Introduction to Machine Learning - Lab 7

Gustav Sternelöv

Friday, November 27, 2015

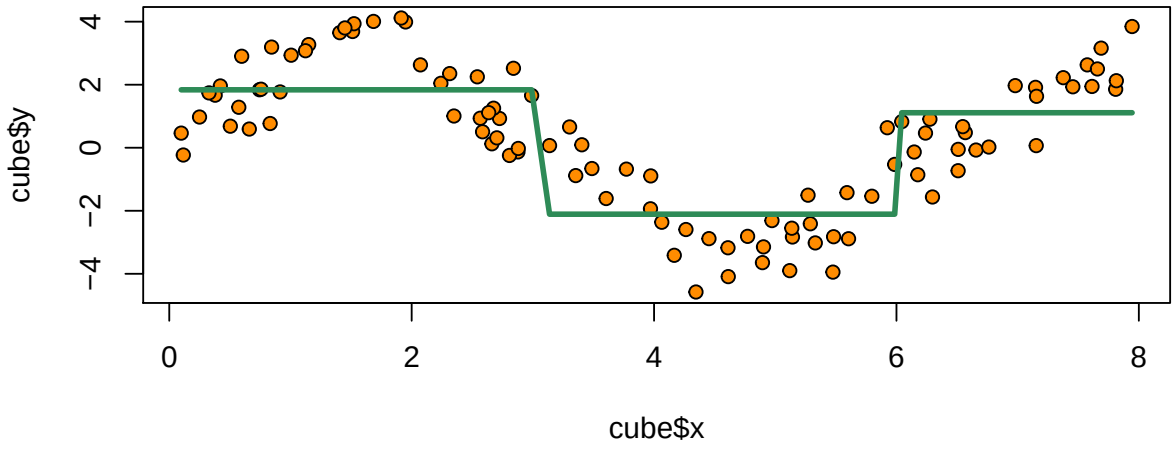
Assignment 1

1.1

In *1.1* a function that conducts a piecewise constant basis expansion is implemented. The function takes the argument *data*, where the first column is supposed to be the input vector and the second column the response vector, and the argument k which is the values for the knots. Returned by the function is a graph showing the results of the conducted expansion.

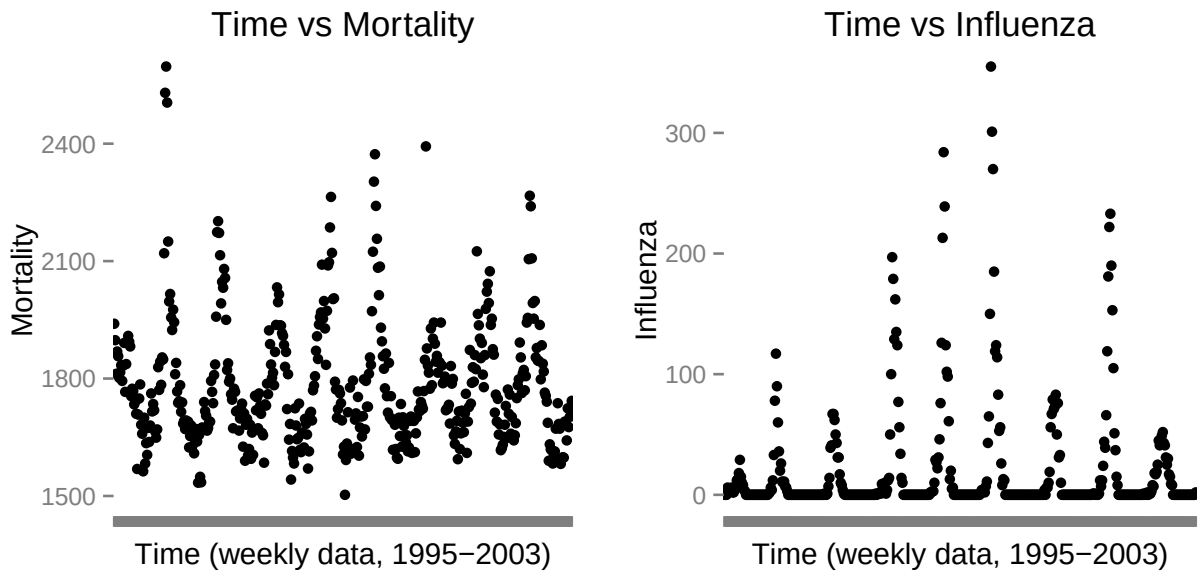
1.2

The created function is now tested on a dataset named *cube* which has two variables, x and y. Two knots are used, at x=3 and at x=6.
The result that is obtained by running the function on this data set is shown below.



