# Lab2Block2

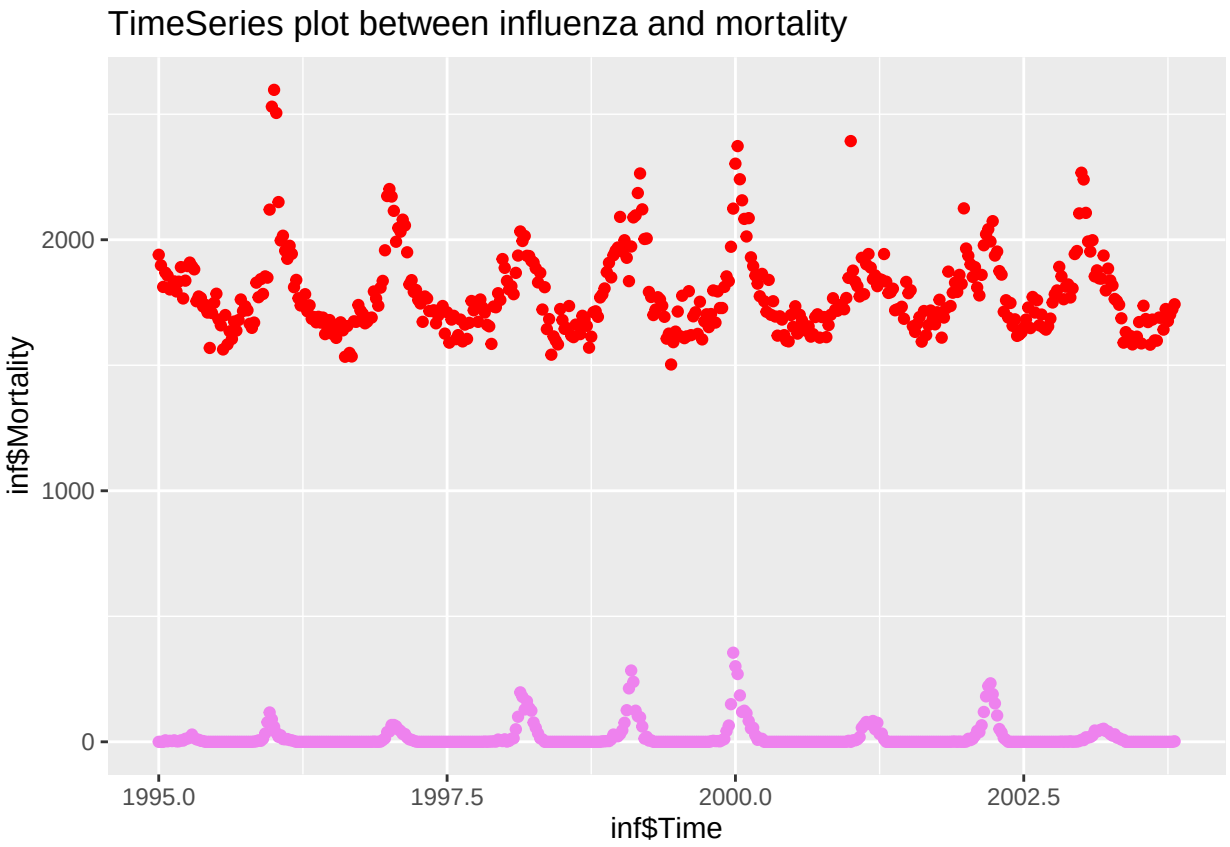
*Sreenand.S*

*10/12/2019*

## Assignment 1

### 1.1:Relation between the mortality rates and the influenza rates

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform
## 'Rounding' sampler used
```

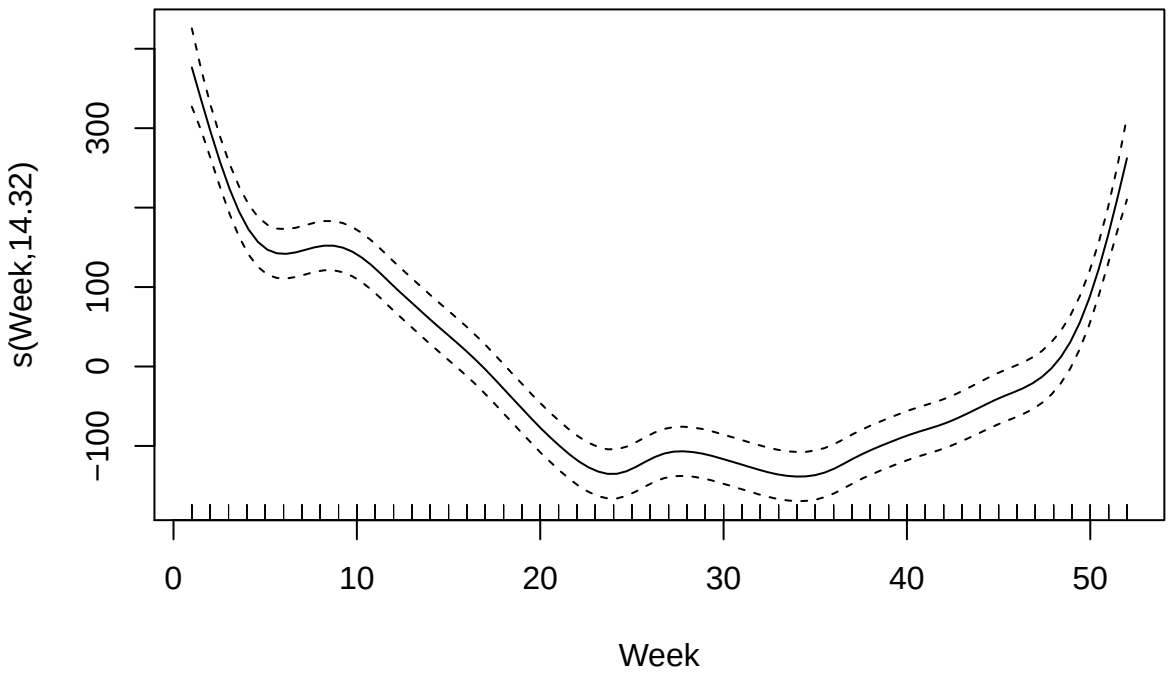



From the graphs above it is observed that there is a slight linear relationship between the mortality rates and the influenza affecting the people. It is a positive relationship.The Mortality seeems to be at a constant rate throughout the year whereas the influenza has a major spike in between.

### 1.2:Gam Function

You can also embed plots, for example:

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform
## 'Rounding' sampler used
```




```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ Year + s(Week, k = length(unique(inf$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
```

