

machine learning(732A99) lab2 block2

Anubhav Dikshit(anudi287)

17 December 2018

Contents

Assignment 1	2
Loading The Libraries	2
1. Use time series plots to visually inspect how the mortality and influenza number vary with time (use Time as X axis). By using this plot, comment how the amounts of influenza cases are related to mortality rates.	2
2. Use gam() function from mgcv package to fit a GAM model in which Mortality is normally distributed and modelled as a linear function of Year and spline function of Week, and make sure that the model parameters are selected by the generalized cross-validation. Report the underlying probabilistic model.	4
3. Plot predicted and observed mortality against time for the fitted model and comment on the quality of the fit. Investigate the output of the GAM model and report which terms appear to be significant in the model. Is there a trend in mortality change from one year to another? Plot the spline component and interpret the plot.	5
4. Examine how the penalty factor of the spline function in the GAM model from step 2 influences the estimated deviance of the model. Make plots of the predicted and observed mortality against time for cases of very high and very low penalty factors. What is the relation of the penalty factor to the degrees of freedom? Do your results confirm this relationship?	8
5. Use the model obtained in step 2 and plot the residuals and the influenza values against time (in one plot). Is the temporal pattern in the residuals correlated to the outbreaks of influenza?	12
6. Fit a GAM model in R in which mortality is be modelled as an additive function of the spline functions of year, week, and the number of confirmed cases of influenza. Use the output of this GAM function to conclude whether or not the mortality is influenced by the outbreaks of influenza. Provide the plot of the original and fitted Mortality against Time and comment whether the model seems to be better than the previous GAM models.	13
Appendix	13

Assignment 1

Loading The Libraries

```
if (!require("pacman")) install.packages("pacman")
pacman::p_load(xlsx, ggplot2, tidyr, dplyr, reshape2, gridExtra,
               mgcv, rgl, akima)

set.seed(12345)
options("jtools-digits" = 2, scipen = 999)

# colours (colour blind friendly)
cbPalette <- c("#999999", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7")

## Making title in the center
theme_update(plot.title = element_text(hjust = 0.5))
```

1. Use time series plots to visually inspect how the mortality and influenza number vary with time (use Time as X axis). By using this plot, comment how the amounts of influenza cases are related to mortality rates.

```
set.seed(12345)

# Importing data
flu_data = read.xlsx("influenza.xlsx", sheetName = "Raw data")
flu_data$Time_fixed <- as.Date(paste(flu_data$Year, flu_data$Week, 1, sep="-"), "%Y-%U-%u")
flu_data$influ_perc <- (flu_data$Influenza/flu_data$Mortality) * 100

# Plot

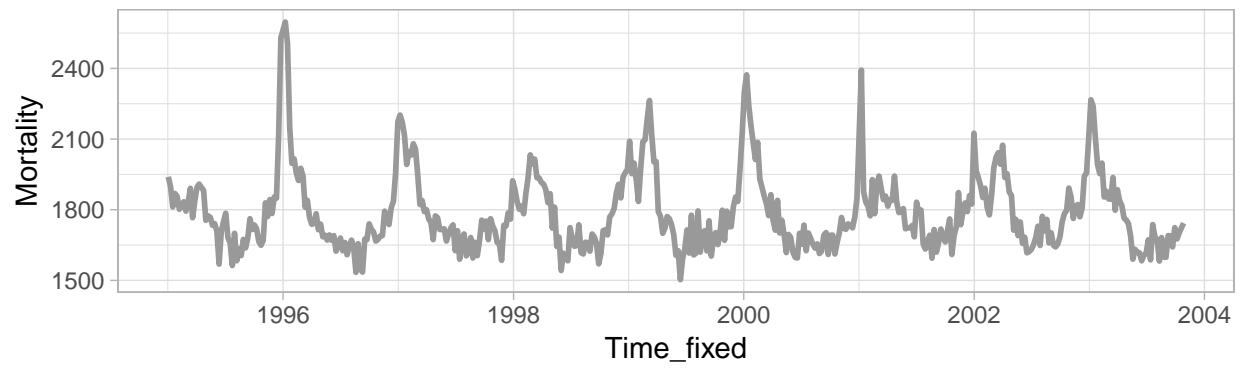
p1 <- ggplot(flu_data, aes(x=Time_fixed, y = Mortality)) +
  geom_line(color = "#999999", size = 1) +
  scale_fill_brewer() +
  theme_light() +
  ggtitle("Time series of Mortality")

p2 <- ggplot(flu_data, aes(x=Time_fixed, y = Influenza)) +
  geom_line(color = "#E69F00", size = 1) +
  scale_fill_brewer() +
  theme_light() +
  ggtitle("Time series of Influenza")

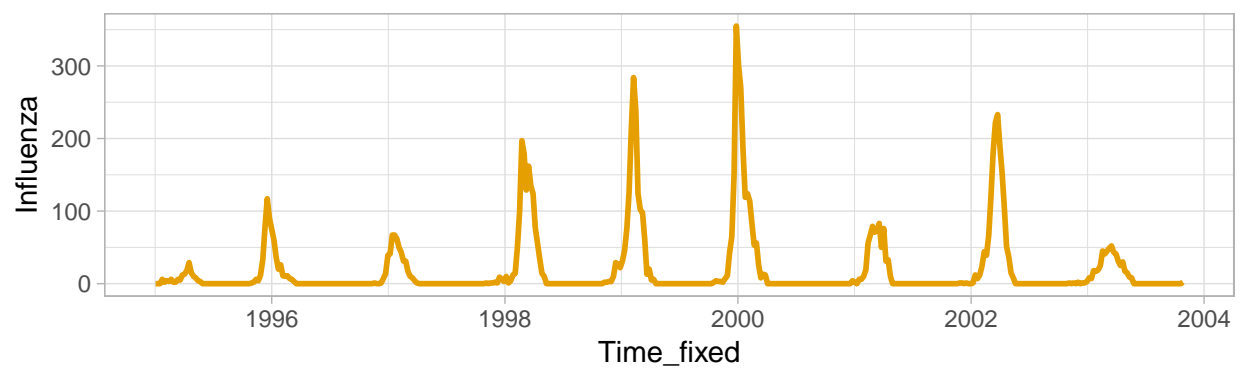
p3 <- ggplot(flu_data, aes(x=Time_fixed, y = influ_perc)) +
  geom_line(color = "#56B4E9", size = 1) +
  scale_fill_brewer() +
  theme_light() +
  ggtitle("Time series of % Mortality due to Influenza")

gridExtra::grid.arrange(p1, p2, ncol=1)
```