Assignment 2

2.1



The graph over time versus mortality seem to follow a seasonal pattern. A closer look gives that the mortality is at its lowest during the summer and at its highest levels during the winter. This pattern is almost identical from year to year.

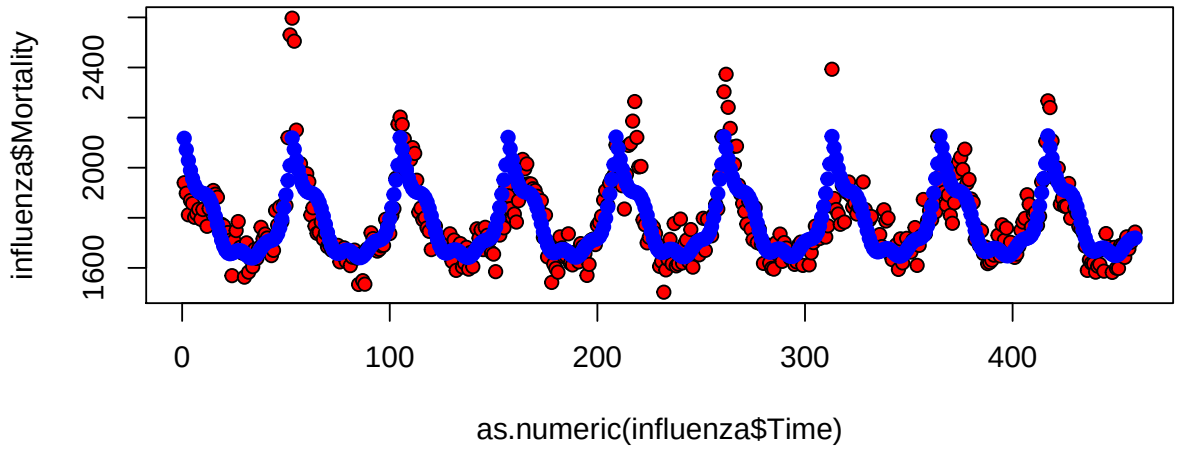
Regarding the graph over time versus influenza it can be seen the values are zero or very close to zero for the majority of the weeks during a year. It is during the coldest weeks of the winter the number of influenza cases rises. How many that is hit by the influenza is varying from year to year.

2.2

Report the probabilistic model...

2.3

The predicted mortality by the fitted model in *2.2*, blue dots, is compared against the observed mortality, red dots.