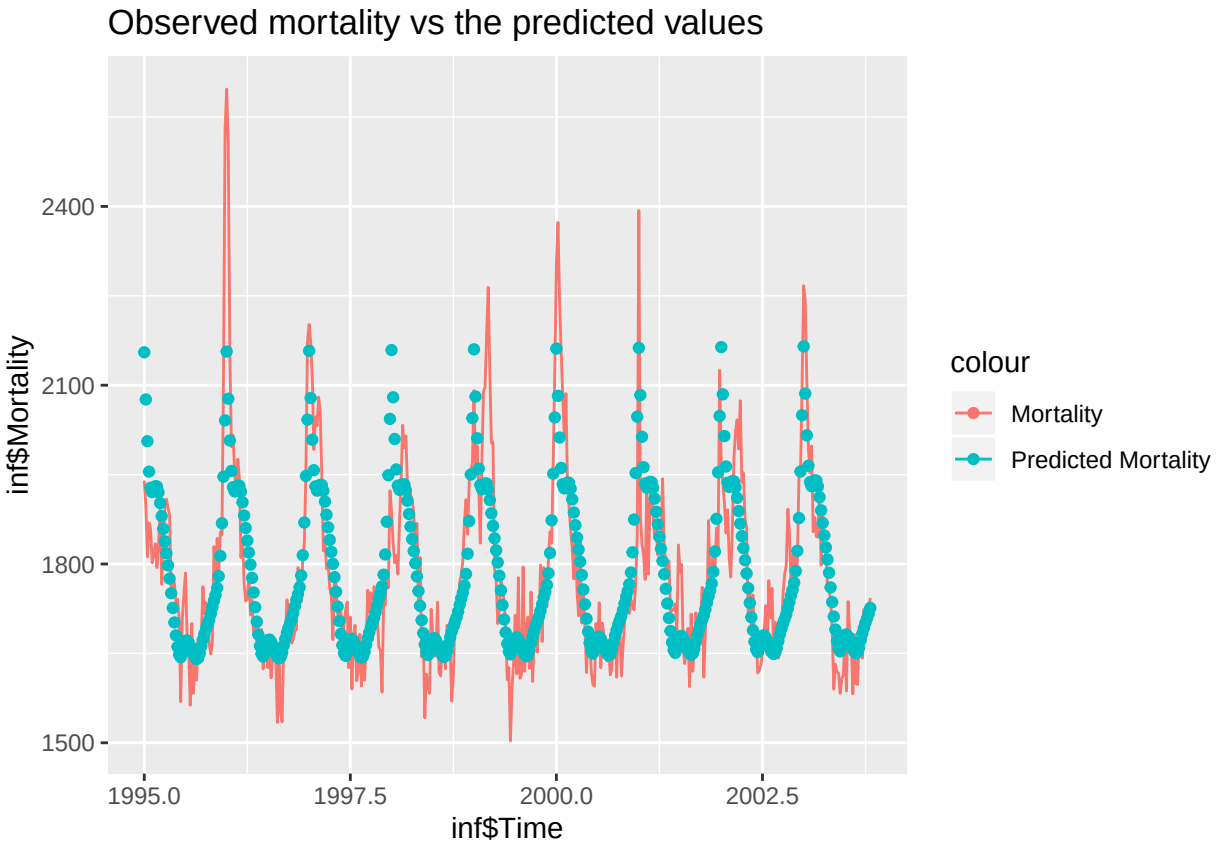
```
## [1] 8398.907
```

The probabilistic model is : Mortality~N(-680.598 + 1.233*Year + spline(week)+std. err

### 1.3:Spline component

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform
## 'Rounding' sampler used
```

```
## Warning in as.data.frame.vector(c(x), row.names, optional, ...): 'row.names' is
## not a character vector of length 459 -- omitting it. Will be an error!
```

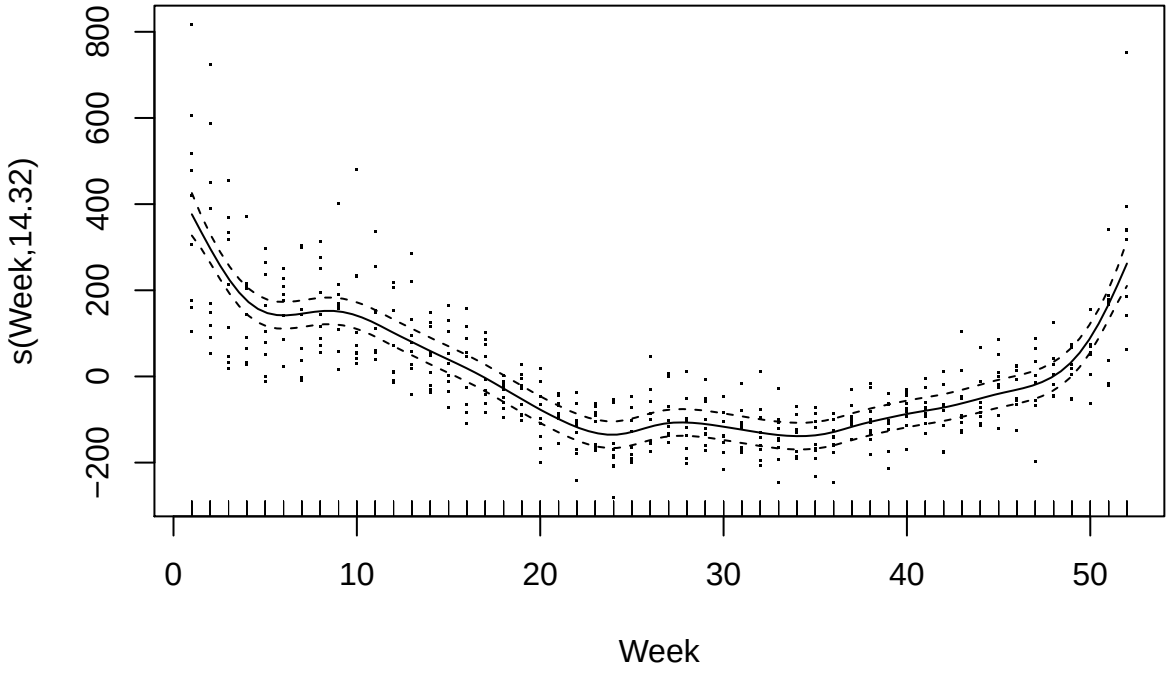



The predicted values are too smooth.The model does not predict any of the data which are above the range of 2200 and the data that are below 1600 from which it can be pointed out that it does not include the outlier values and hence would not be the best fit. It is observed that there is no change in trend over the years and hence the same trend can be observed throughout the years.

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform
## 'Rounding' sampler used
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ Year + s(Week, k = length(unique(inf$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
```




It is observed fron the graph that the rate of mortality is higher in the start of the year and reduces in between and towards the end of the years it peaks again. The inference can be made that in the cold months the rate of mortality is higher during the colder months and it less common when it is sunny. From the model summary it is seen that the p-value of the year is much higher than that of the week and hence it can be inferred that the Year is more significant.

### 1.4:Penalty factor of the spline function

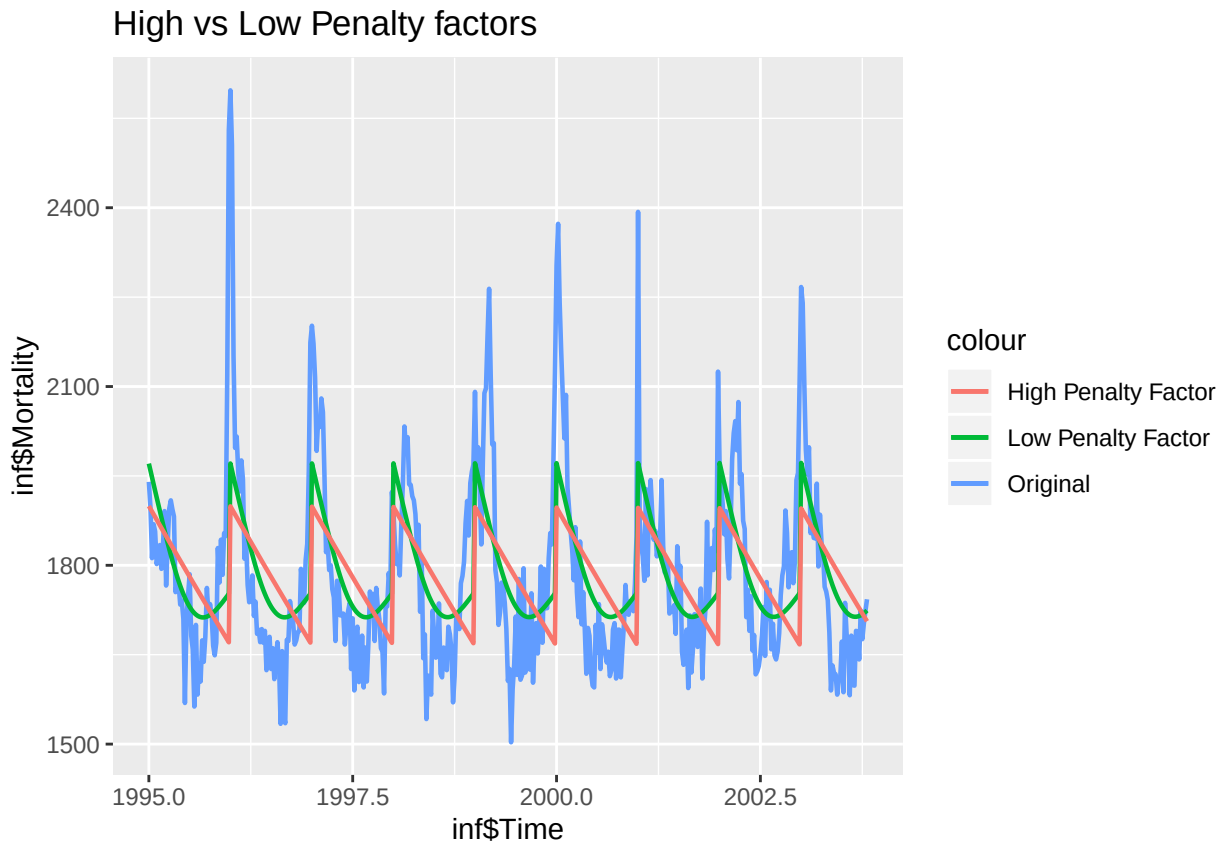
```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform
## 'Rounding' sampler used
```

```
##
##  The sp of the model fitted with the sp of the previous model is 0.0001131932 and the deviance is  3
```

```
## named numeric(0)
```

When the penalty factor of the first model is fit to the new model and when the standard deviation is tested there seems to be no change in the deviance from the first and the second model.

