ISYE 6501 Homework 3

Due: June 10, 2021

1 Exponential smoothing

1.1 Question 7.1

A situation from current events in which using exponential smoothing would be appropriate is analyzing the number of Covid-19 cases in the U.S. I think many of us have seen plots of the raw number of positive cases per day as well as the 7-day moving averages, but even the latter can have awkward data spikes. What if for the sake of analysis one just needs an even more generalized curve. The beauty of using exponential smoothing here, specifically "triple" smoothing, is that the smoothed curve could encompass recent trends i.e. large events (many people close to one another) and seasonality i.e. summer time (less mask wearing) as well as being able to tune parameters such as α to trust what we see vs. assuming random events. For this example, the data one would need is the number of positive cases on every day such that we obtain time-series data. For this particular instance (without getting too political), if one believes in the "science", they could expect α to be closer to 1. This is because a value for α closer to 1 indicates stronger trust in what is observed and less trust in randomness or the previous estimate. The fun aspect of this example as I just eluded to is that the human generating or analyzing the model has a significant affect of the parameters (and hence model) chosen based on their own experiences and biases.

1.2 Question 7.2

Using July-October daily high-temperature data for Atlanta from 1996-2015, I previously attempted to estimate whether or not the unofficial end of summer has gotten later using a CUSUM approach. The CUSUM algorithm detected a change (increase) in 2009 and 2010 which provided good indicators to look further into and visualize the data to see if this had any implications on the 20 year trend. What can be seen in Figure 1, is that it is quite challenging to make that decision with such few data points and one could really argue either side of the argument.

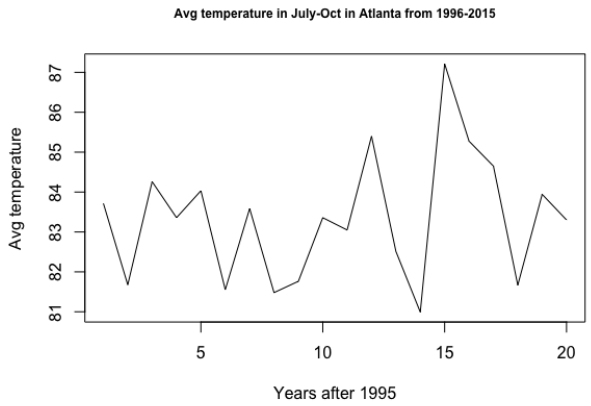

Figure 1: Inspecting when Atlanta's summer has gotten warmer

For this reason, I chose to build an exponential smoothing model to attempt a better prediction of this posed question. Note that one could certainly use exponential smoothing on the average yearly temperature data presented in Figure 1. Although this could certainly work to extract a general trend, because there are so few data points, I wanted to build a model around daily time-series data instead of yearly time-series data. By doing this, I can utilize all 2,460 data points as opposed to 20. Additionally, this expansion gives the exponential smoothing model the ability to use trends, seasonality and simply more data to better model what is actually happening. I completed this by concatenating the data for each year as seen in Appendix A. Figure 2 shows a plot of the cyclic-like daily temperature time-series data.

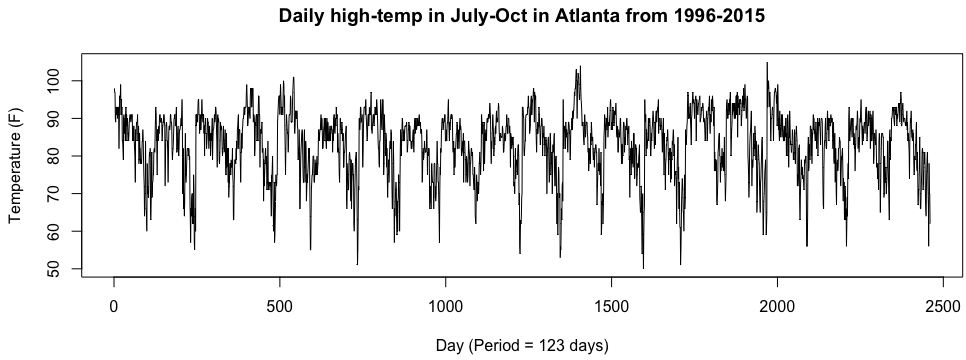

Figure 2: A clear cyclical pattern can be seen, hypothetically every 123 days.