Time series of Mortality

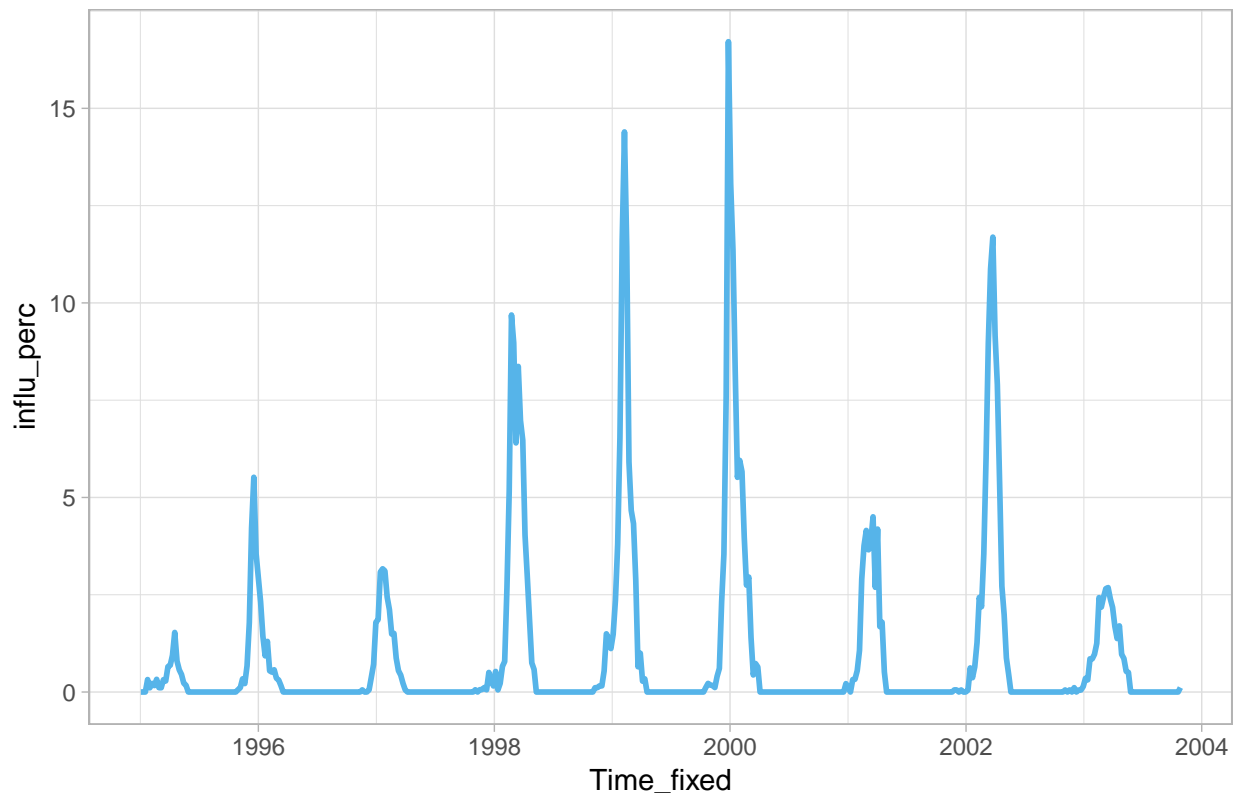


Time series of Influenza



p3

Time series of % Mortality due to Influenza



Analysis: From the plots is we can defintely see that Influenza and Mortality in the given dataset are in sync, everytime Mortality peaks so does influenza, however the magnitiude of peaking is not in sync, that is the highest cases of mortality were observed in '1996' while for influenza its in year '2000'.