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform
## 'Rounding' sampler used
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ Year + s(Week, k = length(unique(inf$Week)))
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1513.2065  4416.6973   0.343    0.732
## Year           0.1354     2.2095   0.061    0.951
##
## Approximate significance of smooth terms:
##           edf Ref.df     F p-value
## s(Week) 1.492  1.833 179.5  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.443   Deviance explained = 44.6%
## GCV =  14587  Scale est. = 14476     n = 459
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ Year + s(Week, k = length(unique(inf$Week)))
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2789.6277  5381.0481   0.518    0.604
## Year          -0.5032     2.6920  -0.187    0.852
##
## Approximate significance of smooth terms:
##           edf Ref.df     F p-value
## s(Week) 1.007  1.014 95.05  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.172   Deviance explained = 17.6%
## GCV =  21643  Scale est. = 21501     n = 459
## Warning in as.data.frame.vector(c(x), row.names, optional, ...): 'row.names' is
## not a character vector of length 459 -- omitting it. Will be an error!
## Warning in if (!optional) names(value) <- nm: the condition has length > 1 and
## only the first element will be used
## Warning in if (!optional) names(value) <- deparse(substitute(x))[[1L]]: the
## condition has length > 1 and only the first element will be used
```




```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ Year + s(Week, k = length(unique(inf$Week)))
##
## Estimated degrees of freedom:
## 1.49  total = 3.49
##
## GCV score: 14586.8
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ Year + s(Week, k = length(unique(inf$Week)))
##
## Estimated degrees of freedom:
## 1.01  total = 3.01
##
## GCV score: 21642.73
```

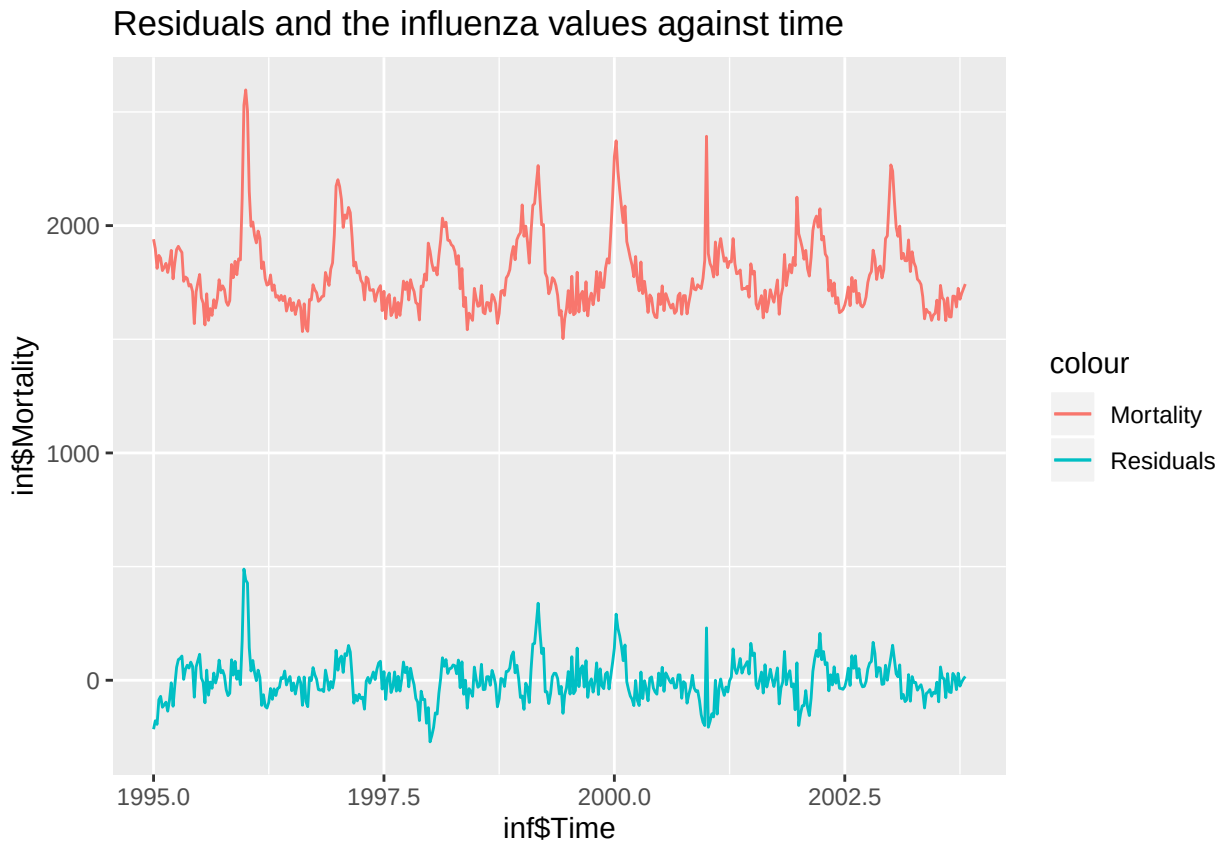
```
## [1] 6593860
```

```
## [1] 9804267
```

It is observed that with the low penalty factor there is a lower deviance and with a high penalty factor there is increase in the deviance level and the when the spline fucntion is used it gives the optimal sp value. When the sp value is high it causes underfitting and when it is low it causes overfitting. The degrees of freedom is observed to reduce with the increase in the sp value.This confirms the relationship df is inversely propotional to the lambda value.

### 1.5:residuals and the influenza values against time

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform
## 'Rounding' sampler used
```



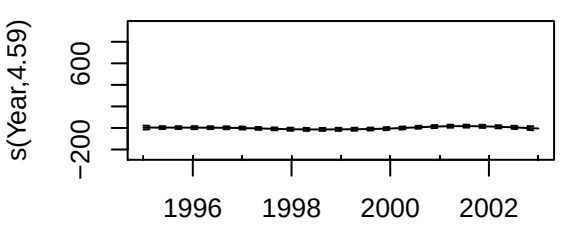

Temporal pattern in the residuals seems to be correlated to the outbreaks of influenza.The spikes in the influenza is being predicted.

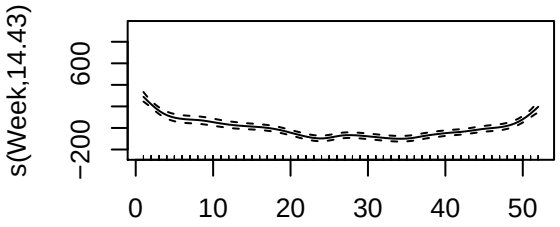
### 1.6:Mortality modelled as an additive function

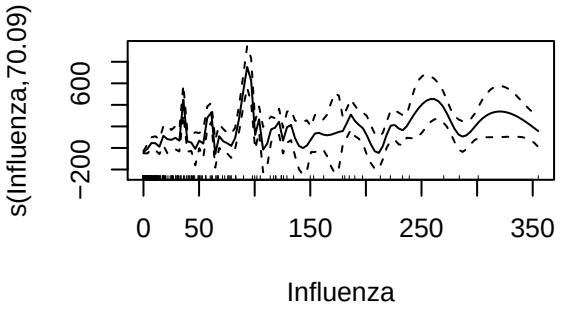
```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform
## 'Rounding' sampler used
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Mortality ~ s(Year, k = length(unique(inf$Year))) + s(Week, k = length(unique(inf$Week))) +
##     s(Influenza, k = length(unique((inf$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
```