As stated, the beauty of using exponential smoothing (triple smoothing or Holt-Winter's method), is the inclusion of trends, T_t at time t, and multiplicative cyclic patterns/seasonality, C_t. Equation 1 summarizes exponential smoothing using these terms.

$$s_t = \alpha \frac{x_t}{C_{t-L}} + (1-\alpha)(s_{t-1} + T_{t-1}) \tag{1}$$

$$T_t = \beta(s_t - s_{t-1}) + (1-\beta)T_{t-1} \tag{2}$$

$$C_t = \gamma \frac{x_t}{s_t} + (1-\gamma)C_{t-L} \tag{3}$$

Here, s_t is the expected baseline response at time t, x_t is the observed temperature, L is the length of a cycle and α, β, γ are tunable parameters between 0 and 1. These parameters can be optimized and solved for in an automated fashion using R as seen in Appendix A. Following [1], I first test for the inherent frequencies for the time series using a periodogram. As expected, the period $(1/f)$ is 123 days or time points. See Figure 3.

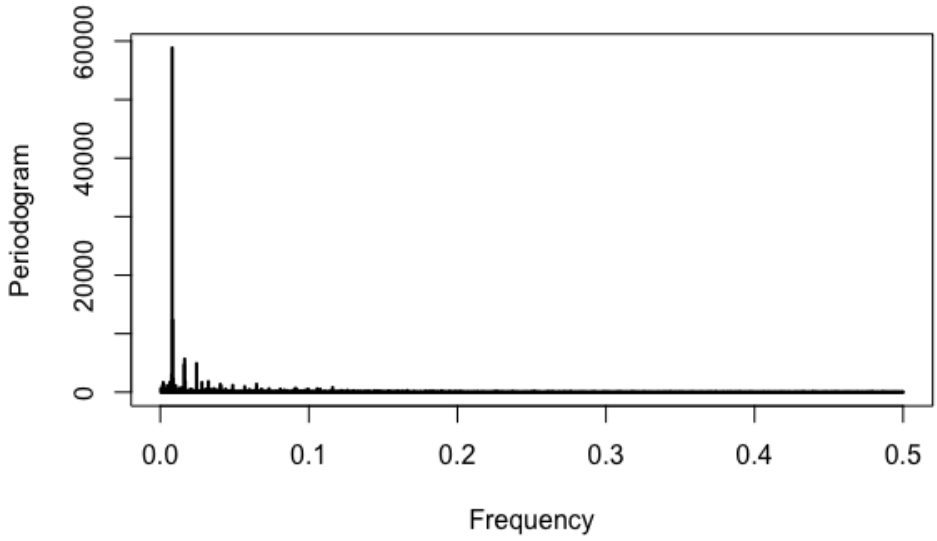

Figure 3: A detected frequency of $\approx 1/123$.

Using this information, I completed an exponential smoothing model using $HoltWinters()$ in R. The results are fascinating and are summarized in Figure 4. After minimizing the squared error, the model outputs optimal parameters: $\alpha = 0.615003$, $\beta = 0.0$, and $\gamma = 0.5495256$. There are a few main takeaways here. 1) The moderate value for α makes sense because after all, it is weather and we expect there to be some randomness as well as trusting the previous day's value. 2) The model does not believe there is any trend in the time-series data. Note that this part of the model is really solving the question of "are temperatures increasing over the years?" which is the real question at hand. I forecasted the temperatures for the next four years to also try and answer this question. Figure 5 shows the results. It is too difficult to make any determination. 3) One could analyze the periodicity of the seasonality data to answer if summer is ending later. It is apparent in Figure 4 that the magnitude of seasonality is getting greater over time. This is also verified with a significant γ value. This means that the 123 period is having more of an affect on the time-series data. As this multiplicative factor increases, the model is accounting for the fact that temperatures are higher simply because it is a certain day. Although there is not a drastic change, it can be argued that summers are infact ending later.

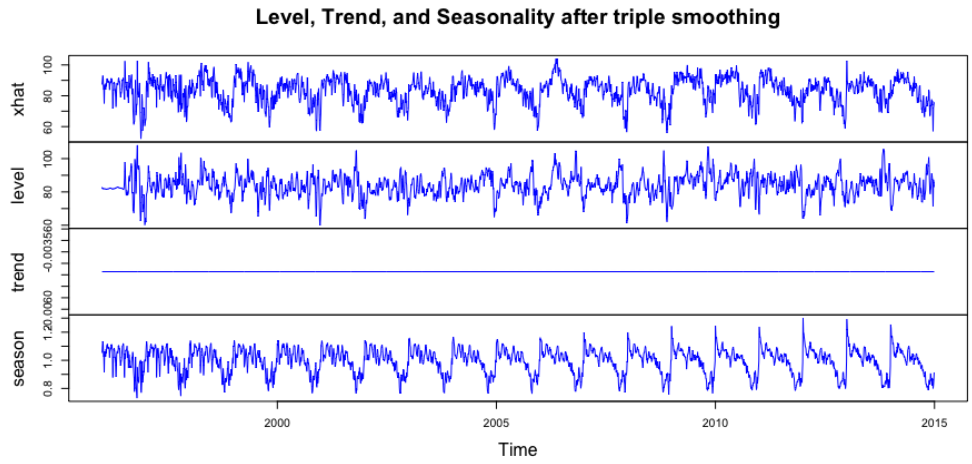

Figure 4: Determining summer's end using exponential smoothing.

