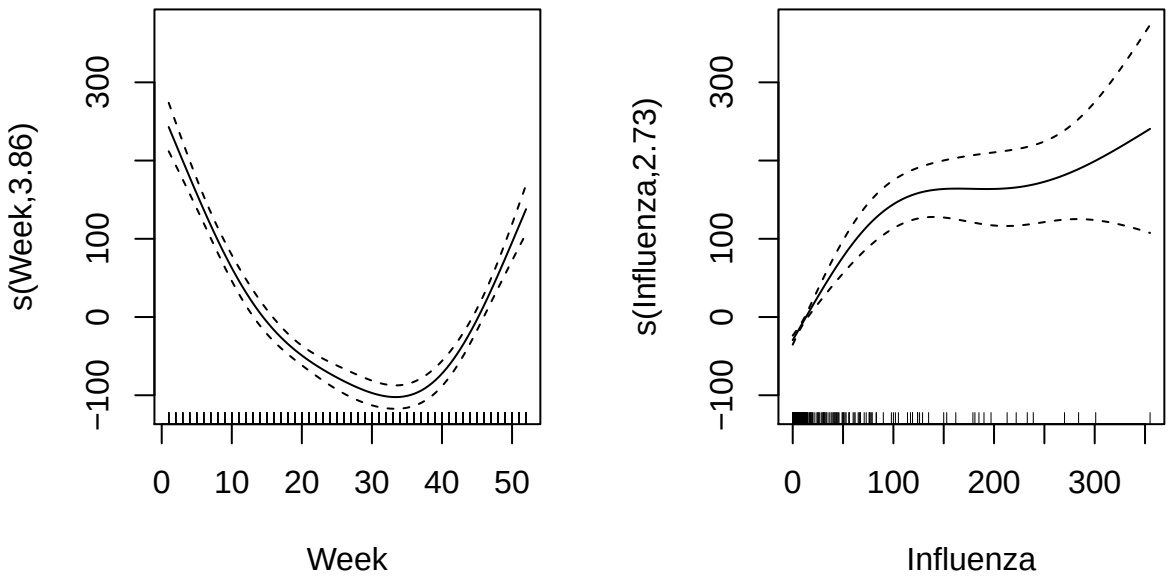
It seems as if increasing the number of basis dimensions for the spline term leads to more overfitting models and a larger part of the deviance explained. I infer that the penalty factor is decreased for higher values of k, the number of basis dimensions for the spline term, and hence higher degrees of freedom.

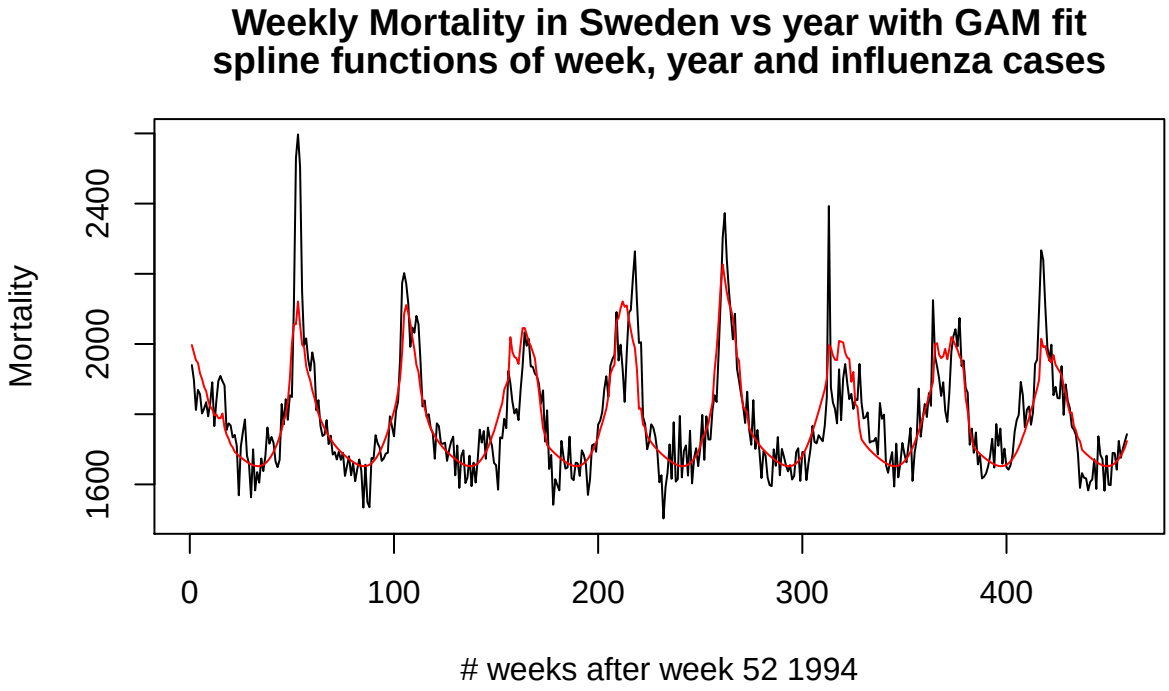
A plot between residuals of our first GAM and the weekly mortality rate is next shown below:



We note that although the residuals appear random, during the influenza spires of year 1996, 1999 and year 2000 the residuals were highly skewed to the positive during the same weeks. This could suggest that influenza cases could be a predictor for mortality rate. We try to make a GAM which takes this suggestion into account, as well as discarding the non-significant year predictor.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ s(Week, k = 5) + s(Influenza, k = 4)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1783.765      4.085   436.7   <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)      3.862  3.988 96.95  <2e-16 ***
## s(Influenza) 2.735  2.947 42.18  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.705   Deviance explained = 70.9%
## GCV =   7788  Scale est. = 7659.1    n = 459
```






