# ISYE 6501 - Week 3 Homework

Ujjawal Madan

28/05/2020

## Contents

### 0.1 Question 7.1

*Describe a situation or problem from your job, everyday life, current events, etc., for which exponential smoothing would be appropriate. What data would you need? Would you expect the value of alpha (the first smoothing parameter) to be closer to 0 or 1, and why?*

An appropriate application for exponential smoothing might be to forecast the number of passengers flying on any given day (or week). Intuitively, we know that the number of passengers is increasing year to year (that would be our trend) and that there would be a greater number or lesser number of passengers depending on the time of the year (seasonality). For example, during Jun - August, or December - January, there might be a greater number of passengers since those tend to be the months during which many people take holiday vacations.

For my exponential smoothing model, I would pick an alpha that is on the higher side so closer to 1. If I have accurately modelled the trend and the seasonality from past data, then the previous data would not have as much relavence in comparison to the past year or two etc.

In this situation, it's important to keep in mind that world events such as 9/11 or the pandemic we are currently experiencing may greatly affect the number of passengers. Such events are quite difficult to predict and so would be very difficult to factor them into our models. It is important to keep in mind that even a finely tuned forecasting model may be completely disconnected from reality if there occur such external events that are beyond our ability to predict.

### 0.2 Question 7.2

*Using the 20 years of daily high temperature data for Atlanta (July through October) from Question 6.2 (file temps.txt), build and use an exponential smoothing model to help make a judgment of whether the unofficial end of summer has gotten later over the 20 years. (Part of the point of this assignment is for you to think about how you might use exponential smoothing to answer this question. Feel free to combine it with other models if you'd like to. There's certainly more than one reasonable approach.)*

The first step to answering this question using exponential smoothing is to import and prepare our data so it can be used by an exponential smoothing function. I will be using the HoltWinters function from the "stats" package

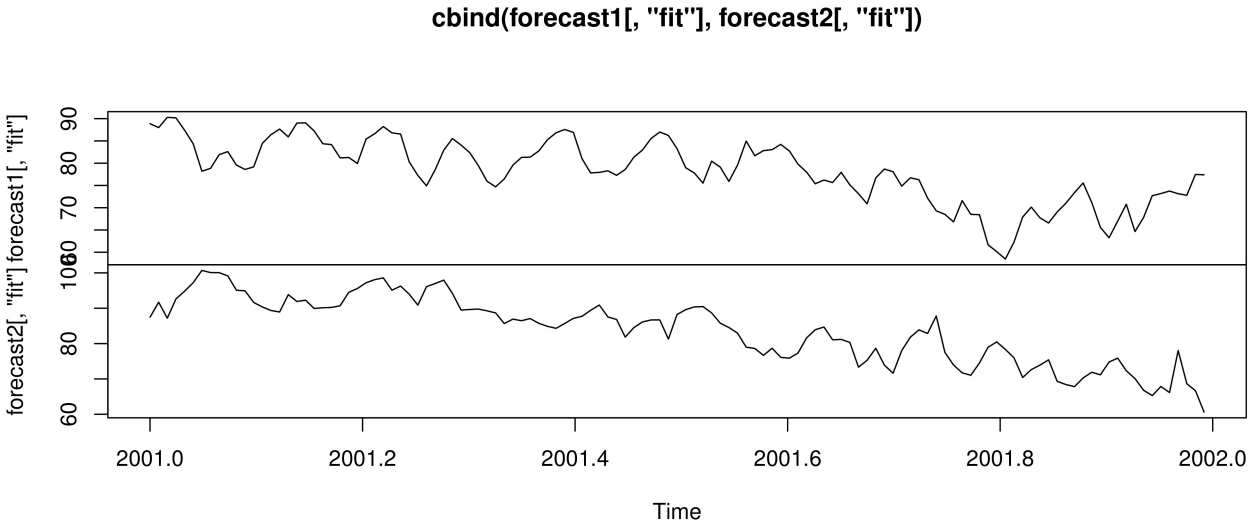
```
#Please change file path
temps <- read_delim("https://raw.githubusercontent.com/ujjawalmadan/ISYE-6501/master/temps.txt",
    "\t", escape_double = FALSE, trim_ws = TRUE)


names <- as.vector(as.character(colnames(temps[,-1])))
temps_agg <- melt(data = temps, id.vars = 'DAY', measure.vars = names)
colnames(temps_agg) <- c('Day', 'Year', 'Temperature')
temps_agg$Day <- as.Date(temps_agg$Day, '%d-%b')
ts1 <- ts(temps_agg$Temperature, frequency=123, start = c(1996, 1))
```