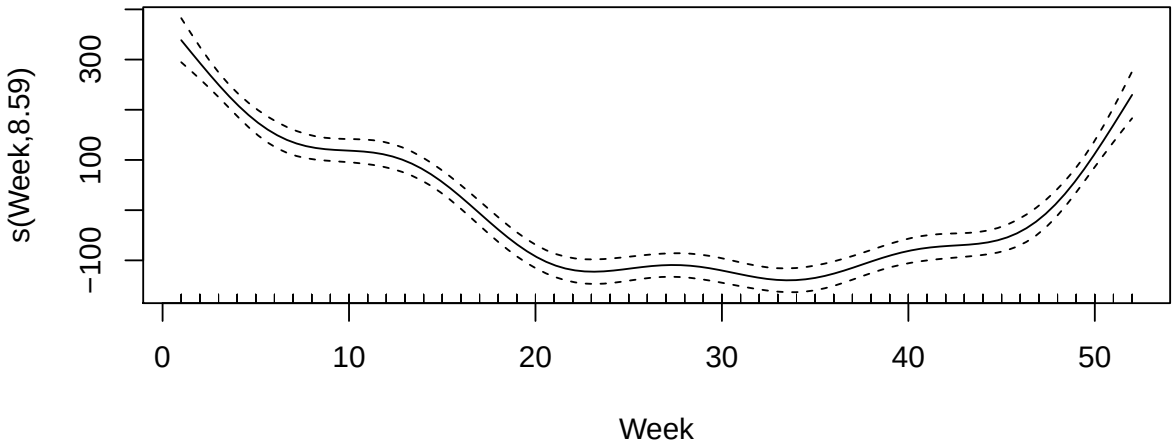


It seems to be a decent fit since the fitted values follows the general pattern for the mortality quite well. (Although it does not capture the variation/more extreme values who changes a bit from year to year.)