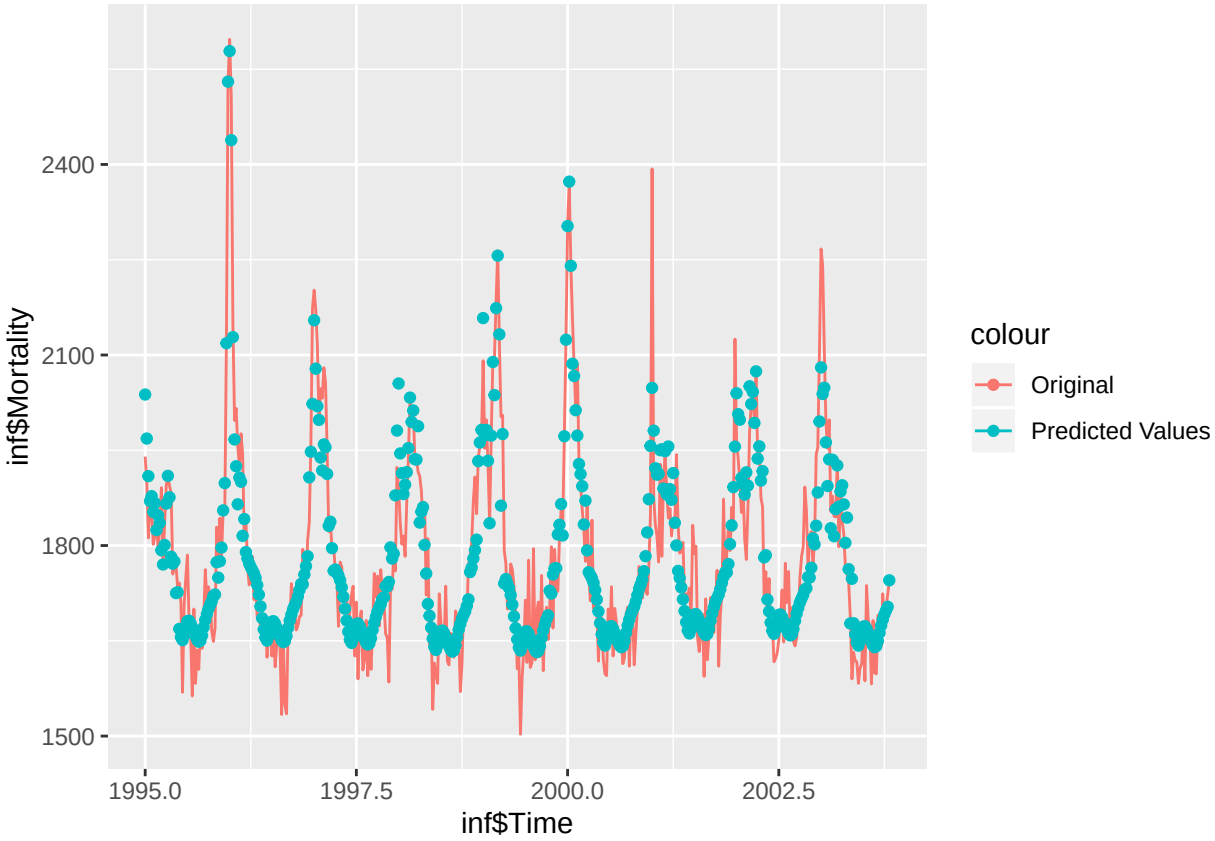



The plot gives a comprehensive analysis of the rates of mortality vs the time, influenza rates and also the weeks . It is observeed that there is hardly any deviance in when plotted with the year and when plotted agaisnt the week it seems to spike only on the onset of the year and toward the end of the year.From the graphs it can be determined that there is a very good relationship woth mortality and Influenza.whereas there is a very poor relationship between the mortality and week and an even worse relationship between the mortlity and the year.The spikes in the moratlity values are correlating to that of the influenza values.

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform
## 'Rounding' sampler used
```

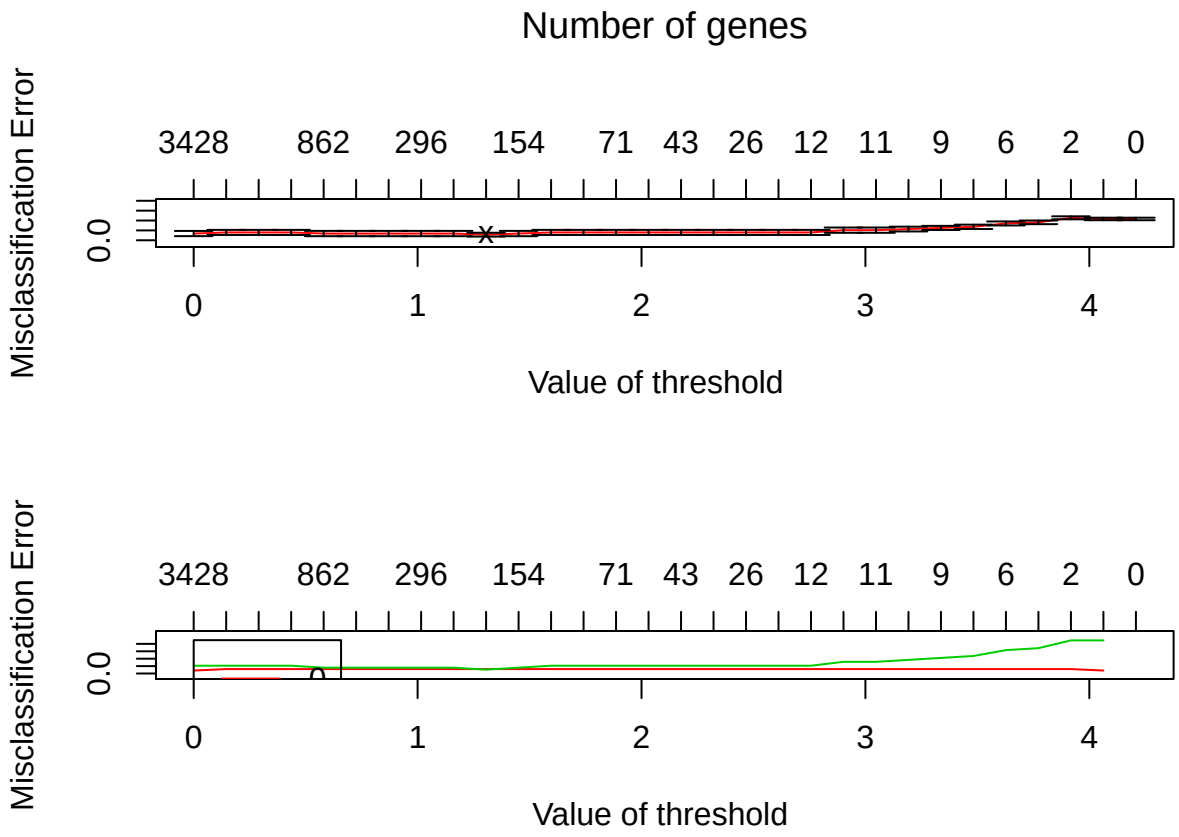



Here in this graph it seems to predict the values very close to the originl values with the minimal error.This shows that the mortality is influenced by the outbreaks of influenza and it has a high impact to the prediction of the mortality.

## Assignment-2

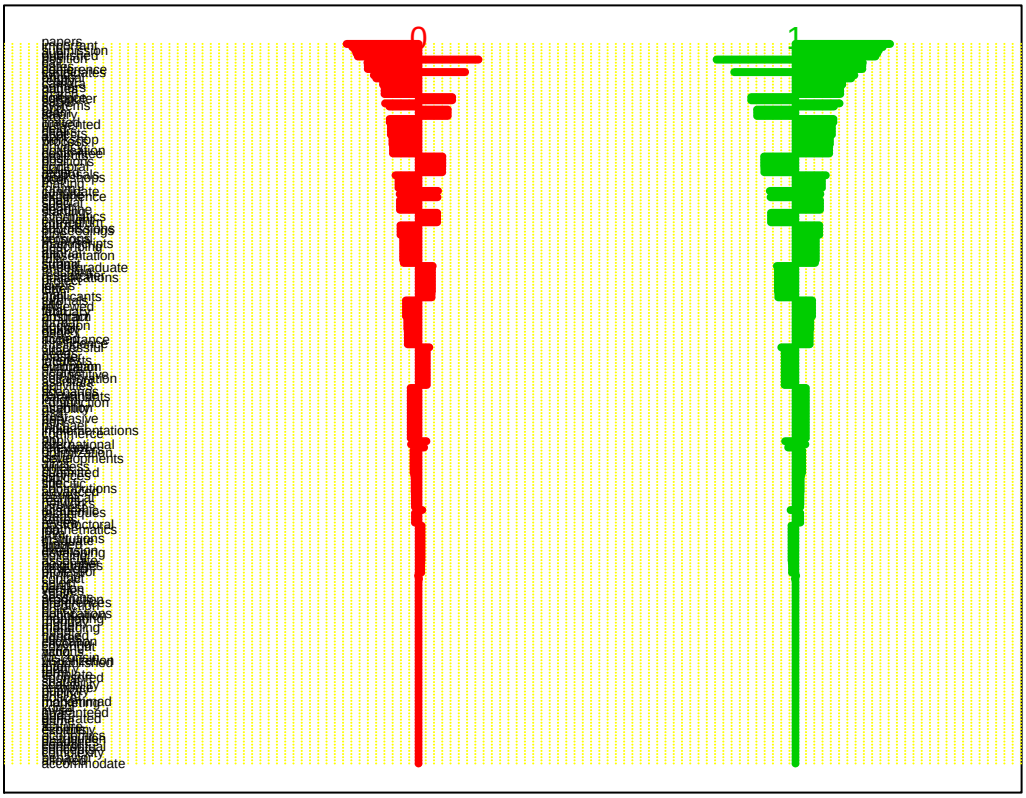
### 2.1

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform
## 'Rounding' sampler used

## 12345678910111213141516171819202122232425262728293O

## 12Fold 1 :12345678910111213141516171819202122232425262728293O
## Fold 2 :12345678910111213141516171819202122232425262728293O
## Fold 3 :12345678910111213141516171819202122232425262728293O
## Fold 4 :12345678910111213141516171819202122232425262728293O
## Fold 5 :12345678910111213141516171819202122232425262728293O
## Fold 6 :12345678910111213141516171819202122232425262728293O
## Fold 7 :12345678910111213141516171819202122232425262728293O
## Fold 8 :12345678910111213141516171819202122232425262728293O
## Fold 9 :12345678910111213141516171819202122232425262728293O
## Fold 10 :12345678910111213141516171819202122232425262728293O
```




```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform
## 'Rounding' sampler used
```

```
## 1
```




