Analytics Modeling HW4

Fall 2024

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 α to be closer to 0 or 1 and why?

I work for a sports & casino gambling company as a data scientist. One problem at work for which an Exponential Smoothing model is appropriate would be to predict how much money a user will deposit into their account each day for the next several days. Since most users deposit money to place bets soon after, a shorter-term predictive model like Exponential Smoothing is appropriate. I would expect the smoothing parameter α to be close to zero since the data will almost always have a lot of random variation in it. Most players do not make deposits every single day, so I would want to give less weight to the most recently observed data point and more weight to the most recent estimated baseline value.

```r
temps <- read.table("temps.txt",  stringsAsFactors = FALSE, header = TRUE)
# remove weird X on column names
names(temps) <- gsub(x=names(temps), pattern = "X", replacement = "")
head(temps)
```

```
##     DAY 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
## 1 1-Jul   98   86   91   84   89   84   90   73   82   91   93   95   85   95
## 2 2-Jul   97   90   88   82   91   87   90   81   81   89   93   85   87   90
## 3 3-Jul   97   93   91   87   93   87   87   87   86   86   93   82   91   89
## 4 4-Jul   90   91   91   88   95   84   89   86   88   86   91   86   90   91
## 5 5-Jul   89   84   91   90   96   86   93   80   90   89   90   88   88   80
## 6 6-Jul   93   84   89   91   96   87   93   84   90   82   81   87   82   87
##   2010 2011 2012 2013 2014 2015
## 1   87   92  105   82   90   85
## 2   84   94   93   85   93   87
## 3   83   95   99   76   87   79
## 4   85   92   98   77   84   85
## 5   88   90  100   83   86   84
## 6   89   90   98   83   87   84
```

```r
# set seed for reproducible results
set.seed(42)
```

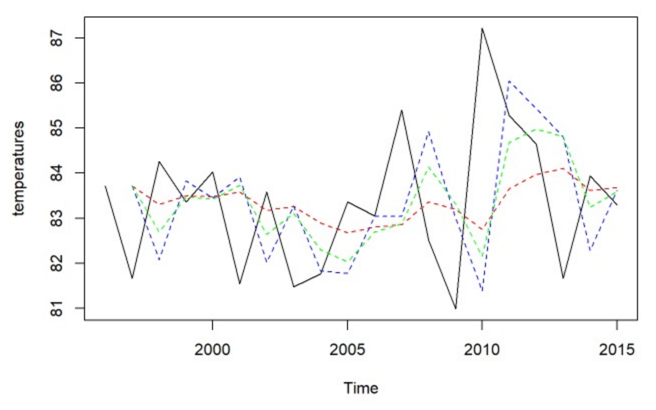
Question 7.2 (daily model)

Using the 20 years of daily high temperature data for Atlanta, build an use an Exponential Smoothing Model to help make a judgment of whether the unofficial end of summer has gotten later over those 20 years.

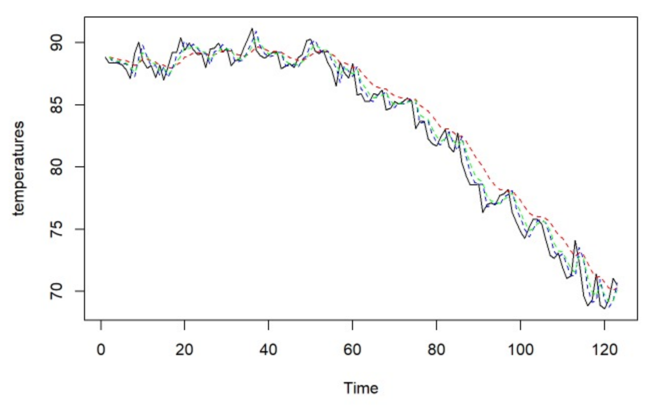
Getting a handle on Exponential Smoothing