```
## [1] "Deviance explained by this fit: 70.9 %"
```

```
## [1] "SSE for best GAM fit: 3457354"
```

It is seen that this model is better, compared to the previous ones. Deviance explained is higher and the plot

looks nicer. The spline terms are both significant. It seems as if the mortality rate is influenced by the out breaks of influenza, even though the influenza spline term experiences high uncertainty for large numbers of influenza cases. This uncertainty could be adressed by more data of influenza outbreaks.

Appendix

R code

```
data <- read.csv2("D:/R_HW/ML-Lab-2/data/cube.csv")
#head(data)

piecew_const <- function(resp,input,knots){
  intervals <- list(c(min(input)-0.01,knots[1]))
  if(length(knots) > 1){
    for(i in 1:(length(knots)-1)){
      intervals[[i+1]] <- c(knots[i],knots[i+1])
    }
    intervals[[length(knots)+1]] <- c(knots[length(knots)],max(input)+0.01)
  } else{
    intervals[[2]] <- c(knots[1],max(input))
  }

#   sortedx <- sort(data$input,index.return = TRUE)
#   sortedy <- data$resp[sortedx$ix]
  listofgroups <- list()


  for(i in 1:length(intervals)){

    for(j in 1:length(input)){
      if(input[j] > intervals[[i]][1] && input[j] <
         intervals[[i]][2]){
        if(length(listofgroups) < i){
          listofgroups[[i]] <- resp[j]
        } else{
        listofgroups[[i]] <- c(listofgroups[[i]],resp[j])
        }
      }
    }
  }

  means <- c()
  for(i in 1:length(listofgroups)){
    means[i] <- mean(listofgroups[[i]])
  }
  plot(input,resp)
  for(i in 1:length(knots)){
    abline(v=knots[i],lty=2,col="gray")
    lines(intervals[[i]],rep(means[i],2),col="red")
  }
  lines(intervals[[length(knots)+1]],rep(means[length(knots)+1],2),col="red")
  return(means)
```

```
}

means <- piecew_const(data$y,data$x,c(3,6))
library(mgcv)
data <- read.csv2("D:/R_HW/ML-Lab-2/data/influenza.csv")

influ_data_ts <- ts(data$Influenza,start=c(data$Year[1],
                                    data$Week[1]),freq=52)
mort_data_ts <- ts(data$Mortality,start=c(data$Year[1],
                                          data$Week[1]),freq=52)
plot(influ_data_ts,main="Weekly Influenza cases in Sweden vs year",
     xlab="year",ylab="Confirmed Influenza cases")
plot(mort_data_ts,main="Weekly Mortality in Sweden vs year",
     xlab="year",ylab="Mortality")
acf(influ_data_ts,
    main="Auto-correlation function of weekly influenza cases",
    xlab="Years lag",
    lag.max=52*3)
acf(mort_data_ts,
    main="Auto-correlation function of weekly Mortality",
    xlab="Years lag",
    lag.max=52*3)


gam_mort <- gam(Mortality ~ Year + s(Week),data=data)

#str(gam_mort)
plot(data$Mortality,type="l",
     main=c("Weekly Mortality in Sweden vs year with GAM fit",
            "linear component for year,spline function for week"),
     xlab="# weeks after week 52 1994",ylab="Mortality")
lines(gam_mort$fitted.values,col="red")
summary(gam_mort)

plot(gam_mort, main="Week spline component over a year")
gam_mort2 <- gam(formula=Mortality ~ Year + s(Week, k=3), data=data)
gam_mort3 <- gam(formula=Mortality ~ Year + s(Week, k=20), data=data)



plot(data$Mortality,type="l",
     main=c("Weekly Mortality in Sweden vs year with GAM fit",
     "week spline function space has dimension 3"),
     xlab="# weeks after week 52 1994",ylab="Mortality")
lines(gam_mort2$fitted.values, col="red")
paste("Deviance explained by fit, Week spline basis dimension 3:",61.6,"%")
plot(data$Mortality,type="l",
     main=c("Weekly Mortality in Sweden vs year with GAM fit",
            "week spline function space has dimension 20"),
     xlab="# weeks after week 52 1994",ylab="Mortality")
lines(gam_mort3$fitted.values, col="red")
paste("Deviance explained by fit, Week spline basis dimension 20:",68.8,"%")
plot(data$Time,data$Influenza,ylim=c(-250,600),
```

```
     main="Weekly influenza and residuals of mortality GAM fit",
     xlab="year")
points(data$Time,gam_mort$residuals,col="red")
legend("topleft",legend=c("influenza cases","Residual of GAM fit"),
       pch=c("o","o"),col=c("black","red"))

gam_mort4 <- gam(formula=Mortality ~ s(Week, k=5) + s(Influenza, k=4), data=data)
summary(gam_mort4)
par(mfrow=c(1,2))
plot(gam_mort4)
par(mfrow=c(1,1))



plot(data$Mortality,type="l",
     main=c("Weekly Mortality in Sweden vs year with GAM fit",
            "spline functions of week, year and influenza cases"),
     xlab="# weeks after week 52 1994",ylab="Mortality")
lines(gam_mort4$fitted.values, col="red")
paste("Deviance explained by this fit:",70.9,"%")
sse <- sum((data$Mortality-gam_mort4$fitted.values)^2)
paste("SSE for best GAM fit:",round(sse,0))
## NA
```