After applying the minimum threshold it can be observed that features on the top contribute more than the faetures towards the bottom.

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform 'Rounding' sampler used
```

```
##        id   name            0-score 1-score
##   [1,] 3036 papers          -0.3814 0.5019
##   [2,] 2049 important       -0.3519 0.4631
##   [3,] 4060 submission      -0.3368 0.4431
##   [4,] 1262 due             -0.3301 0.4344
##   [5,] 3364 published       -0.3223 0.4241
##   [6,] 3187 position        0.318   -0.4184
##   [7,] 596  call            -0.2717 0.3575
##   [8,] 869  conference      -0.2698 0.355
##   [9,] 1045 dates           -0.2698 0.355
##  [10,] 607  candidates      0.2468  -0.3247
##  [11,] 4282 topics          -0.2376 0.3126
##  [12,] 2990 original        -0.2246 0.2956
##  [13,] 599  camera          -0.1889 0.2485
##  [14,] 3433 ready           -0.1889 0.2485
##  [15,] 389  authors         -0.1808 0.2378
##  [16,] 2588 march           -0.1808 0.2378
##  [17,] 3022 pages           -0.1808 0.2378
##  [18,] 850  computer        0.1785  -0.2349
##  [19,] 3725 science         0.1785  -0.2349
##  [20,] 3035 paper           -0.1777 0.2338
##  [21,] 4129 systems         -0.1551 0.204
```

```
##  [22,] 3125 phd            0.1551  -0.204
##  [23,] 4177 team           0.1548  -0.2037
##  [24,] 3671 salary         0.1548  -0.2037
##  [25,] 2974 org            -0.1534 0.2018
##  [26,] 2463 limited        -0.1534 0.2018
##  [27,] 329  aspects        -0.1472 0.1937
##  [28,] 681  chairs         -0.1472 0.1937
##  [29,] 1891 held           -0.1472 0.1937
##  [30,] 3243 presented      -0.1472 0.1937
##  [31,] 283  april          -0.1392 0.1832
##  [32,] 4628 workshop       -0.1392 0.1832
##  [33,] 3286 process        -0.1392 0.1832
##  [34,] 3274 privacy        -0.136  0.179
##  [35,] 810  committee      -0.136  0.179
##  [36,] 2889 notification   -0.136  0.179
##  [37,] 1233 doctoral       0.1264  -0.1664
##  [38,] 3188 positions      0.1264  -0.1664
##  [39,] 3191 post           0.1264  -0.1664
##  [40,] 3312 projects       0.1264  -0.1664
##  [41,] 3891 skills         0.1257  -0.1654
##  [42,] 3458 record         0.1257  -0.1654
##  [43,] 3324 proposals      -0.1224 0.161
##  [44,] 1643 forum          -0.1069 0.1407
##  [45,] 2561 making         -0.1069 0.1407
##  [46,] 3090 peer           -0.1069 0.1407
##  [47,] 4629 workshops      -0.1069 0.1407
##  [48,] 606  candidate      0.1034  -0.136
##  [49,] 2058 include        -0.1004 0.1322
##  [50,] 1501 experience     0.1004  -0.1322
##  [51,] 680  chair          -0.0992 0.1306
##  [52,] 3952 special        -0.0992 0.1306
##  [53,] 3836 short          -0.0992 0.1306
##  [54,] 1061 deadline       -0.0991 0.1303
##  [55,] 1007 curriculum     0.0987  -0.1299
##  [56,] 1477 excellent      0.0987  -0.1299
##  [57,] 2103 informatics    0.0987  -0.1299
##  [58,] 3992 starting       0.0987  -0.1299
##  [59,] 2295 journal        -0.0964 0.1268
##  [60,] 4061 submissions    -0.0964 0.1268
##  [61,] 2305 june           -0.0962 0.1266
##  [62,] 3285 proceedings    -0.0962 0.1266
##  [63,] 92   acm            -0.0824 0.1084
##  [64,] 1127 describing     -0.0824 0.1084
##  [65,] 2583 manuscripts    -0.0824 0.1084
##  [66,] 3323 proposal       -0.0824 0.1084
##  [67,] 4500 versions       -0.0824 0.1084
##  [68,] 1698 fully          -0.0824 0.1084
##  [69,] 3241 presentation   -0.0824 0.1084
##  [70,] 4364 tutorial       -0.0824 0.1084
##  [71,] 4062 submit         -0.0788 0.1037
##  [72,] 4039 strong         0.0749  -0.0986
##  [73,] 740  city           0.0714  -0.0939
##  [74,] 2438 letter         0.0714  -0.0939
##  [75,] 2442 levels         0.0714  -0.0939
```

```
##  [76,] 3311 project         0.0714  -0.0939
##  [77,] 3383 qualifications  0.0714  -0.0939
##  [78,] 3559 researcher      0.0714  -0.0939
##  [79,] 4176 teaching        0.0714  -0.0939
##  [80,] 4402 undergraduate   0.0714  -0.0939
##  [81,] 267  applicants      0.0693  -0.0912
##  [82,] 2553 mail            0.0693  -0.0912
##  [83,] 63   abstract        -0.0673 0.0885
##  [84,] 1563 february        -0.0673 0.0885
##  [85,] 1594 final           -0.0673 0.0885
##  [86,] 3589 reviewed        -0.0673 0.0885
##  [87,] 3882 site            -0.0673 0.0885
##  [88,] 4365 tutorials       -0.0673 0.0885
##  [89,] 3301 program         -0.0604 0.0795
##  [90,] 1636 format          -0.0602 0.0792
##  [91,] 1072 decision        -0.0602 0.0792
##  [92,] 386  author          -0.0602 0.0792
##  [93,] 2198 invited         -0.0579 0.0761
##  [94,] 3021 page            -0.0579 0.0761
##  [95,] 3386 quality         -0.0579 0.0761
##  [96,] 76   acceptance      -0.0575 0.0757
##  [97,] 2150 intelligence    -0.0575 0.0757
##  [98,] 4075 successful      0.0571  -0.0752
##  [99,] 107  activities      0.044   -0.0579
## [100,] 336  assistant       0.044   -0.0579
## [101,] 776  collaboration   0.044   -0.0579
## [102,] 831  competitive     0.044   -0.0579
## [103,] 1088 degree          0.044   -0.0579
## [104,] 1450 european        0.044   -0.0579
## [105,] 1456 evaluation      0.044   -0.0579
## [106,] 1542 faculty         0.044   -0.0579
## [107,] 2170 interests       0.044   -0.0579
## [108,] 2613 master          0.044   -0.0579
## [109,] 2837 needs           0.044   -0.0579
## [110,] 4529 vitae           0.044   -0.0579
## [111,] 363  attention       -0.0421 0.0554
## [112,] 879  conjunction     -0.0421 0.0554
## [113,] 2433 length          -0.0421 0.0554
## [114,] 3051 participants    -0.0421 0.0554
## [115,] 3514 relevance       -0.0421 0.0554
## [116,] 3711 scenarios       -0.0421 0.0554
## [117,] 4449 url             -0.0421 0.0554
## [118,] 501  bio             -0.0421 0.0554
## [119,] 803  commerce        -0.0421 0.0554
## [120,] 2046 implementations -0.0421 0.0554
## [121,] 2082 india           -0.0421 0.0554
## [122,] 2690 michael         -0.0421 0.0554
## [123,] 2877 non             -0.0421 0.0554
## [124,] 3118 pervasive       -0.0421 0.0554
## [125,] 4342 trust           -0.0421 0.0554
## [126,] 4451 usa             -0.0421 0.0554
## [127,] 4452 usability       -0.0421 0.0554
## [128,] 272  apply           0.0417  -0.0549
## [129,] 2175 international   -0.04   0.0526
```