Next the output of the GAM model is investigated. According to the output the parametric coefficient for the variable *Year* is insignificant and the non-parametric coefficient for the variable *week* is significant. Hence, there is no significant trend saying that the mortality rate changes from one year to another.

```
##
## 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 spline component for the variable *Week* is visualized with the following plot.

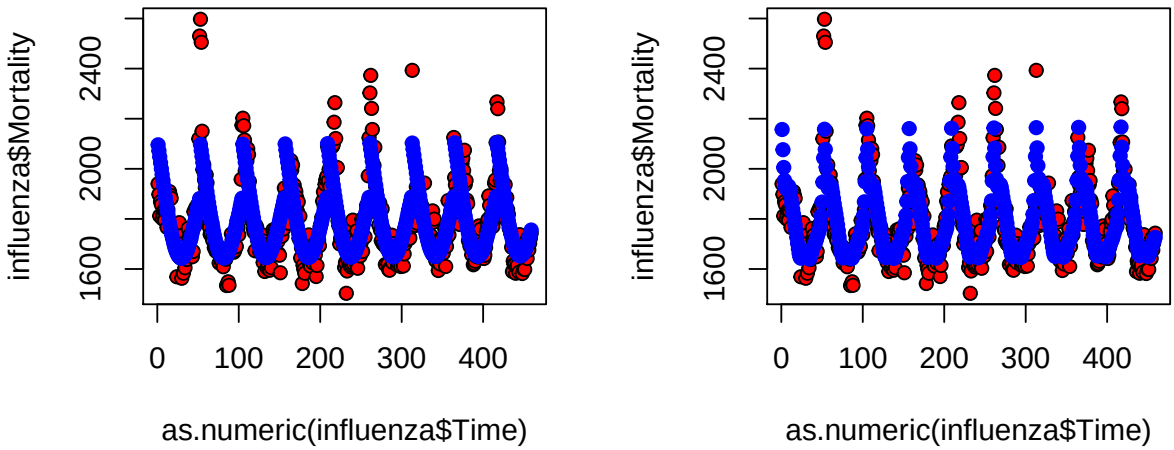


The plot over the spline component shows how this component affects the fitted values given by the model. For weeks at the start and at the end of the year it raises the mortality and during mid-year it lowers the mortality.

The conlcusion about the model fit is that it seem to be a decent model. The comparsion between the observed and fitted values showed that the general pattern is captured. A disadvantage is that one of the variables, the parametric coefficient for *Year*, is insignificant.

2.4

Two different adjustments are done for the model created in *2.2*. In the first case the penalty factor of the spline function(the number of knots) is set to 3 and in the second case it is set to 30. A comparison of the fitted values with the observed values is given by the plots below.



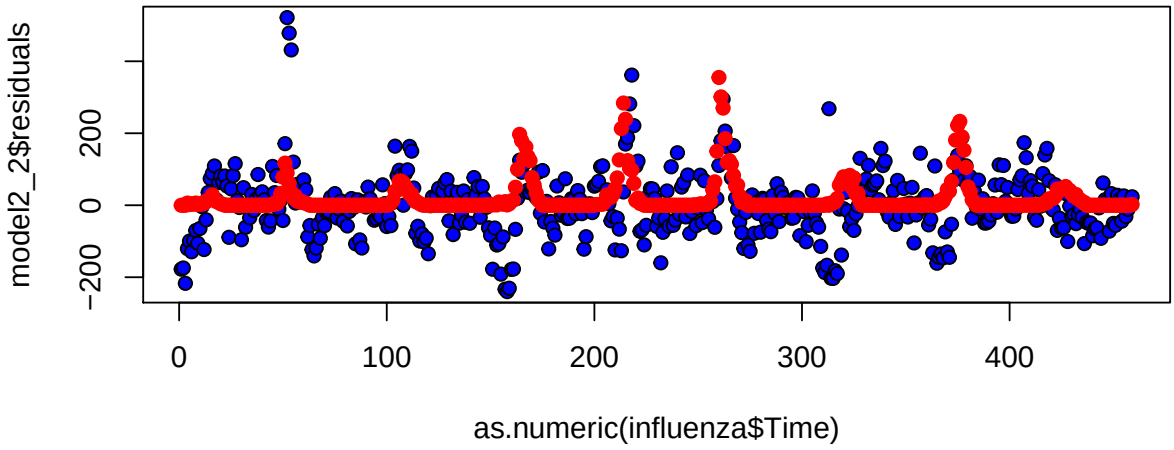
When the penalty factor (or number of knots) is low the fitted values almost looks exactly the same from

year to year. For a higher value of the penalty factor a bit more of the yearly variation is captured.

The relation between the penalty factor and the degrees of freedom...

2.5

The residuals for the model created in *2.2* and the observed influenza values are plotted against time.



The highest values of the residuals seem to be correlated to the periods of outbreaks of influenza(?) Shows well that the model are quite general. For years with either a low or high number of influenza outbreaks the values of the residuals becomes high.

2.6

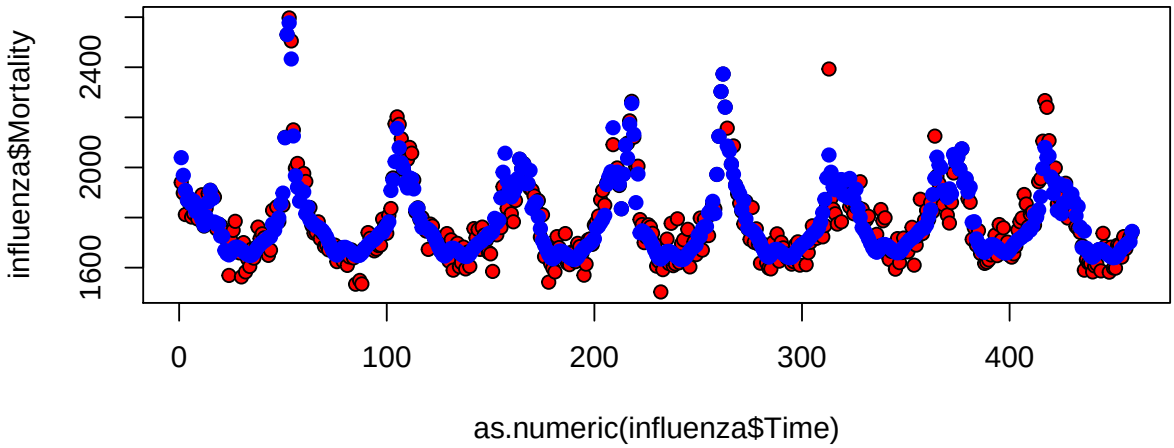
Next, a GAM model is fitted where mortality is described by spline functions of *year*, *week* and the number of confirmed cases of influenza.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ s(Week, k = 52) + s(Year, k = 9) + s(Influenza, k = 85)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1783.8        3.2   557.5   <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)     14.641 18.248 18.533  <2e-16 ***
```

```
## s(Year)       4.663  5.677  1.487   0.181
## s(Influenza) 69.735 72.840  5.601  <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 = 5846.7  Scale est. = 4699.8    n = 459
```

Looks like the mortality is influenced by the outbreaks of influenza(?).

The fitted values for the created model is compared against the observed valuse for mortality.



The model have a hard time predicting right for the lowest mortality values durint the respective years. Overall a more flexible model than the model created in *2.2*.

The model have a hard time predicting right for the lowest mortality values durint the respective years. Overall a more flexible model than the model created in *2.2*. 4.