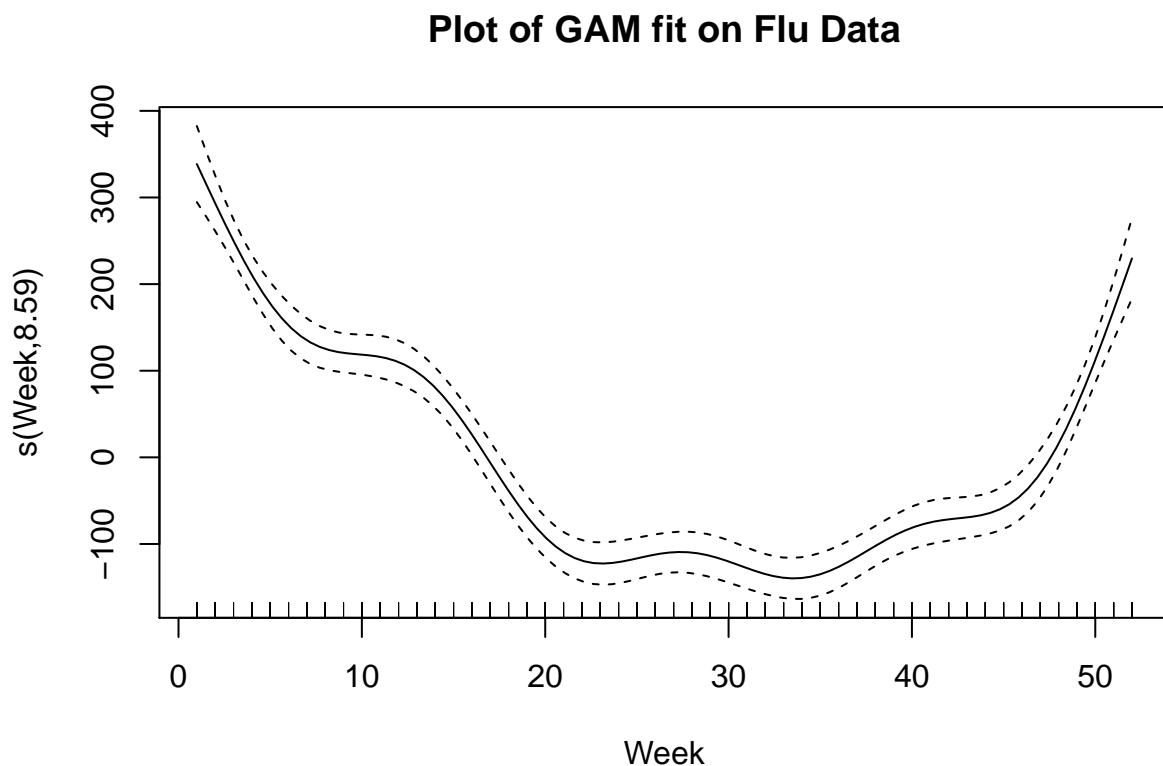
From the third plot, we can see the percentage of mortality due to influenza, here also the peaks match with the other plots, suggests that these two events are closely correlated.

2. Use `gam()` function from `mgcv` package to fit a GAM model in which Mortality is normally distributed and modelled as a linear function of Year and spline function of Week, and make sure that the model parameters are selected by the generalized cross-validation. Report the underlying probabilistic model.

```
gam_model <- mgcv::gam(data = flu_data, Mortality~Year+s(Week), method = "GCV.Cp")
summary(gam_model)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ Year + s(Week)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -652.060   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  8.951 100.3 <0.0000000000000002 ***
## ---
## 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
#plot the fit
p4 <- plot(gam_model, main= "Plot of GAM fit on Flu Data")
```



3. Plot predicted and observed mortality against time for the fitted model and comment on the quality of the fit. Investigate the output of the GAM model and report which terms appear to be significant in the model. Is there a trend in mortality change from one year to another? Plot the spline component and interpret the plot.

```
temp <- flu_data
temp$Fitted_Mortality <- gam_model$fitted.values

p5 <- ggplot(data=temp, aes(x = Time_fixed, y = Fitted_Mortality)) +
  geom_line(color = "#009E73", size = 1) +
```

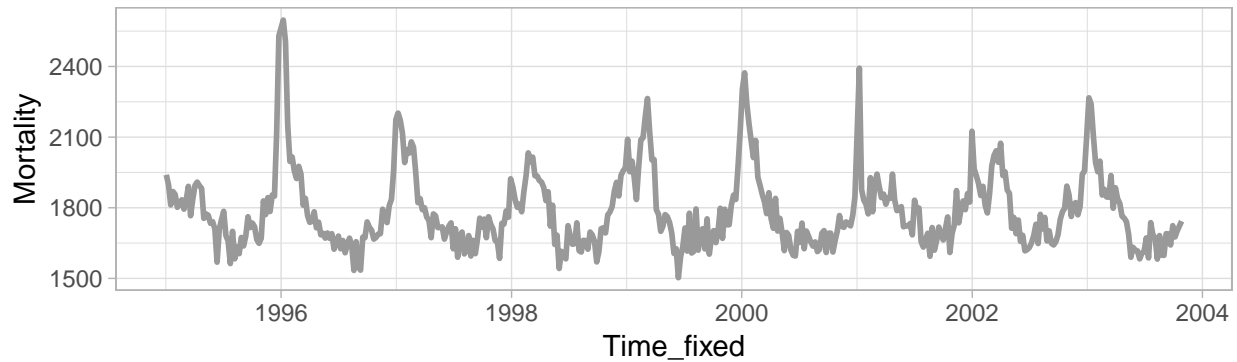
```

scale_fill_brewer() +
  theme_light() +
  ggtitle("Time series of Fitted Mortality")

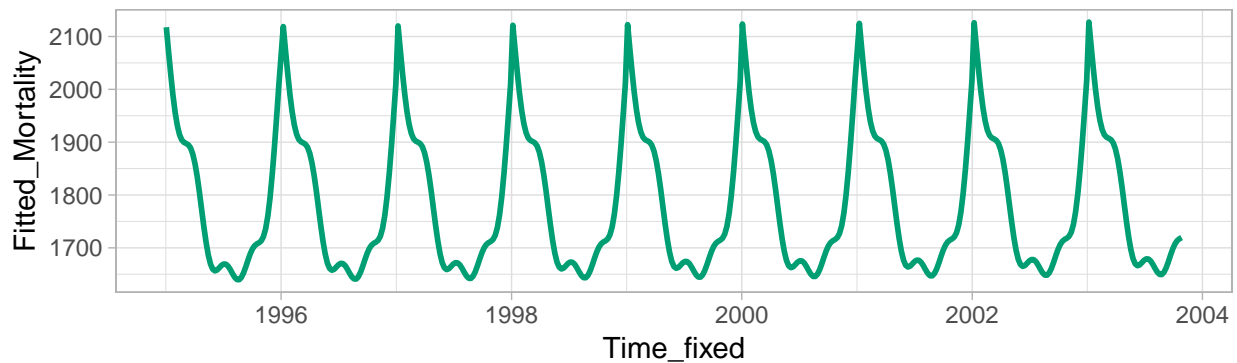
grid.arrange(p1, p5, nrow = 2)