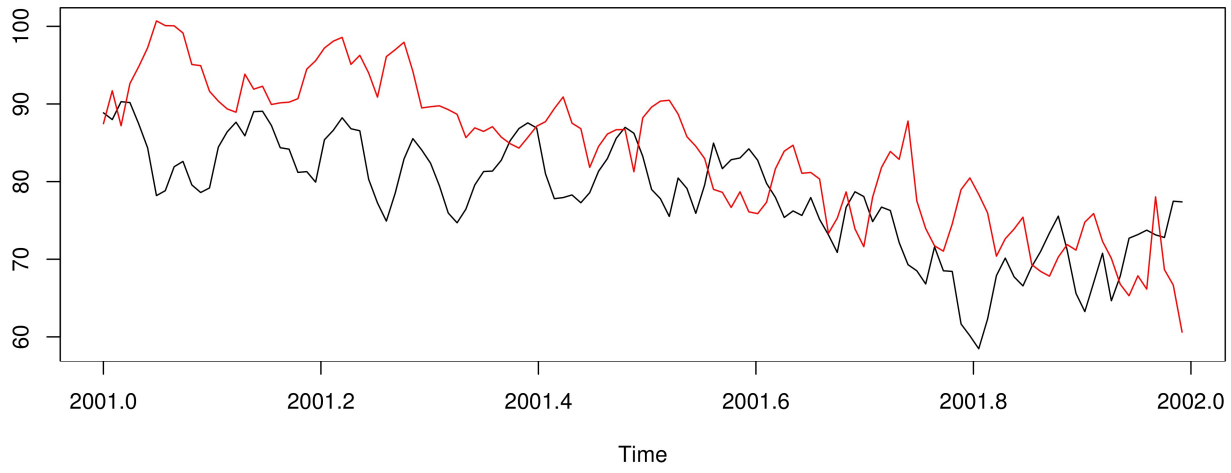
Now that our data has been imported and prepared, let's take a look at how we can use exponential smoothing to help determine whether the unofficial end of summer has gotten later over the past 20 years.

### 0.2.1 Method 1

We compile two groups of data, one containing the temperatures of the first five years of temps, and the one containing the data from the last five years of temps. We forecast temperatures for both sets using exponential smoothing (specifically Holt-Winters) and use CUMUS to determine at what point does a change occuring according to a standardized set of parameters.




The first line on top is of the forecast of the first dataset (first five years) while the second lie is the forecast of the second dataset (last five years).




| Fit 1 SSE | Fit 2 SSE |
|---|---|
| 17918.37 | 18044.39 |

The red line above is the forecast of the later dataset (last five years) and the black line is of the first five years.

I have run these models based on the fact that each period consists of 123 days and that seasonality is additive rather than multiplicative. Previous experimentation revealed that this produced the lowest SSE and intuitively this makes sense. Taking a look at SSE for both fits reveals that they both models are quite similar in terms of overall fit and should be similar enough to each other to make a fair comparison.

It seems that according to the plots, the forecast of temperatures of 2016 (red line) seems to indicate that overall temperatures are higher in the summer in comparison to the forecast of 2001. Please note that the axis of the overlayed plot is set to start from 2001 only for the purpose of being able to visualize both trends on top of one another.

It also appears that the transition from summer to fall appears to happen sooner in the forecast of 2016 rather than the forecast of 2001. However, one could argue that the forecast for 2016 also has higher temperatures in the summer so it could also be possible that the transition actually occurs at the same time, but that summer just seems to be a bit hotter than usual in 2016.