```
## [130,] 3515 relevant        0.0294  -0.0387
## [131,] 172  aims            -0.0276 0.0363
## [132,] 1149 developments    -0.0276 0.0363
## [133,] 2219 issue           -0.0276 0.0363
## [134,] 2964 optimization    -0.0276 0.0363
## [135,] 2984 organizers      -0.0276 0.0363
## [136,] 2887 notes           -0.0276 0.0363
## [137,] 4605 wireless        -0.0276 0.0363
## [138,] 4064 submitted       -0.0272 0.0358
## [139,] 3800 services        -0.023  0.0302
## [140,] 134  advanced        -0.0214 0.0282
## [141,] 919  contributions   -0.0214 0.0282
## [142,] 3957 specific        -0.0214 0.0282
## [143,] 4268 title           -0.0214 0.0282
## [144,] 4281 topic           -0.0214 0.0282
## [145,] 2220 issues          -0.0203 0.0267
## [146,] 2847 networks        -0.0203 0.0267
## [147,] 3582 results         -0.0203 0.0267
## [148,] 4181 technical       -0.0203 0.0267
## [149,] 2167 interest        -0.0196 0.0258
## [150,] 67   academic        0.0196  -0.0258
## [151,] 2005 ideas           -0.0195 0.0257
## [152,] 4185 techniques      -0.0195 0.0257
## [153,] 3588 review          -0.0195 0.0257
## [154,] 3794 series          -0.0195 0.0257
## [155,] 579  building        0.0162  -0.0213
## [156,] 1147 developing      0.0162  -0.0213
## [157,] 1524 extension       0.0162  -0.0213
## [158,] 1591 filled          0.0162  -0.0213
## [159,] 1702 funded          0.0162  -0.0213
## [160,] 1797 graduate        0.0162  -0.0213
## [161,] 2141 institutions    0.0162  -0.0213
## [162,] 2251 java            0.0162  -0.0213
## [163,] 2278 job             0.0162  -0.0213
## [164,] 2619 mathematics     0.0162  -0.0213
## [165,] 3194 postdoctoral    0.0162  -0.0213
## [166,] 340  associate       0.0162  -0.0213
## [167,] 2894 november        0.0148  -0.0194
## [168,] 1144 develop         0.0141  -0.0185
## [169,] 2392 languages       0.0141  -0.0185
## [170,] 3295 professor       0.0141  -0.0185
## [171,] 2713 mining          -0.0028 0.0037
## [172,] 899  contact         0.0024  -0.0032
## [173,] 1859 hand            -4e-04  5e-04
## [174,] 3764 select          -4e-04  5e-04
## [175,] 104  actions         -4e-04  5e-04
## [176,] 940  copyright       -4e-04  5e-04
## [177,] 967  covering        -4e-04  5e-04
## [178,] 1343 elicitation     -4e-04  5e-04
## [179,] 1587 figures         -4e-04  5e-04
## [180,] 1861 handled         -4e-04  5e-04
## [181,] 1965 huge            -4e-04  5e-04
## [182,] 2574 managing        -4e-04  5e-04
## [183,] 2623 matter          -4e-04  5e-04
```

```
## [184,] 2754 monday          -4e-04  5e-04
## [185,] 2757 monitoring      -4e-04  5e-04
## [186,] 2839 negotiation     -4e-04  5e-04
## [187,] 2890 notifications   -4e-04  5e-04
## [188,] 3169 policy          -4e-04  5e-04
## [189,] 3224 prediction      -4e-04  5e-04
## [190,] 3231 preferences     -4e-04  5e-04
## [191,] 3289 production      -4e-04  5e-04
## [192,] 3802 sessions        -4e-04  5e-04
## [193,] 3943 spain           -4e-04  5e-04
## [194,] 4490 venues          -4e-04  5e-04
## [195,] 4499 version         -4e-04  5e-04
## [196,] 84   accommodate     -4e-04  5e-04
## [197,] 196  allowed         -4e-04  5e-04
## [198,] 455  behavior        -4e-04  5e-04
## [199,] 837  complexity      -4e-04  5e-04
## [200,] 856  concepts        -4e-04  5e-04
## [201,] 857  conceptual      -4e-04  5e-04
## [202,] 920  control         -4e-04  5e-04
## [203,] 1062 deadlines       -4e-04  5e-04
## [204,] 1214 distribution    -4e-04  5e-04
## [205,] 1291 economics       -4e-04  5e-04
## [206,] 1292 economy         -4e-04  5e-04
## [207,] 1490 exhibits        -4e-04  5e-04
## [208,] 1560 feature         -4e-04  5e-04
## [209,] 1721 game            -4e-04  5e-04
## [210,] 1745 generated       -4e-04  5e-04
## [211,] 1818 grid            -4e-04  5e-04
## [212,] 1833 guaranteed      -4e-04  5e-04
## [213,] 2197 invite          -4e-04  5e-04
## [214,] 2359 korea           -4e-04  5e-04
## [215,] 2598 marketing       -4e-04  5e-04
## [216,] 2746 mohammad        -4e-04  5e-04
## [217,] 2888 notice          -4e-04  5e-04
## [218,] 3259 pricing         -4e-04  5e-04
## [219,] 3361 publicity       -4e-04  5e-04
## [220,] 3570 resource        -4e-04  5e-04
## [221,] 3703 scalability     -4e-04  5e-04
## [222,] 3948 spatial         -4e-04  5e-04
## [223,] 3966 sponsored       -4e-04  5e-04
## [224,] 4202 template        -4e-04  5e-04
## [225,] 4212 term            -4e-04  5e-04
## [226,] 4236 theory          -4e-04  5e-04
## [227,] 4363 tutor           -4e-04  5e-04
## [228,] 4435 unpublished     -4e-04  5e-04
## [229,] 4526 visualization   -4e-04  5e-04
## [230,] 4606 wisconsin       -4e-04  5e-04
## [231,] 4664 yang            -4e-04  5e-04
```

### 1.1-No. of Features Selected

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform 'Rounding' sampler used
```

```
##
## Total Features Selected:  231
```

231 fratures were selected after the cross validation method.