```

Time series of Mortality



Time series of Fitted Mortality



```
summary(gam_model)
```

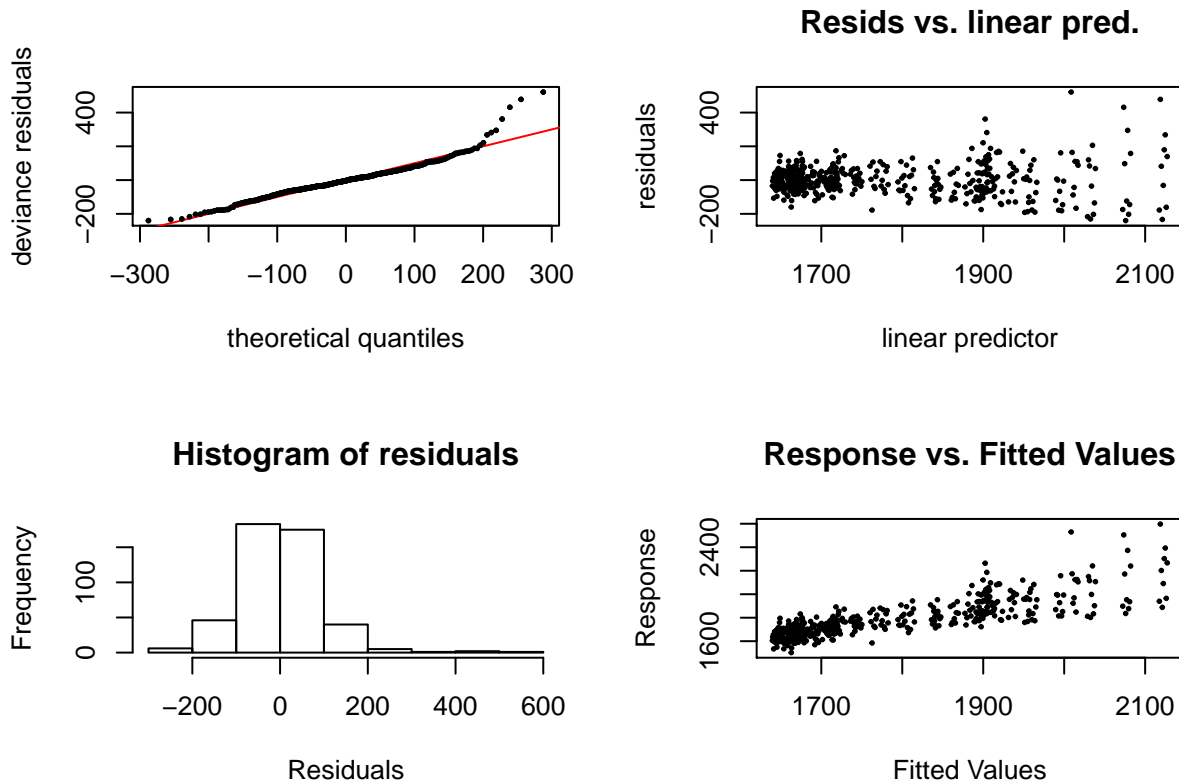
```

##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ Year + s(Week)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -652.060   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   8.951 100.3 <0.000000000000002 ***
## ---
## 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
```

```
gam.check(gam_model,pch=19,cex=.3)
```



```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 7 iterations.
## The RMS GCV score gradient at convergence was 0.114505 .
## The Hessian was positive definite.
## Model rank = 11 / 11
##
## 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) 9.00 8.59   1.04   0.74
```

```
s=interp(temp$Year,temp$Week, fitted(gam_model))
persp3d(s$x, s$y, s$z, col="red")
```