Let us run CUMUS models to more accurately detect whether the changes are ocurring at the same time. The cost is set to 5 and threshold set to 25 which are the correct parameters as noted in Week 2 Homework.

```
temp1 <- as.vector(forecast1[,'fit'])
temp2 <- as.vector(forecast2[,'fit'])
Day <- temps_agg$Day[1:123]

#Function that performs the CUSUM
cusum <- function(data, average, cost, threshold){

  count <- 0
  for (i in 1:length(data)){
    count <- min(data[i] - average + cost + count, 0)
    if (abs(count) > abs(threshold)){
      Day <- Day[i]
      break
    }
  }

  if (i != length(data)){
    return(c(cost, threshold, Day))
  } else{
    return ('No change detected')
  }
}

model1 <- cusum(temp1, mean(temp1[1:31]), 5, 25)[3]
print('Change detection on Forecast of 2001:')
```

```
## [1] "Change detection on Forecast of 2001:"
```

```
print(as.Date(model1, origin = "1970-01-01"))
```

```
## [1] "2020-09-21"
```

```
print('Change detection on Forecast of 2016:')
```

```
## [1] "Change detection on Forecast of 2016:"
```

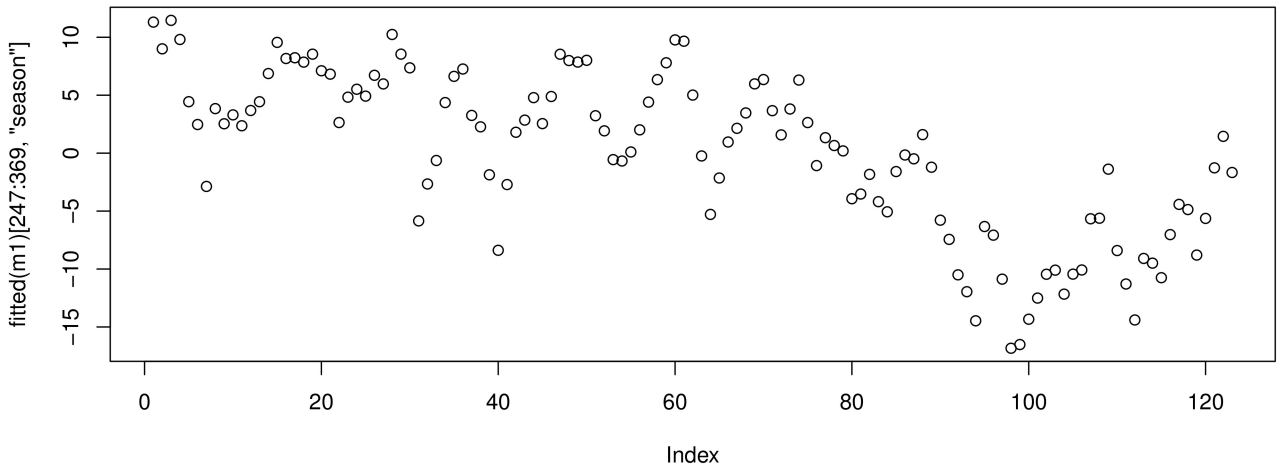
```r
model2 <- cusum(temp2, mean(temp2[1:31]), 5, 25)[3]
print(as.Date(model2, origin = "1970-01-01"))
```

```
## [1] "2020-08-19"
```