To get a better visual of the potential changes each year, I looked at the daily and yearly averages for trends. Like with the previous daily model, there does not appear to be any consistent trend that indicates that summer is ending later and later each year. I used this opportunity to compare three different models for each of the daily and yearly averages with low (0.2), medium (0.5), and high (0.8) α values to get a better understanding of how Exponential Smoothing works. Based off this alone, I don't think there is enough data here to conclude with any confidence that summer is ending later and later as time goes on in Atlanta, especially since some of the increases in average temperature may be caused by randomness.

```r
# get average daily and yearly temperatures
yearly_avg <- colMeans(as.matrix(temps[,-1]))
daily_avg <- rowMeans(as.matrix(temps[,-1]))

# convert to time series data format
ts_avg_year <- ts(data = yearly_avg, start = 1996)
ts_avg_daily <- ts(data = daily_avg)

# create Exponential Smoothing model with Holt Winters
avg_year_1 <- HoltWinters(ts_avg_year, alpha = .2, beta = FALSE, gamma = FALSE)
avg_year_2 <- HoltWinters(ts_avg_year, alpha = .8, beta = FALSE, gamma = FALSE)
avg_year_3 <- HoltWinters(ts_avg_year, alpha = 0.5, beta = FALSE, gamma = FALSE)

avg_daily_1 <- HoltWinters(ts_avg_daily, alpha = .2, beta = FALSE, gamma = FALSE)
avg_daily_2 <- HoltWinters(ts_avg_daily, alpha = .8, beta = FALSE, gamma = FALSE)
avg_daily_3 <- HoltWinters(ts_avg_daily, alpha = 0.5, beta = FALSE, gamma = FALSE)

# plot the observed yearly average values with three different smoothed models (low, medium, high alpha)
plot(ts_avg_year, ylab = "temperatures", xlim = c(1996, 2015))
lines(avg_year_1$fitted[,1], lty = 2, col = "red")
lines(avg_year_2$fitted[,1], lty = 2, col = "blue")
lines(avg_year_3$fitted[,1], lty = 2, col = "green")
```




```r
# plot the observed daily average values with three different smoothed models (low, medium, high alpha)
plot(ts_avg_daily, ylab = "temperatures")
lines(avg_daily_1$fitted[,1], lty = 2, col = "red")
lines(avg_daily_2$fitted[,1], lty = 2, col = "blue")
lines(avg_daily_3$fitted[,1], lty = 2, col = "green")
```



Reformatting Data & HoltWinters

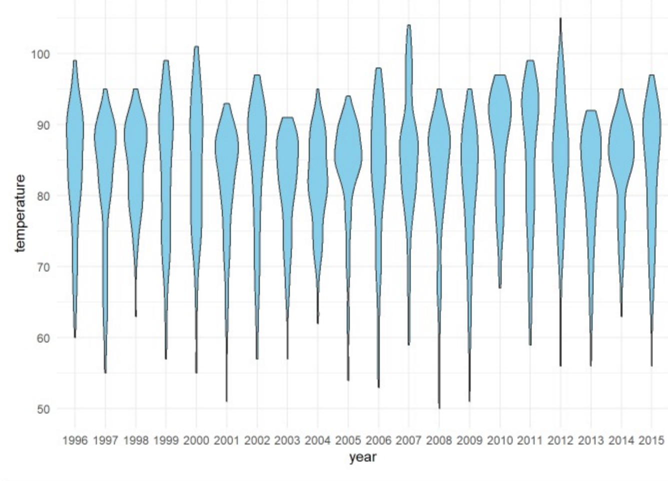
First I reformatted the `temps` data to be in a time-series format and converted it into a time-series object so that I can make use of some of the functions R has to analyze time series data. The decomposition of the the data suggests that there has been a trend of increasing high temperatures over the years and the seasonal cycle of temperatures appears pretty consistent to the eye.

I tried two different models on the daily time series data, one with a low α value of 0.2 and one with a high α value of 0.8. The higher α gives more weight to the observed value and the end result is that the model with the higher α is much less smooth than the model with the lower α. However, with so many data points and such a stuffed graph, it's difficult to observe any real yearly trends with the naked eye. It doesn't appear that there is a trend towards a later and later end of summer visually.

```r
# reformat data into two columns: one data point for each day and its recorded high temperature
temps_year <- melt(temps
                   , id.vars = c("DAY")
                   , variable.name= "YEAR"
                   , value.name="TEMP")

# violin plot of yearly temperatures
ggplot(temps_year, aes(x=temps_year$YEAR, y=temps_year$TEMP)) +
  geom_violin(fill="skyblue") +
  xlab("year") +
  ylab("temperature") +
  theme_minimal()
```

```
## Warning: Use of `temps_year$YEAR` is discouraged.
## ℹ Use `YEAR` instead.
```