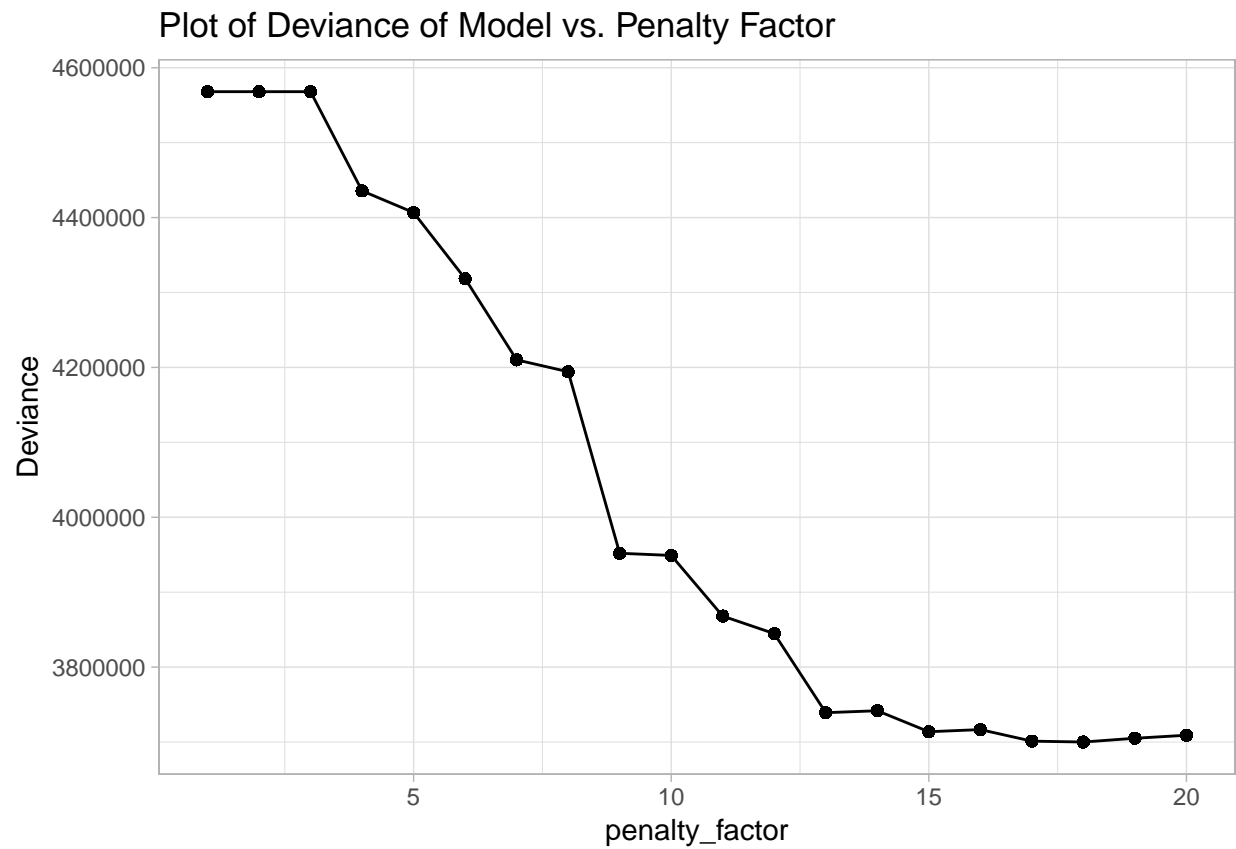
Analysis: From the plot of residuals we can see that the residuals are normally distributed. Thus this is a good fit.