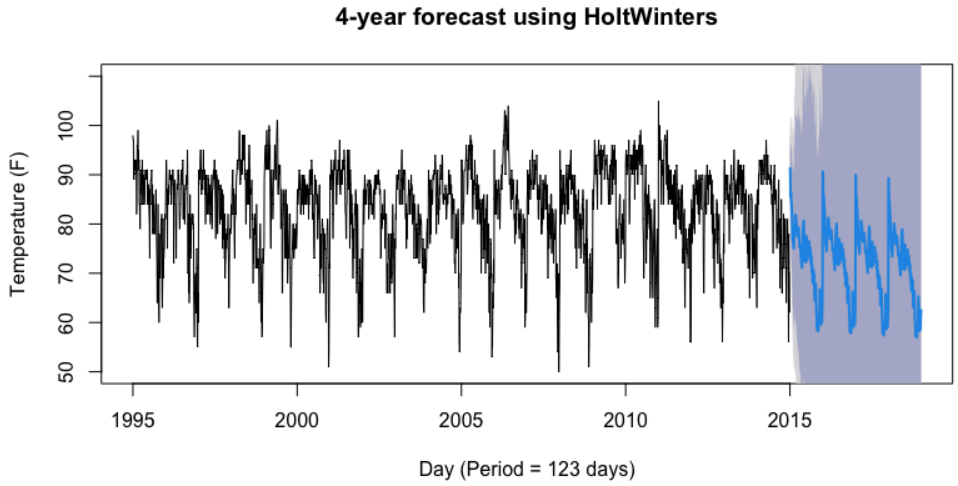

Figure 5: 4-year prediction of daily temperature data.

2 Linear regression

2.1 Question 8.1

An example of where a linear regression model can be applied in everyday life is the housing market. The cost or expected cost of a home could be based on several predictors. The reason I phrase the previous sentence as such is because one could use linear regression in this instance to answer either descriptive or prescriptive questions. I.e. how does the neighbor's home price affect the model? vs. using the set of predictors, how much is this home worth? There are several predictors that can be used to analyze or predict home price: cost of the nearest k neighbors homes, square footage, property acreage, proximity to city center, proximity to nearest school district, renovations completed etc.

2.2 Question 8.2

Using the *uscrime* dataset, I used multiple linear regression to predict the crime rate (number of crimes per 100,000 people) in a city with the predictors shown in Table 1.

Predictor	Value
M	14.0
So	0.0
Ed	10.0
Po1	12.0
Po2	15.5
LF	0.640
M.F	94.0
Pop	150.0
NW	1.1
U1	0.120
U2	3.6
Wealth	3200
Ineq	20.1
Prob	0.04
Time	39.0

Table 1: Data point to predict crime rate.

As shown in Appendix B, I used the $lm()$ function in R to first complete a regression model using all of the predictors with the goal of discovering which coefficients showed significance. Recall that the model aims to minimize the squared error. The model can be described mathematically as seen in Equation 4.

$$y = a_0 + a_1x_1 + a_2x_2 + ... + a_mx_m \tag{4}$$

Here, y is the response, a_i is the coefficient of predictor i, and x_i is the value of the i^{th} predictor like in Table 1. Completing the regression model yielded the following results shown in Figure 6 and Equation 5.

```
Residuals:
    Min      1Q  Median      3Q     Max
-395.74  -98.09   -6.69  112.99  512.67

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.984e+03  1.628e+03  -3.675 0.000893 ***
M            8.783e+01  4.171e+01   2.106 0.043443 *
So          -3.803e+00  1.488e+02  -0.026 0.979765
Ed           1.883e+02  6.209e+01   3.033 0.004861 **
Po1          1.928e+02  1.061e+02   1.817 0.078892 .
Po2         -1.094e+02  1.175e+02  -0.931 0.358830
LF          -6.638e+02  1.470e+03  -0.452 0.654654
M.F          1.741e+01  2.035e+01   0.855 0.398995
Pop         -7.330e-01  1.290e+00  -0.568 0.573845
NW           4.204e+00  6.481e+00   0.649 0.521279
U1          -5.827e+03  4.210e+03  -1.384 0.176238
U2           1.678e+02  8.234e+01   2.038 0.050161 .
Wealth       9.617e-02  1.037e-01   0.928 0.360754
Ineq         7.067e+01  2.272e+01   3.111 0.003983 **
Prob        -4.855e+03  2.272e+03  -2.137 0.040627 *
Time        -3.479e+00  7.165e+00  -0.486 0.630708
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 209.1 on 31 degrees of freedom
Multiple R-squared:  0.8031,    Adjusted R-squared:  0.7078
F-statistic: 8.429 on 15 and 31 DF,  p-value: 3.539e-07
```