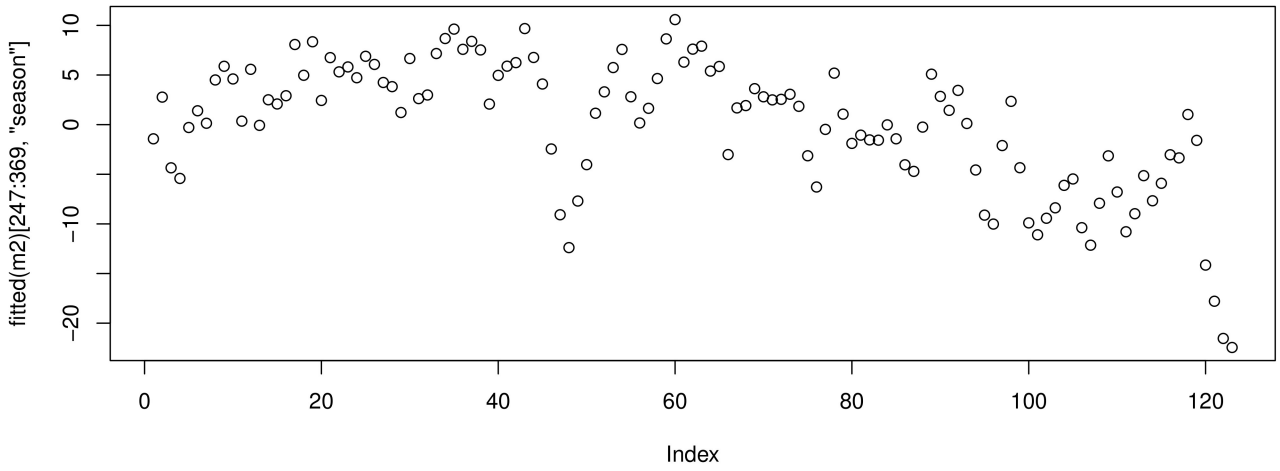
As predicted, the CUMUS, using the same parameters as provided in the answers of Homework Week 2 yields us different dates. As such, one could use this forecast to make an assessment that the transition into fall is ocurring earlier in laters years as compared to previous years.

### 0.2.2 Method 2

We use exponential smoothing to fit a model that incorporates seasonality and run a CUMUS algorithm on the seasonality of year 4 (incorporating data from year 1) and year 20 (incorporating year 17, 18, 19) using a standardized set of parameters to determine whether the change is occurring at different times. This is not a forecast but is rather based on the seasonality of those years as fitted by our models. They are also of different years as compared to method 1, so as to not incorporate year 5 and 20 (2000 and 2015) into these models.




This first plot is of the seasonality, according to our model, of 1999 which contains four years of prior data used to make our forecast. As gamma is set to approximately 0.62, which means it contains a considerable amount of data from the first 3 years.




This second plot is of the seasonality, according to our model, of 2015 which is the the last year. This incorporates the seasonlity of 2012 - 2015 as well.

Let's now run the CUMUS models to see when the change detection occurs. The cost is set to 5 and threshold set to 25.

```
#Year 1999
temp1 <- (fitted(m1)[247:369,'season'])

#Year 2015
temp2 <- (fitted(m2)[247:369,'season'])
Day <- temps_agg$Day[1:123]

model1 <- cusum(temp1, mean(temp1[1:31]), 5, 25)[3]
print('Change detection on seasonality of 1999:')
```

```
## [1] "Change detection on seasonality of 1999:"
```

```
print(as.Date(model1, origin = "1970-01-01"))
```

```
## [1] "2020-09-23"
```

```
print('Change detection on seasonality of 2015: ')
```

```
## [1] "Change detection on seasonality of 2015: "
```

```
model2 <- cusum(temp2, mean(temp2[1:31]), 5, 25)[3]
print(as.Date(model2, origin = "1970-01-01"))
```

```
## [1] "2020-08-19"
```

As we can see, even with a different method the summer is apparently ocurring earlier rather than later. As both methods are still experiments designed to make us explore how exponential smoothing can help us with change detection, I am not in a position to make any conclusions. However, in both experiments, the results told us that summer is indeed ocurring earlier. Whether that is true, not true, or not possible to prove at this time is up to more erudite statisticians.

## 0.3 Question 8.1

*Describe a situation or problem from your job, everyday life, current events, etc., for which a linear regression model would be appropriate. List some (up to 5) predictors that you might use.*

There are inumerable examples of how linear regression might be applicable to data analysis and predictive modeling. One example is predicting a person's salary based on certain attributes. These could include: profession, type of industry, years of experience, location, prior schools attended etc. By compiling enough training data and fitting a linear regression model on it, I believe we would be able to predict the salary of an individual with a relatively high degree of confidence.