4. Examine how the penalty factor of the spline function in the GAM model from step 2 influences the estimated deviance of the model. Make plots of the predicted and observed mortality against time for cases of very high and very low penalty factors. What is the relation of the penalty factor to the degrees of freedom? Do your results confirm this relationship?

```
model_deviance <- NULL
for(k in seq(from = 1, to = 20, by = 1))
{
  gam_model <- mgcv::gam(data = flu_data, Mortality~Year+s(Week, k=k), method = "GCV.Cp")
  temp <- cbind(gam_model$deviance, gam_model$fitted.values, gam_model$y, flu_data$Time_fixed,
               k, sum(influence(gam_model)))

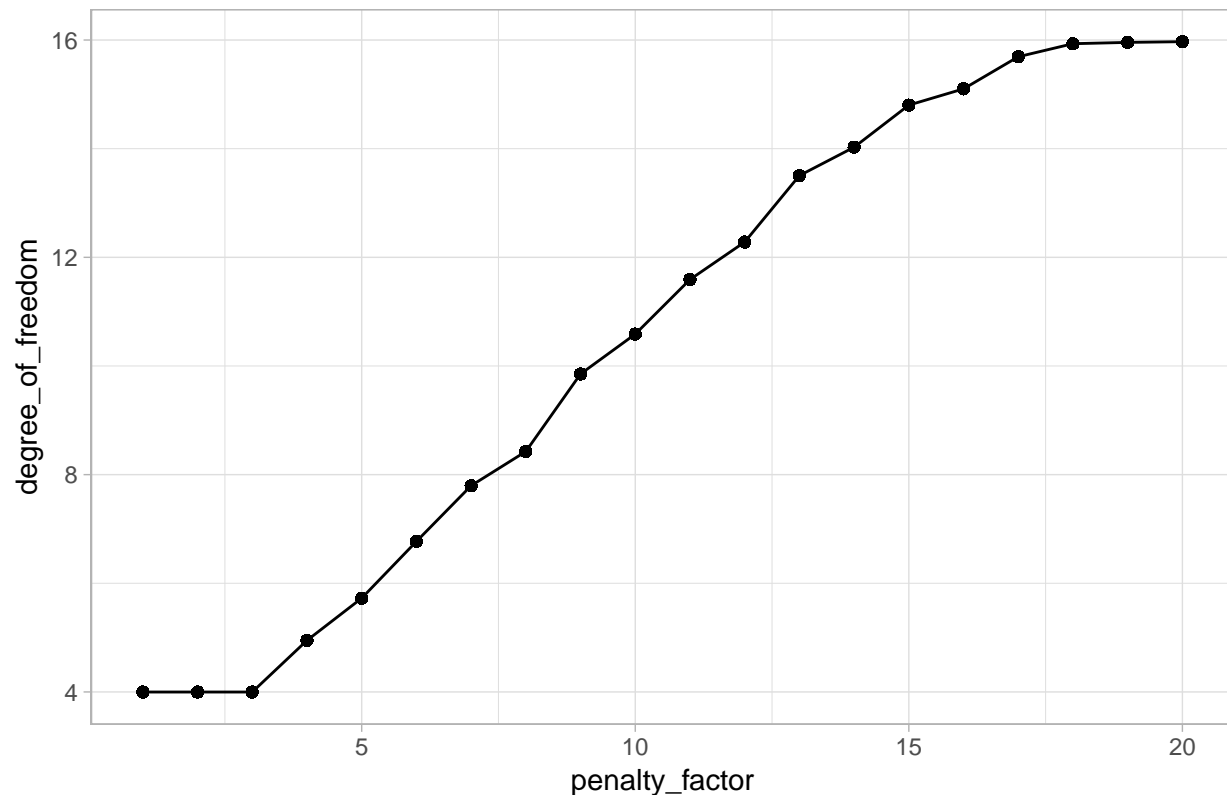
  model_deviance <- rbind(temp, model_deviance)
}
model_deviance <- as.data.frame(model_deviance)
colnames(model_deviance) <- c("Deviance", "Predicted_Mortality", "Mortality", "Time",
                             "penalty_factor", "degree_of_freedom")
model_deviance$Time <- as.Date(model_deviance$Time, origin = '1970-01-01')

# plot of deviance
p6 <- ggplot(data=model_deviance, aes(x = penalty_factor, y = Deviance)) +
  geom_point() +
  geom_line() +
  theme_light() +
  ggtitle("Plot of Deviance of Model vs. Penalty Factor")
p6
```

```
# plot of degree of freedom
p7 <- ggplot(data=model_deviance, aes(x = penalty_factor, y = degree_of_freedom)) +
  geom_point() +
  geom_line() +
  theme_light() +
  ggtitle("Plot of degree_of_freedom of Model vs. Penalty Factor")
p7
```

Plot of degree_of_freedom of Model vs. Penalty Factor



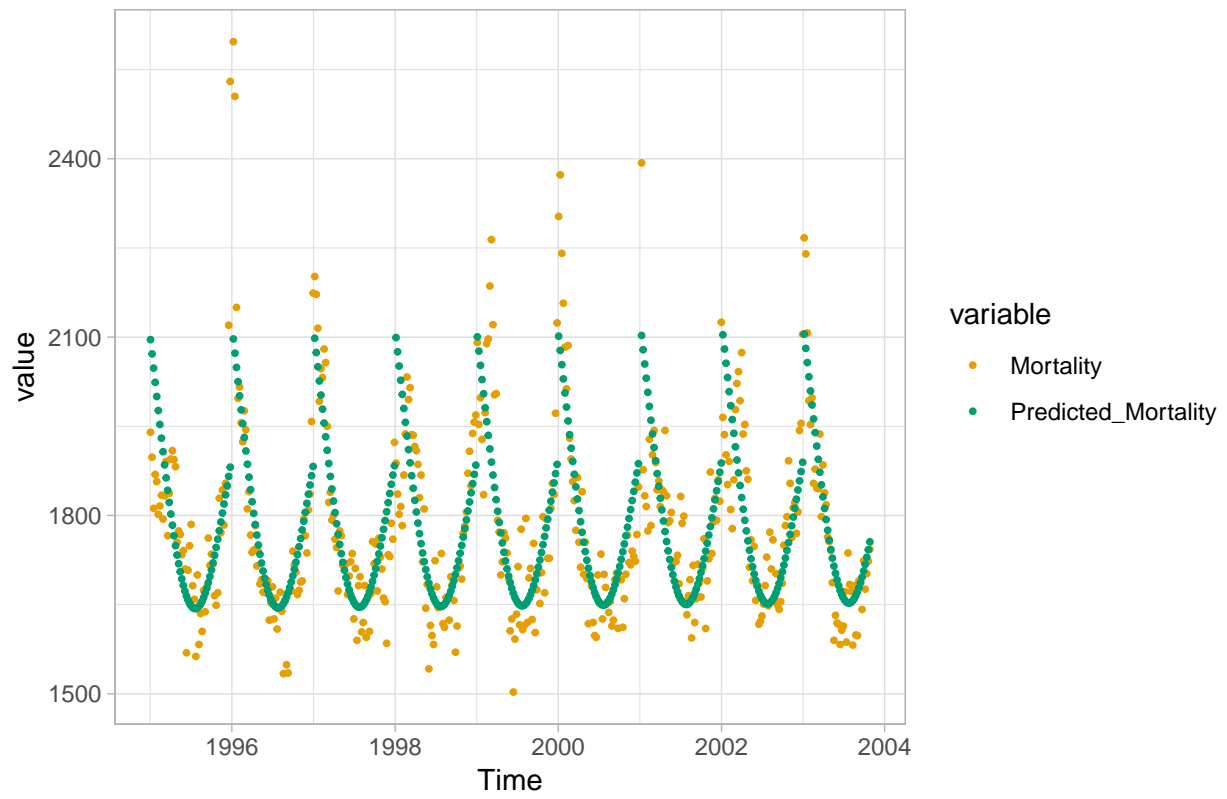
```
model_deviance_wide <- melt(model_deviance[,c("Time", "penalty_factor",
                                              "Mortality", "Predicted_Mortality")],
                           id.vars = c("Time", "penalty_factor")) %>%
  filter(penalty_factor %in% c("1", "5", "15", "20"))

# plot of predicted vs. observed mortality
p8 <- ggplot(data=model_deviance_wide[model_deviance_wide$penalty_factor == 1,],
            aes(x= Time, y = value)) +
  geom_point(aes(color = variable), size=0.7) +
  scale_color_manual(values=c("#E69F00", "#009E73")) +
  theme_light() +
  ggtitle("Plot of Mortality vs. Time(Penalty 1)")

p9 <- ggplot(data=model_deviance_wide[model_deviance_wide$penalty_factor == 20,],
            aes(x= Time, y = value)) +
  geom_point(aes(color = variable), size=0.7) +
  scale_color_manual(values=c("#E69F00", "#009E73")) +
  theme_light() +
  ggtitle("Plot of Mortality vs. Time(Penalty 20)")
```