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform 'Rounding' sampler used
```

```
##
## Most contributing featuers are
```

```
##      Features
## 1      papers
## 2   important
## 3  submission
## 4         due
## 5   published
## 6    position
## 7        call
## 8  conference
## 9       dates
## 10 candidates
```

The most contibuting features are printed above.It is reasonable to say that they have strong effect on the discrimination between the conference since they are all specific words.

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform 'Rounding' sampler used
```

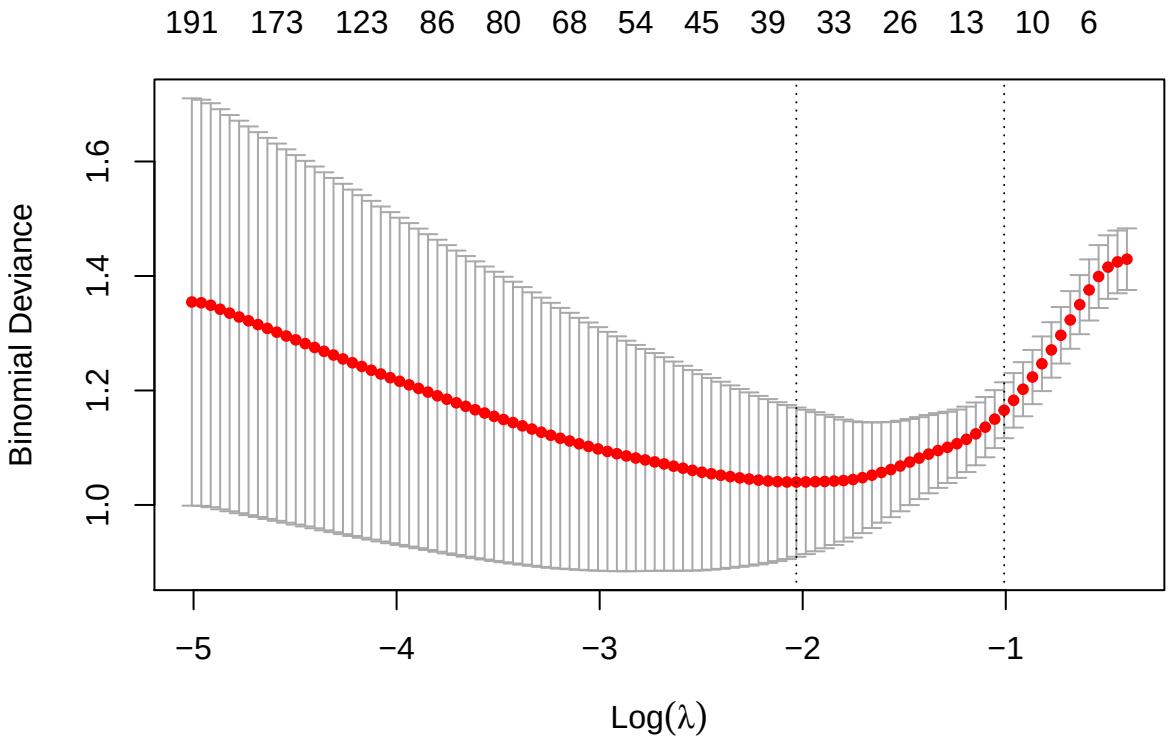
```
## [1] "Test set performance"
```

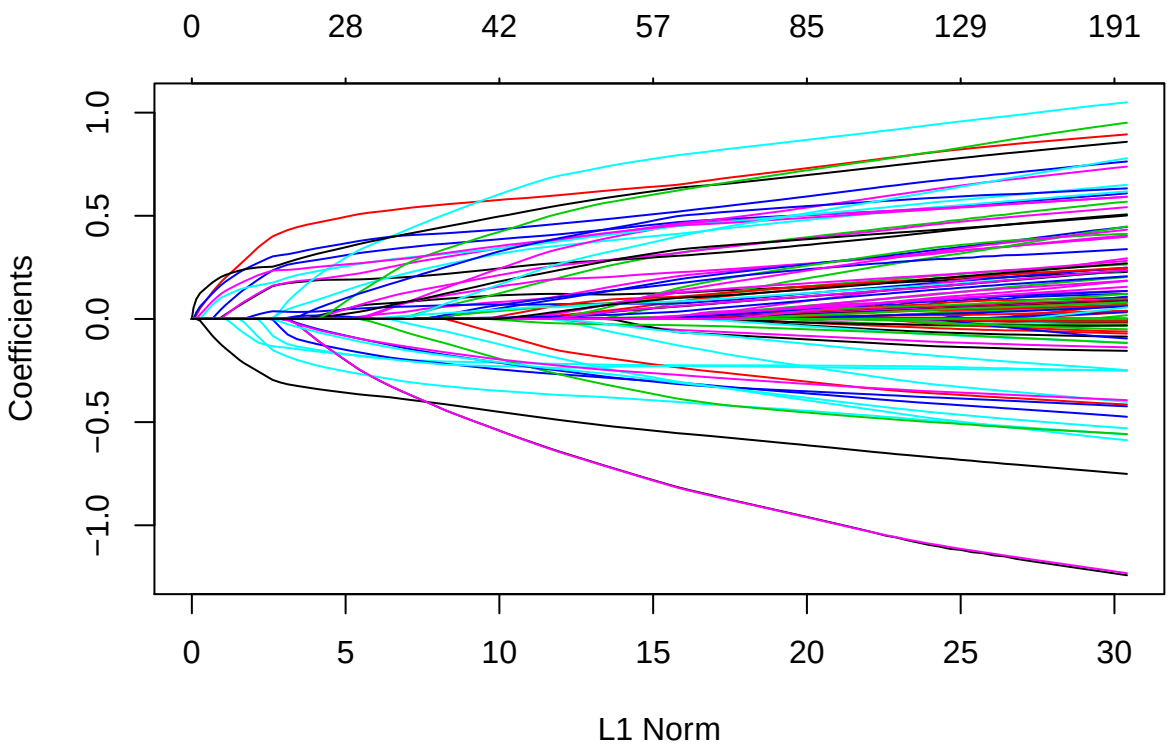
```
##       Shrink_pred
## y_test  0  1
##      0 10  0
##      1  2  8
```

```
##
## Misclassification rate on test:  0.1
```

### 2.2-Part A:Elastic net

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform 'Rounding' sampler used
```


191 173 123 86 80 68 54 45 39 33 26 13 10 6
Binomial Deviance
1.0
1.2
1.4
1.6
−5
−4
−3
−2
−1
Log(λ)




```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform 'Rounding' sampler used
```

```
##        glm_test_pred
## y_test1  0  1
##       0 10  0
##       1  2  8
```

```
##
##  the misclassification rate using the test set was found out to be  0.1
```

```
##
##  The no of features selected are  38
```

**2.2-Part B:Support vector machine with "vanilladot" kernel.**

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform 'Rounding' sampler used
```

```
## Warning in any(scaled): coercing argument of type 'character' to logical
```

```
##  Setting default kernel parameters
```

```
## Warning in any(scaled): coercing argument of type 'character' to logical
```

```
## Warning in any(scaled): coercing argument of type 'character' to logical
```

```
## Support Vector Machine object of class "ksvm"
##
## SV type: C-svc  (classification)
##  parameter : cost C = 1
##
```

```
## Linear (vanilla) kernel function.
##
## Number of Support Vectors : 43
##
## Objective Function Value : -2.0817
## Training error : 0.022727
```

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform 'Rounding' sampler used
```

```
##    kernel_pred
##      0  1
##   0 10  0
##   1  1  9
```

```
## The misclassification rate for the svm is  0.05
```

```
##
## The number of selected features 42
```

**Comparison between all models**

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform 'Rounding' sampler used
```

```
##                 Model MisClassificationRate FeaturesSelected
## 1   Shrinken Centroid                 0.10             231
## 2 Elastic net,Support                 0.10              38
## 3      Vector Machine                 0.05              42
```

**2.3:Benjamini-Hochberg method**

```
## Warning in RNGkind("Mersenne-Twister", "Inversion", "Rounding"): non-uniform 'Rounding' sampler used
```

```
## Total features rejected are : 39 and the features are
```

```
##    p_val[which(l_vec < 0), 2]
## 1                      papers
## 2                  submission
## 3                    position
## 4                   published
## 5                   important
## 6                        call
## 7                  conference
## 8                  candidates
## 9                       dates
## 10                      paper
## 11                     topics
## 12                    limited
## 13                  candidate
## 14                     camera
## 15                      ready
## 16                    authors
## 17                        phd
## 18                   projects
## 19                        org
## 20                     chairs
## 21                        due
## 22                   original
## 23               notification
```

```
## 24                    salary
## 25                    record
## 26                    skills
## 27                      held
## 28                      team
## 29                     pages
## 30                  workshop
## 31                 committee
## 32               proceedings
## 33                     apply
## 34                    strong
## 35             international
## 36                    degree
## 37                 excellent
## 38                      post
## 39                 presented
```

From the results we can conclude that 39 features correspond to the rejected hypothesis.