## 0.4 Question 9.2

*Using crime data from http://www.statsci.org/data/general/uscrime.txt (file uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html ), use regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the following data:*

*M = 14.0 So = 0 Ed = 10.0 Po1 = 12.0 Po2 = 15.5 LF = 0.640 M.F = 94.0 Pop = 150 NW = 1.1 U1 = 0.120 U2 = 3.6 Wealth = 3200 Ineq = 20.1 Prob = 0.04 Time = 39.0*

*Show your model (factors used and their coefficients), the software output, and the quality of fit.*

Ordinarily, a feature selection process would be performed to better optimize our models. Methods might include PCA or RFE. However, as this is not expected of us at this point nor asked by the assignment, I am assuming I am expected to run all the variables as is. Let us then run a linear model using the lm function.

```
#Linear Model Fit
fit <- lm(Crime ~., data = us_crime)

kable(fit$coefficients)
```

| | x |
|---|---|
| (Intercept) | -5984.2876045 |
| M | 87.8301732 |
| So | -3.8034503 |
| Ed | 188.3243148 |
| Po1 | 192.8043383 |
| Po2 | -109.4219254 |
| LF | -663.8261451 |
| M.F | 17.4068555 |
| Pop | -0.7330081 |
| NW | 4.2044610 |
| U1 | -5827.1027244 |
| U2 | 167.7996722 |
| Wealth | 0.0961662 |
| Ineq | 70.6720995 |
| Prob | -4855.2658155 |
| Time | -3.4790178 |

This table lays out for us the variables and their coefficients of our linear model. Again, an improvement on such a model would be perhaps to consider kernels functions (non-linear functions) as well, but I am going to continue with the linear model for now.

Let us now examine the quality of the linear model. There are several metrics we can utilize: MAE (Mean Absolute Error) which can be used if we are solely interested in seeing what the mean of the residuals is, RMSE (Root Mean Squared Error) which can be used if we are interested penalizing larger errors and R^2 which can help us see how

much of the variance is explained by the linear model.

The first table below tells us the actual value and predicted value according to our linear model for each data point in the us_crimes dataset. The second table gives us performance metric of our linear regression model

| Actual Value | Predicted Value |
|---:|---:|
| 791 | 755.0322 |
| 1635 | 1473.6764 |
| 578 | 322.2615 |
| 1969 | 1791.3619 |
| 1234 | 1166.6840 |
| 682 | 792.9301 |
| 963 | 934.1637 |
| 1555 | 1361.7468 |
| 856 | 688.8682 |
| 705 | 736.5080 |
| 1674 | 1161.3291 |
| 849 | 722.0408 |
| 511 | 732.6412 |
| 664 | 780.0401 |
| 798 | 903.3541 |
| 946 | 1005.6569 |
| 539 | 393.3633 |
| 929 | 843.8072 |
| 750 | 1145.7379 |
| 1225 | 1227.8387 |
| 742 | 774.8506 |
| 439 | 657.2092 |
| 1216 | 957.9918 |
| 968 | 868.9805 |
| 523 | 605.8824 |
| 1993 | 1977.3707 |
| 342 | 279.4772 |
| 1216 | 1258.4842 |
| 1043 | 1287.3917 |
| 696 | 702.6945 |
| 373 | 388.0334 |
| 754 | 807.8167 |
| 1072 | 840.9992 |
| 923 | 971.4558 |
| 653 | 737.7888 |
| 1272 | 1137.6171 |
| 831 | 971.1513 |
| 566 | 562.6934 |
| 826 | 839.2864 |
| 1151 | 1131.4533 |
| 880 | 823.7419 |
| 542 | 326.3324 |
| 823 | 1134.4172 |
| 1030 | 1120.8227 |
| 455 | 616.8983 |
| 508 | 827.3543 |
| 849 | 991.7629 |

| MAE | RMSE | R^2 |
|---|---|---|
| 129.9152 | 169.79 | 0.8030868 |

The second table gives us an idea of the quality of the fit. The R^2 is particularly illuminating as we can see that the linear model explains 80% of the variance which is relatively high.

Let us now make a prediction for the values provided by the question.

| fit | lwr | upr |
|---|---|---|
| 155.4349 | -1310.076 | 1620.946 |

As we can see, the predicted value is 155 although the confidence interval is quite large. As such, using our model to try and predict this particular data point (and possibly others) is not very helpful to us. The crime value could be anywhere between 0 and 1620 if we want to be confident in our assessment.