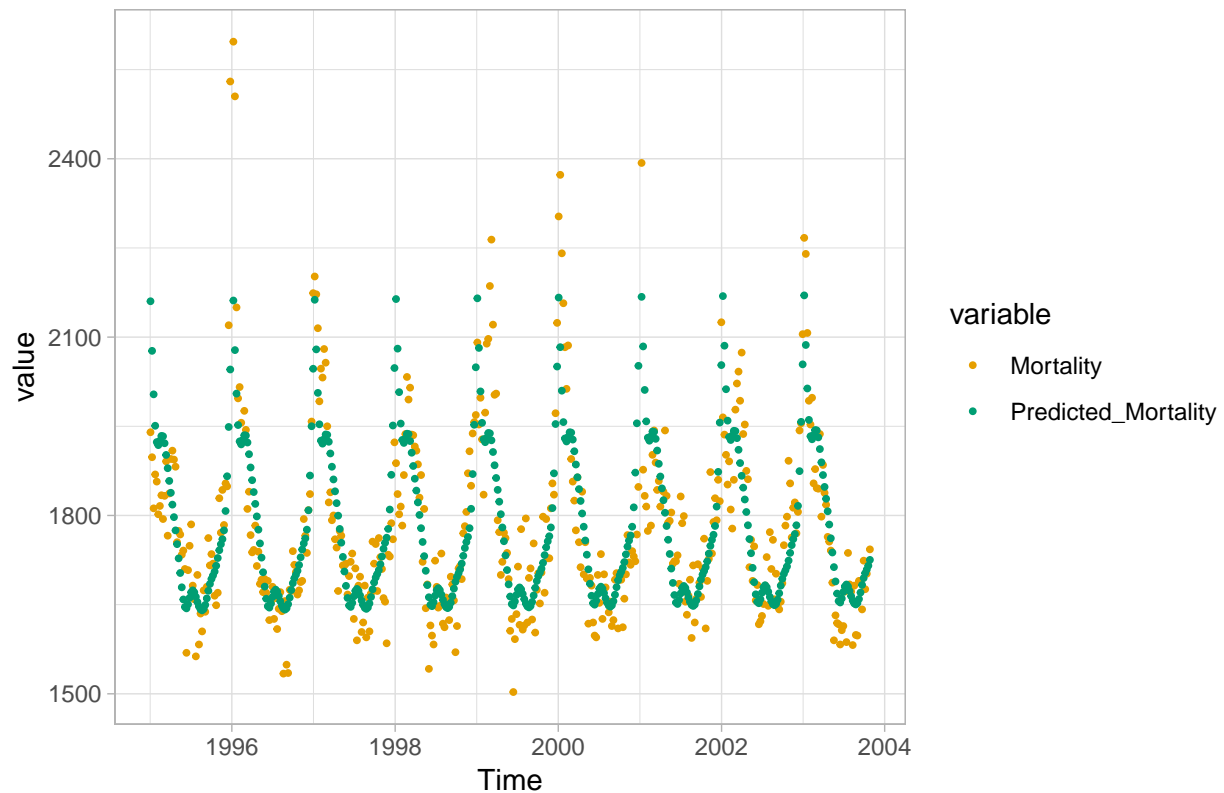
p8

Plot of Mortality vs. Time(Penalty 1)



p9

Plot of Mortality vs. Time(Penalty 20)



Analysis: theoretical maximum degree of freedom is $k-1$.

