GT Introduction to Analytics Modeling - Week 3 HW

Robert Phillips

May 29, 2017

The following contains the original questions along with answers and the R code that was used.

Question 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?

Answer: One of my daily duties is to monitor, capture, and analyze system performance. The data is extract from log files. The performance metrics can vary signficantly from hour to hour, therefore the default visualizations of the data can appear to be too noisy and difficult to read. Being time series data, we could use exponential smoothing to visualize and better interpret the metrics. Regarding a value for alpha, we would likely start with a value closer to .5 as recent values aren't necessarily more important than older values.

Question 2

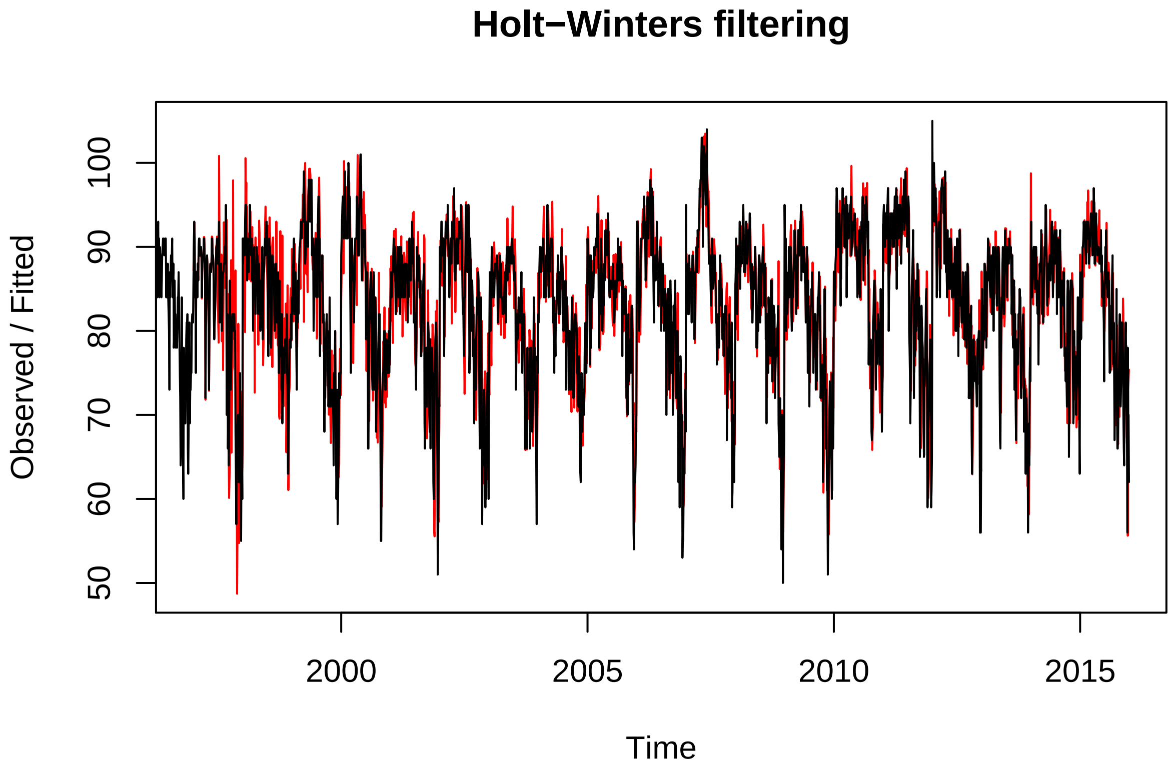
Using the 20 years of daily high temperature data for Atlanta (July through October) from Homework 2 Question 5, build and use an exponential smoothing model to help make a judgment of whether the unofficial end of summer has gotten later over the 20 years. (Part of the point of this assignment is for you to think about how you might use exponential smoothing to answer this question.)

We first load the data. We also capture the years for plotting purposes.