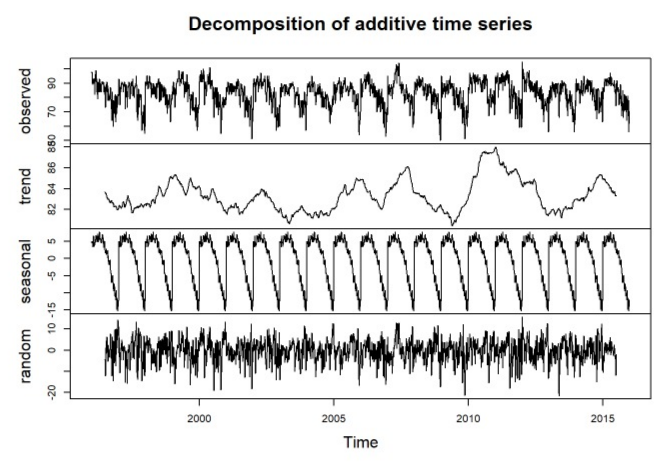
```
## Warning: Use of `temps_year$TEMP` is discouraged.
## ℹ Use `TEMP` instead.
```




```r
# create date column
temps_year$DATE <- paste(temps_year$DAY, temps_year$YEAR)

# reformat date column as date datatype instead of character
temps_year$DATE <- dmy(temps_year$DATE)
head(temps_year)
```

```
##     DAY YEAR TEMP       DATE
## 1 1-Jul 1996   98 1996-07-01
## 2 2-Jul 1996   97 1996-07-02
## 3 3-Jul 1996   97 1996-07-03
## 4 4-Jul 1996   90 1996-07-04
## 5 5-Jul 1996   89 1996-07-05
## 6 6-Jul 1996   93 1996-07-06
```

```r
# create time-series object out of the data
ts_temps <- ts(data = temps_year$TEMP, frequency=123, start=1996)

# decompose the time series data into its three main components: long term trends, seasonal cycle, and random and plot it
plot(decompose(ts_temps))
```



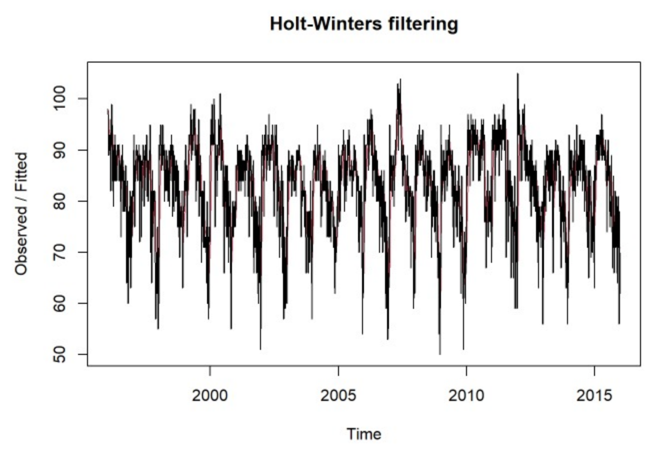

```r
# create Exponential Smoothing model with Holt Winters
lowa_daily_forecast <- HoltWinters(ts_temps, alpha = .2, beta = FALSE, gamma = FALSE)
higha_daily_forecast <-  HoltWinters(ts_temps, alpha = .8, beta = FALSE, gamma = FALSE)

# look at the smoothed data
lowa_fitted <- lowa_daily_forecast$fitted
higha_fitted <- higha_daily_forecast$fitted
head(lowa_fitted)
```

```
## Time Series:
## Start = c(1996, 2)
## End = c(1996, 7)
## Frequency = 123
##              xhat    level
## 1996.008 98.00000 98.00000
## 1996.016 97.80000 97.80000
## 1996.024 97.64000 97.64000
## 1996.033 96.11200 96.11200
## 1996.041 94.68960 94.68960
## 1996.049 94.35168 94.35168
```

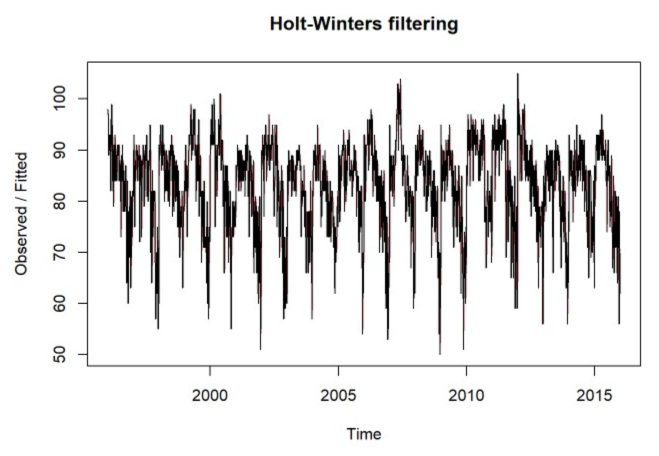
```r
head(higha_fitted)
```

```
## Time Series:
## Start = c(1996, 2)
## End = c(1996, 7)
## Frequency = 123
##              xhat    level
## 1996.008 98.00000 98.00000
## 1996.016 97.20000 97.20000
## 1996.024 97.04000 97.04000
## 1996.033 91.40800 91.40800
## 1996.041 89.48160 89.48160
## 1996.049 92.29632 92.29632
```

```r
# add s1 = x1 for first smoothed data point
# lowa_fitted[,-1]
#temps_year[,3]


# plot the forecasts
plot(lowa_daily_forecast)
```




```r
plot(higha_daily_forecast)
```