5. Use the model obtained in step 2 and plot the residuals and the influenza values against time (in one plot). Is the temporal pattern in the residuals correlated to the outbreaks of influenza?

```
gam_model <- mgcv::gam(data = flu_data, Mortality~Year+s(Week), method = "GCV.Cp")

temp <- flu_data
temp <- cbind(temp, gam_model$residuals)

#ggplot(data =temp)
```

6. Fit a GAM model in R in which mortality is be modelled as an additive function of the spline functions of year, week, and the number of confirmed cases of influenza. Use the output of this GAM function to conclude whether or not the mortality is influenced by the outbreaks of influenza. Provide the plot of the original and fitted Mortality against Time and comment whether the model seems to be better than the previous GAM models.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
if (!require("pacman")) install.packages("pacman")
pacman::p_load(xlsx, ggplot2, tidyr, dplyr, reshape2, gridExtra,
               mgcv, rgl, akima)

set.seed(12345)
options("jtools-digits" = 2, scipen = 999)

# colours (colour blind friendly)
cbPalette <- c("#999999", "#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7")

## Making title in the center
theme_update(plot.title = element_text(hjust = 0.5))
set.seed(12345)

# Importing data
flu_data = read.xlsx("influenza.xlsx", sheetName = "Raw data")
flu_data$Time_fixed <- as.Date(paste(flu_data$Year, flu_data$Week, 1, sep="-"), "%Y-%U-%u")
flu_data$influ_perc <- (flu_data$Influenza/flu_data$Mortality) * 100

# Plot

p1 <- ggplot(flu_data, aes(x=Time_fixed, y = Mortality)) +
  geom_line(color = "#999999", size = 1) +
  scale_fill_brewer() +
  theme_light() +
  ggtitle("Time series of Mortality")

p2 <- ggplot(flu_data, aes(x=Time_fixed, y = Influenza)) +
  geom_line(color = "#E69F00", size = 1) +
  scale_fill_brewer() +
  theme_light() +
  ggtitle("Time series of Influenza")

p3 <- ggplot(flu_data, aes(x=Time_fixed, y = influ_perc)) +
  geom_line(color = "#56B4E9", size = 1) +
  scale_fill_brewer() +
  theme_light() +
  ggtitle("Time series of % Mortality due to Influenza")

gridExtra::grid.arrange(p1, p2, ncol=1)
p3
```

```

gam_model <- mgcv::gam(data = flu_data, Mortality~Year+s(Week), method = "GCV.Cp")
summary(gam_model)
#plot the fit

p4 <- plot(gam_model, main= "Plot of GAM fit on Flu Data")
temp <- flu_data
temp$Fitted_Mortality <- gam_model$fitted.values

p5 <- ggplot(data=temp, aes(x = Time_fixed, y = Fitted_Mortality)) +
  geom_line(color = "#009E73", size = 1) +
  scale_fill_brewer() +
  theme_light() +
  ggtitle("Time series of Fitted Mortality")

grid.arrange(p1, p5, nrow = 2)

summary(gam_model)
gam.check(gam_model,pch=19,cex=.3)

s=interp(temp$Year,temp$Week, fitted(gam_model))
persp3d(s$x, s$y, s$z, col="red")
model_deviance <- NULL
for(k in seq(from = 1, to = 20, by = 1))
{
  gam_model <- mgcv::gam(data = flu_data, Mortality~Year+s(Week, k=k), method = "GCV.Cp")
  temp <- cbind(gam_model$deviance, gam_model$fitted.values, gam_model$y, flu_data$Time_fixed,
    k, sum(influence(gam_model)))

  model_deviance <- rbind(temp, model_deviance)
}
model_deviance <- as.data.frame(model_deviance)
colnames(model_deviance) <- c("Deviance", "Predicted_Mortality", "Mortality", "Time",
  "penalty_factor", "degree_of_freedom")
model_deviance$Time <- as.Date(model_deviance$Time, origin = '1970-01-01')

# plot of deviance
p6 <- ggplot(data=model_deviance, aes(x = penalty_factor, y = Deviance)) +
  geom_point() +
  geom_line() +
  theme_light() +
  ggtitle("Plot of Deviance of Model vs. Penalty Factor")
p6

# plot of degree of freedom
p7 <- ggplot(data=model_deviance, aes(x = penalty_factor, y = degree_of_freedom)) +
  geom_point() +
  geom_line() +
  theme_light() +
  ggtitle("Plot of degree_of_freedom of Model vs. Penalty Factor")
p7

model_deviance_wide <- melt(model_deviance[,c("Time", "penalty_factor",

```

```

                                "Mortality", "Predicted_Mortality"]],
                                id.vars = c("Time", "penalty_factor")) %>%
filter(penalty_factor %in% c("1", "5", "15", "20"))

# plot of predicted vs. observed mortality
p8 <- ggplot(data=model_deviance_wide[model_deviance_wide$penalty_factor == 1,],
             aes(x= Time, y = value)) +
  geom_point(aes(color = variable), size=0.7) +
  scale_color_manual(values=c("#E69F00", "#009E73")) +
  theme_light() +
  ggtitle("Plot of Mortality vs. Time(Penalty 1)")

p9 <- ggplot(data=model_deviance_wide[model_deviance_wide$penalty_factor == 20,],
             aes(x= Time, y = value)) +
  geom_point(aes(color = variable), size=0.7) +
  scale_color_manual(values=c("#E69F00", "#009E73")) +
  theme_light() +
  ggtitle("Plot of Mortality vs. Time(Penalty 20)")

p8
p9

gam_model <- mgcv::gam(data = flu_data, Mortality~Year+s(Week), method = "GCV.Cp")

temp <- flu_data
temp <- cbind(temp, gam_model$residuals)

#ggplot(data =temp)

```