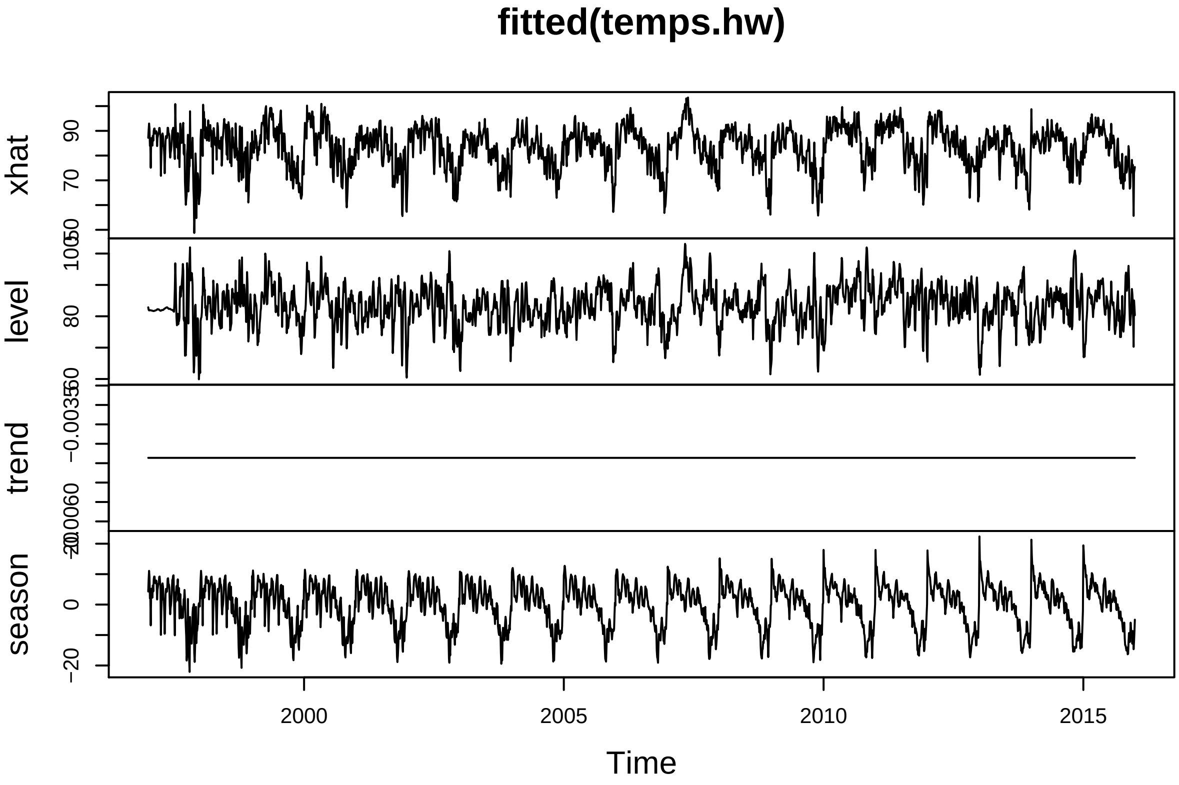
```r
temps = read.csv('temps.txt', header = T, sep = '\t')
years = as.numeric(substring(colnames(temps[-1]), 2))
```

For this execise, we'll convert all of the temperature data into a time series and use that as input to the Holt Winters function. The function will determine optimal parameter values (alpha, beta and gamma). We can then inspect the parameters paying attention to beta to determine if any trend is detected. If a trend is detected, and the official end of summer is occuring later in the year, then we would expect the temperative values to move in an upward trend since the temperatures would be increasing over time.

