Homework 4

SP 25: Introduction to Analytics Modeling

Student: Tyler Wacker, tjwac27@gmail.com

My solutions to this week's homework are based in last week's homework. As a result, my final program for question 7.2 is very long! Do not worry, I have taken care to explain my methods in both the code, and my writeup.

> **Question 1: '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 (the first smoothing parameter) to be closer to 0 or 1, and why?*

One of the key advantages of exponential smoothing was revealed in lesson 7.3. In this lesson, Joel states that exponential smoothing ensures that "All past observations are accounted for, with the more recent observations being more important to the current baseline estimate." I was recently hired as an intern in the Biotech space to build websites and establish web presence.

Using this method may be pivotal to understanding current trends, while still valuing the cultural norms of past semantic data, to maximize Search Engine Optimization (SEO).

I could source data using various methods. Google provides a suite of nice analytic tools to measure web presence along with histories of popular search terms. If I were to use these datasets, I would have to take care in transforming the text data into a workable form. I hypothesize that the value would be closer to 0, as cultural trends are somewhat unpredictable and random.

> **Question 2: '7.1'**
>
> *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.*
>
> - Note: *in R, you can use either HoltWinters (simpler to use) or the smooth package's es function (harder to use, but more general). If you use es, the Holt-Winters model uses model="AAM" in the function call (the first and second constants are used "A"dditively, and the third (seasonality) is used "M"ultiplicatively; the documentation doesn't make that clear).*

My program utilizes the CUSUM and Linear Regression patterns to predict where the 'end of summer' is occuring, predicts the trend, and provides an R value to judge statistical significance.

Regarding my Holtwinters implementation, the gamma is set to False. The reason why the 'seasonal' aspect is disregarded is because our seasons are improperly defined in the original dataset, along with our dataset only providing 123/365 days of the year.

Figure 1: Year 2000 End of Summer (EOS) predicted using CUSUM and Linear Regression methods with no data smoothing. Predicted EOS is AUG-31. Lin Reg Slope = 0.053 deg/year with too low P-Value. (0.40027, expected was 0.95+)

The Holtwinters exponential data smoothing is used an an optional data cleaning step within my program. This allows a user to select whether or not they would like to use the data smoothing to attempt to mitigate the 'randomness' found within the dataset.

Holtwinters was implemented using the es package within my program. By default, this implementation uses an optimizer to find idealized values.

Attached to this documents are figures produced within my webapp comparing the use of no data smoothing (the 'none' option), and data smoothing (the 'es Holtwinters' option).

Curiously, this pattern of lower P-Values for the Holtwinters data smoothing continues through year by year implementations. While I only attached a single year to this writeup, it may be valuable to look at various other years by running this code on your local machine.

At least with my implementation, this data smoothing method has not seemed to produce a quality dataset and we're still bogged down by the issues of last week's implementations. While there does seem to be a positive slope in EOS predictions, and the CUSUM predicted date does consistently move forwards, this is done with extremely low P-Values. I don't believe we can prove that summer is ending later beyond a reasonable doubt.

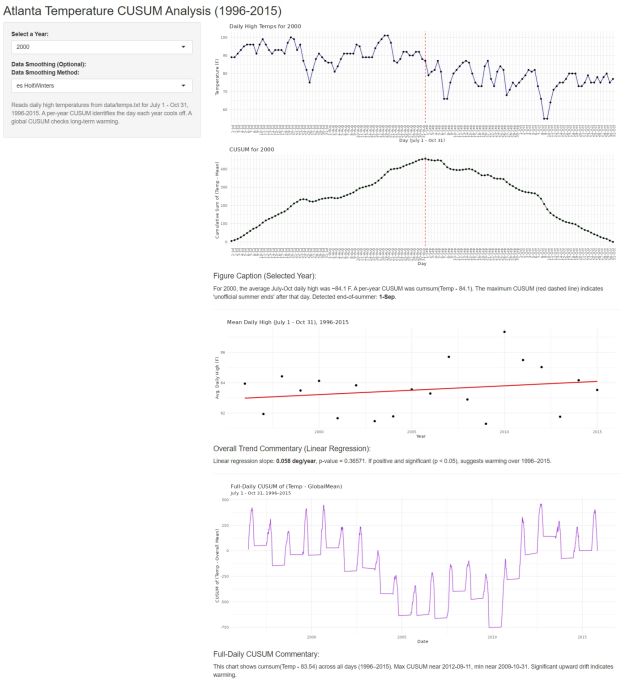

Figure 2: Year 2000 End of Summer (EOS) predicted using CUSUM and Linear Regression methods with exponential data smoothing. Predicted EOS is SEP-1. Lin Reg Slope = 0.058 deg/year with too low P-Value. (0.36571, expected was 0.95+)

References

ChatGPT for Code Review
Overleaf