CUSUM with smoothed HoltWinters data

To see if summer appears to be ending later and later (mathematically), I used the smoothed data from the HoltWinters model to do a CUSUM analysis using the model made with the lower α, giving more weight to the previous baseline rather than the observed data points. I split the smoothed data into dataframes by year and did individual CUSUM analyses on each dataframe. From each analysis, I stored the first date that was identified as a decrease in temperature for each year.

The results of that show that, for most years, the end of summer falls somewhere in October, usually early-mid October. This analysis did not give me any concrete evidence that summer is definitely ending later and later as time goes by. Maybe there is a slight trend towards it ending later, but nothing so definitive as to make that conclusion definite.

```r
# get smoothed datapoints from low a HoltWinters
smoothed_points <- as.matrix(lowa_fitted[,-1])
length(smoothed_points) <- length(temps_year[,1])

# rule of thumb: C = half of the standard deviation of the data points
stdev<- sd(as.matrix(lowa_fitted[,-1]))
C <- .5*stdev           # Critical Value
T <- 5*stdev            # Threshold

smoothed_data <- data.frame(date_x = temps_year[,4]
                            , year = temps_year[,2]
                            , day = temps_year[,1]
                            , xi = smoothed_points
                            , mu = mean(as.matrix(lowa_fitted[,-1]))
                            , dDiff = mean(as.matrix(lowa_fitted[,-1]))-smoothed_points-C
                            )
head(smoothed_data)
```

```
##       date_x year   day       xi       mu     dDiff
## 1 1996-07-01 1996 1-Jul 98.00000 83.39188 -18.09509
## 2 1996-07-02 1996 2-Jul 97.80000 83.39188 -17.89509
## 3 1996-07-03 1996 3-Jul 97.64000 83.39188 -17.73509
## 4 1996-07-04 1996 4-Jul 96.11200 83.39188 -16.20709
## 5 1996-07-05 1996 5-Jul 94.68960 83.39188 -14.78469
## 6 1996-07-06 1996 6-Jul 94.35168 83.39188 -14.44677
```

```r
# split data into subsets by year
split_data <- split(smoothed_data, smoothed_data$year)

results <- vector() # empty vector to store data

for (i in 1:length(split_data)) {
# low a HoltWinters CUSUM
  cusum <- split_data[[i]]

  # calculate CUSUM metric, but set to zero if the metric is less than zero
  cusum <- cusum %>% mutate(decrease = accumulate(dDiff,  ~ ifelse(.x + .y < 0, 0, .x + .y)))

  # if the metric >= T, mark TRUE
  cusum$dChange <- ifelse(cusum$decrease>=T, TRUE, FALSE)

  # get all rows after the first increase change has been identified
  decrease_identified <- cusum[which(cusum$dChange == TRUE),]

  # get first identified decrease for each year
  first_identified <- head(decrease_identified,1)

  # store the first identified change of each year in the results vector
  results <- as.vector(c(results, first_identified$day))
}

df_results <- data.frame(year = seq(1996, 2015)
                         , day = results)
df_results
```

```
##    year    day
## 1  1996  6-Oct
## 2  1997  4-Oct
## 3  1998 18-Oct
## 4  1999  6-Oct
## 5  2000  6-Oct
## 6  2001  5-Oct
## 7  2002 15-Oct
## 8  2003  5-Oct
## 9  2004 10-Oct
## 10 2005 23-Oct
## 11 2006 15-Oct
## 12 2007 23-Oct
## 13 2008 21-Oct
## 14 2009  9-Oct
## 15 2010 27-Oct
## 16 2011 12-Oct
## 17 2012 12-Oct
## 18 2013  5-Jul
## 19 2014 23-Oct
## 20 2015  3-Oct
```