Figure 6: Summary of regression results using all predictors.

$$y = -5984.288 + 87.83017x_1 - 3.803450x_2 + 188.3243x_3 + 192.8043x_4 - 109.4219x_5$$
$$-663.8261x_6 + 17.40686x_7 - 0.7330081x_8 + 4.204461x_9 - 5827.103x_{10} + 167.7997x_{11} \quad (5)$$
$$+0.09616624x_{12} + 70.67210x_{13} - 4855.266x_{14} - 3.479018x_{15}$$

For the sake of comparison, I used all of the coefficients of the model from Figure 6 and predicted the crime rate of the data point in Table 1. This prediction yielded a value of 155.4349. The summary in Figure 6 makes it very clear that only four of the 15 predictor's coefficients (not including the intercept) have p-values < 0.05 and hence are significant when building the model, minimizing the squared error and predicting the actual response. I then completely redid the multiple regression using solely the data of significant predictors, M, Ed, Ineq, and Prob. The results of this newer concise model are summarized in Figure 7 and Equation 6.

```
Residuals:
    Min      1Q  Median      3Q     Max
-532.97 -254.03  -55.72  137.80  960.21

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1339.35    1247.01  -1.074  0.28893
M              35.97      53.39   0.674  0.50417
Ed            148.61      71.92   2.066  0.04499 *
Ineq           26.87      22.77   1.180  0.24458
Prob        -7331.92    2560.27  -2.864  0.00651 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 347.5 on 42 degrees of freedom
Multiple R-squared:  0.2629,    Adjusted R-squared:  0.1927
F-statistic: 3.745 on 4 and 42 DF,  p-value: 0.01077
```

Figure 7: Summary of regression results using M, Ed, Ineq, and Prob predictors.

$$y = -1339.34621 + 35.97296x_1 + 148.60531x_2 + 26.87457x_3 - 7331.91531x_4 \quad (6)$$

Using this concise model, I predicted the crime rate of the data point in Table 1 and now obtained a value of 897.2307. By visually inspecting the raw data, this value seems to be way more appropriate. There are several more takeaways from inspecting this model. 1) Only two of the p-values for the four coefficients are now less than 0.05. 2) the R-squared value drop significantly from the model which included all predictors, from 0.8031 to 0.2629. This is in fact alright, it just needs to be made aware that the model might not account for much variability. Lastly, I visualized the regression in each significant dimension of the data with the corresponding confidence/prediction interval. See Figures 8-11.

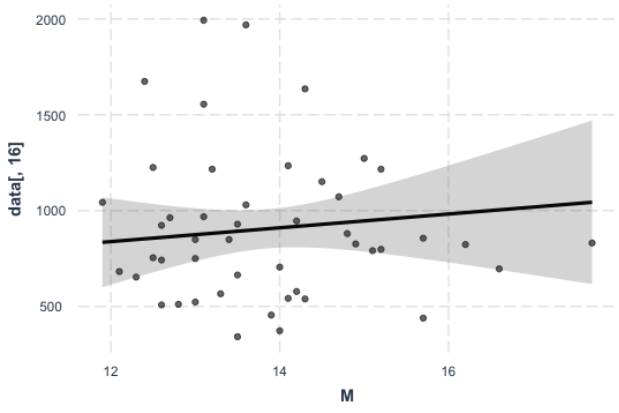

Figure 8: Regression in M dimension.