### Appendix

```
knitr::opts_chunk$set(echo = TRUE)
RNGversion("3.5.1")
library(readxl)
library(mgcv)
library(interp)
library(grid)
library(glmnet)
library(ggplot2)
library(plotly)
library(pamr)
library(caret)
library(kernlab)
inf<-read_xlsx("C:\\Users\\MahmoodS\\Documents\\Machine Learning\\Lab2-Block2\\Influenza.xlsx")

RNGversion("3.5.1")
inf<-read_xlsx("C:\\Users\\MahmoodS\\Documents\\Machine Learning\\Lab2-Block2\\Influenza.xlsx")



ggplot(inf)+geom_point(aes(x= inf$Time,y= inf$Mortality),col = 'RED')+geom_point(aes(x=inf$Time,y= inf$
RNGversion("3.5.1")
Gamfit<- gam(Mortality~Year+s(Week, k=length(unique(inf$Week))),data = inf,method = "GCV.Cp")

plot(Gamfit)
s=interp(inf$Year,inf$Week, fitted(Gamfit))
summary(Gamfit)
Gamfit$sig2
RNGversion("3.5.1")
gampred<- predict(Gamfit,inf)
MvTdf<-as.data.frame(gampred,inf)
ggplot(MvTdf)+geom_line(aes(x=inf$Time,y=inf$Mortality,color = 'Mortality'))+geom_point(aes(x=inf$Time,
```

```
RNGversion("3.5.1")

s=interp(inf$Year,inf$Week, fitted(Gamfit))
# plot_ly(x=~s$x, y=~s$y, z=~s$z, type="surface")
summary(Gamfit)
plot.gam(Gamfit,residuals = TRUE)
RNGversion("3.5.1")
Gamfit1<- gam(Mortality~Year+s(Week, k=length(unique(inf$Week))),sp=Gamfit$sp, data = inf,method = "GCV
cat("\n The sp of the model fitted with the sp of the previous model is",Gamfit$sp , "and the deviance

Gamfit1$sp
RNGversion("3.5.1")
Gamfit3<- gam(Mortality~Year+s(Week, k=length(unique(inf$Week))),sp = 1, data = inf,method = "GCV.Cp")
Gamfit4<- gam(Mortality~Year+s(Week, k=length(unique(inf$Week))),sp = 100, data = inf,method = "GCV.Cp")
summary(Gamfit3)
summary(Gamfit4)

gam_pred3<-predict(Gamfit3,inf)
gam_pred4<-predict(Gamfit4,inf)
HighLowDf<-as.data.frame(gam_pred3,gam_pred4,inf)
ggplot(HighLowDf)+geom_line(aes(x=inf$Time,y=inf$Mortality,color = 'Original'),size=0.8)+geom_line(aes(
print(Gamfit3)
print(Gamfit4)
Gamfit3$deviance
Gamfit4$deviance
RNGversion("3.5.1")
ggplot(MvTdf)+geom_line(aes(x=inf$Time,y=inf$Mortality,color = 'Mortality'))+geom_line(aes(x=inf$Time,y
RNGversion("3.5.1")
Gamfit3<-gam(Mortality~s(Year,k=length(unique(inf$Year)))+s(Week,k=length(unique(inf$Week)))+s(Influenza
summary(Gamfit3)
par(mfrow=c(2,2))
plot.gam(Gamfit3)
RNGversion("3.5.1")
Gampred_3<-predict(Gamfit3,inf)
df<-data.frame(inf$Mortality,inf$Time,Gampred_3)

ggplot(df)+geom_line(aes(x=inf$Time,y=inf$Mortality,color='Original'))+geom_point(aes(x=inf$Time,y=Gampr
RNGversion("3.5.1")
set.seed(12345)
df2<-read.csv2("C:\\Users\\MahmoodS\\Documents\\Machine Learning\\Lab2-Block2\\data.csv",sep = ";")
df2$Conference = as.factor(df2$Conference)
n=dim(df2)[1]
id = sample(1:n,floor(n*0.7))
train = df2[id,]
test = df2[-id,]
rownames(train)<-1:nrow(train)

x_train<-t(as.matrix(train[,-4703]))
y_train<-as.matrix(train[,4703])
mydata<-list(x=x_train,y=y_train, geneid=as.character(1:nrow(x_train)),genenames=rownames(x_train))


mytrain<-pamr.train(mydata)
```

```
mycv <-  pamr.cv(mytrain, mydata,nfold = 10)
pamr.plotcv(mycv)
RNGversion("3.5.1")
mint<- mycv$threshold[which.min(mycv$error)]
mytrain2<-pamr.train(mydata,threshold = mint)
#Centroid Plot
pamr.plotcen(mytrain,mydata,threshold = mint)
RNGversion("3.5.1")
cent_fea <- pamr.listgenes(mytrain, mydata, threshold = mint, genenames=TRUE)

RNGversion("3.5.1")
cat("\nTotal Features Selected: ",dim(cent_fea)[1])
RNGversion("3.5.1")
# featorder<-cent_fea[order(id),]

cat("\nMost contributing featuers are")
Features<-cent_fea[1:10,"name"]
print(as.data.frame(Features))
RNGversion("3.5.1")
#Test error
print("Test set performance")
x_test <- t(as.matrix(test[,-4703]))
y_test <- as.matrix(test[,4703])
Shrink_pred <- pamr.predict(mytrain,newx=x_test,threshold = mint,type="class")
cfMat <- table(y_test,Shrink_pred)
print(cfMat)
misCalc <- 1- sum(diag(cfMat))/sum(cfMat)
cat("\nMisclassification rate on test: ",misCalc)
RNGversion("3.5.1")
set.seed(12345)
x_train1<-as.matrix(train[,-4703])
y_train1<-as.matrix(train$Conference)
glmfit<-glmnet(x=x_train1,y=y_train1,family = "binomial",alpha = 0.5)
gmlfitcv<-cv.glmnet(x=x_train1,y=y_train1,family = "binomial",alpha = 0.5)
plot(gmlfitcv)
plot(glmfit)

RNGversion("3.5.1")
set.seed(12345)
x_test1<-as.matrix(test[,-4703])
y_test1<-as.matrix(test$Conference)
glm_test_pred<-predict(glmfit,x_test1,s=gmlfitcv$lambda.min,type = 'class')
cfmat2<-table(y_test1,glm_test_pred)
print(cfmat2)
misCalc1<-1-sum(diag(cfmat2))/sum(cfmat2)
glm_fea<-coef(gmlfitcv,s = "lambda.min")
glm_fea<-as.matrix(glm_fea)
index<-which(glm_fea != 0)
glm_feat<-as.matrix(glm_fea[index,])
feat_length<-length(glm_feat)
cat("\n the misclassification rate using the test set was found out to be ",misCalc1)
cat("\n The no of features selected are ",feat_length-1)
```

```
RNGversion("3.5.1")
set.seed(12345)
kernel_fit<-ksvm(Conference~.,train,kernel = "vanilladot",scaled = 'FALSE')
print(kernel_fit)
RNGversion("3.5.1")
kernel_pred<-predict(kernel_fit,test,type = 'response')
cfMat3<-table(test$Conference,kernel_pred)
cfMat3
misCalc2<-1 - sum(diag(cfMat3))/sum(cfMat3)
cat("The misclassification rate for the svm is ",misCalc2)
cat("\nThe number of selected features",length(as.matrix(kernel_fit@coef[[1]]))-1)
RNGversion("3.5.1")
misCalcRates<-c(misCalc,misCalc1,misCalc2)
models<-c("Shrinken Centroid","Elastic net,Support","Vector Machine")
fea_sel<-c(dim(cent_fea)[1],feat_length-1,length(as.matrix(kernel_fit@coef[[1]]))-1)
resultant_df<-data.frame(Model = models,MisClassificationRate = misCalcRates,FeaturesSelected = fea_sel
print(resultant_df)
RNGversion("3.5.1")
p_val<-data.frame(nrow = ncol(df2),ncol = 2)
for (i in 1:4702) {
  x<-df2[,i]
  p_val[i,1]<-t.test(x~Conference,data = df2,alternative = "two.sided")$p.value
  p_val[i,2] <- colnames(df2)[i]
}

#p_matrix<-data.frame('p-values' = p_val,'features' = 1:length(p_val))
#order the p-values
p_val <- p_val[order(p_val[,1]),]
l_vec<-c()
count = 1

for (j in 1:nrow(p_val))
{
  l_vec[j] <- p_val[j,1] - ((0.05 * j)/4702)
}



 retrieve<-max(l_vec)


cat("Total features rejected are :",length(p_val[which(l_vec <0),2]),"and the features are\n")
rejfea=as.data.frame(p_val[which(l_vec <0),2])
rejfea
```