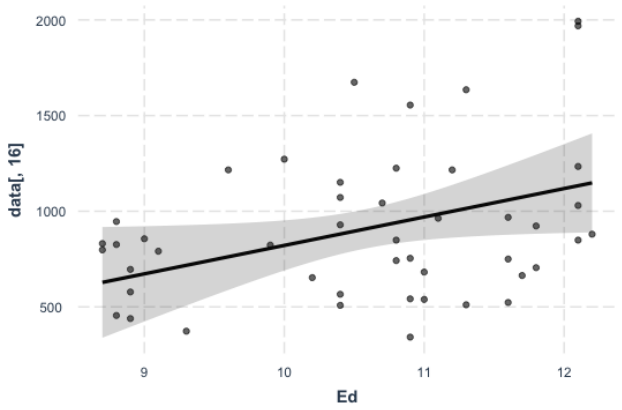

Figure 9: Regression in Ed dimension.

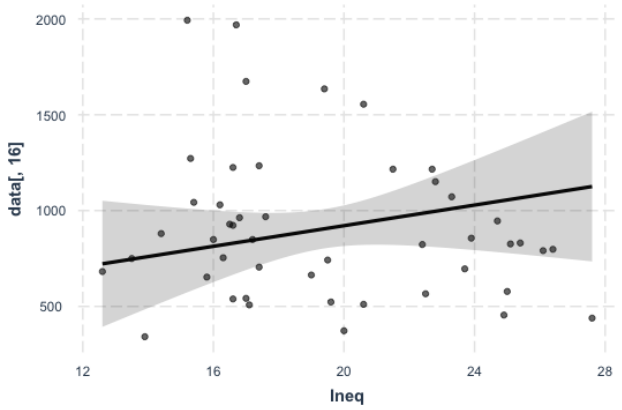

Figure 10: Regression in Ineq dimension.

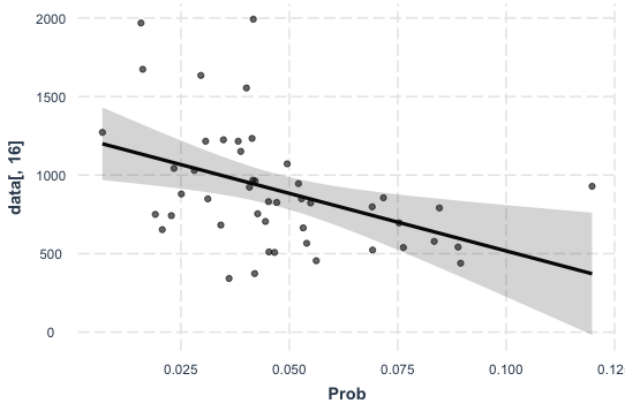

Figure 11: Regression in Prob dimension.

References

[1] https://rstudio-pubs-static.s3.amazonaws.com/366011_3ee069277eb84547824f8f4022973823.html

Appendix A Exponential smoothing code

```
data_load <- read.table("temps.txt", header=TRUE)
data <- c()
for (year in 2:ncol(data_load)) {
  for (day in 1:nrow(data_load)) {
  data <- c(data, data_load[day,year])
  }
}
plot(data, type="l", main='Daily high-temp in July-Oct in Atlanta from 1996-2015',
     xlab='Day (Period = 123 days)', ylab='Temperature (F)')

library(TSA)
print(periodogram(data))

ts <- ts(data, start = c(1995,1), frequency = 123)
print(ts)

hw <- HoltWinters(ts, seasonal = "multiplicative")
print(hw)
fit <-(fitted(hw))
print(fit)
plot(fit, main='Level, Trend, and Seasonality after triple smoothing', col = "blue")
library(forecast)
plot(forecast(hw, 492), main='4-year forecast using HoltWinters',
     xlab='Day (Period = 123 days)', ylab='Temperature (F)', ylim=c(50,110))
```

Appendix B Linear regression code

```
data <- read.table("uscrime.txt", header=TRUE)
crimes <- data[,1:15]

model <- lm(data[,16]~., data=crimes)
summary(model)
#anova(model)

test_point <- c(14.0, 0, 10.0, 12.0, 15.5, 0.640, 94.0, 150, 1.1,
                0.120, 3.6, 3200, 20.1, 0.04, 39.0)

#pred <- predict(model, test_point)
coeffs <- model$coefficients
print(coeffs)
keep <- c(2, 4, 14, 15)

print(as.matrix(coeffs[keep]))
print(as.matrix(point))

print((t(as.matrix(coeffs[-1])) %*% as.matrix(test_point)) + coeffs[1])

model_concise <- lm(data[,16]~., data=crimes[,keep-1])
summary(model_concise)
```

```
coeffs2 <- model_concise$coefficients
print(coeffs2)
library(jtools)
point <- test_point[keep-1]
print(t(as.matrix(coeffs2[-1])) %*% as.matrix(point) + coeffs2[1])
effect_plot(model_concise, pred = M, interval = TRUE, plot.points = TRUE)
```