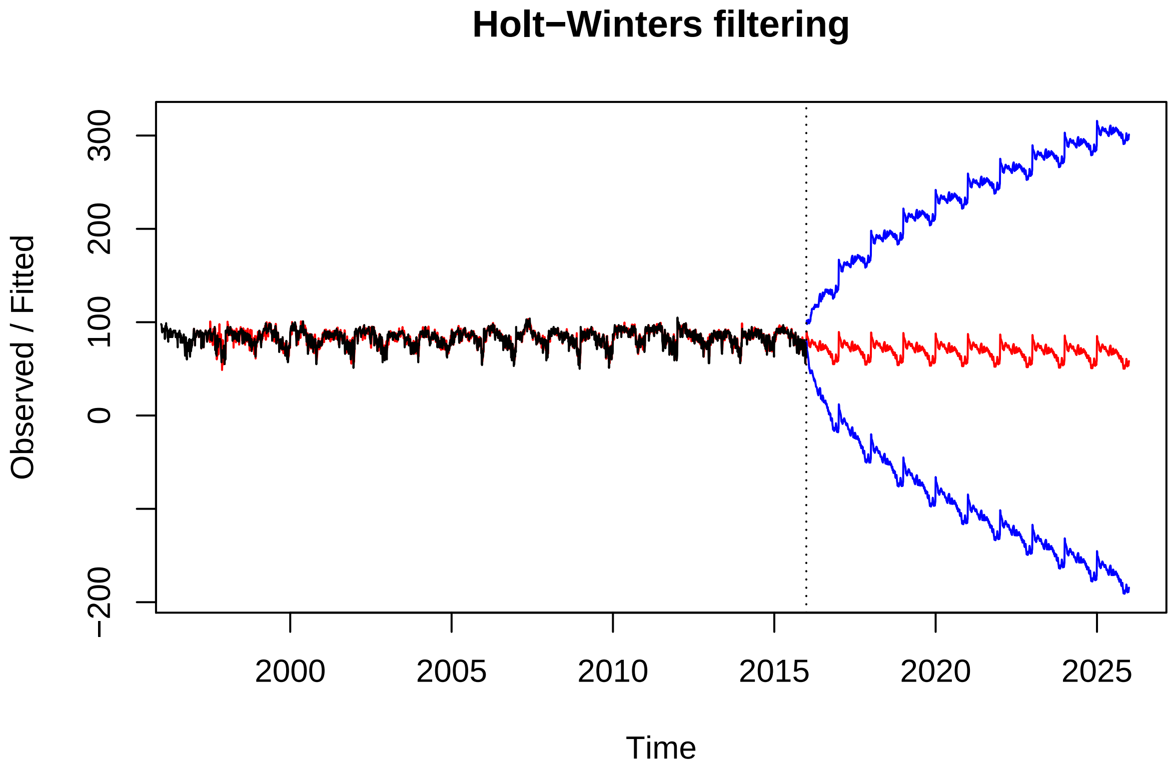
```r
#fit with all temps
temps.all = unname(unlist(temps[,-1]))

#convert to a time series
temps.count = length(temps$DAY)
temps.ts = ts(temps.all, frequency=temps.count, start=c(years[1], 1))

#calculate and plot Holt Winters
temps.hw <- HoltWinters(temps.ts)
plot(temps.hw)
```



```
plot(fitted(temps.hw))
```



```
#forcast 10 years into the future
temps.forecast <- predict(temps.hw, n.ahead = temps.count*10, prediction.interval = T, level = 0.95)
plot(temps.hw, temps.forecast)
```



The value of beta is 0. This indicates that a trend was not detected. We can also inspect the forecast based plot which is 10 years into the future. The fitted plot (in red) does not appear to move up or down in terms of a trend which further indicates the lack of a trend.

Question 3

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.

Answer: Related to Question 1, in addition to monitoring performance metrics, we could also use log data to build a data set from which a linear regression model could be created to predict future performance metrics. The model may also indicate which metrics are better predictors.

An example performance metric is “run time” for particular financial calculations. Predictors would be quantifications of items such as transactions per account, transaction types, data size, process configuration, and transaction complexity.

Question 4

Using crime data from http://www.statsci.org/data/general/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.

We can load the data and create a model in a few lines of code. As requested in the assignment, we create a model with Crime as the dependent variable and all other features as independent variables. We can then view the coefficients along with corresponding statistics.

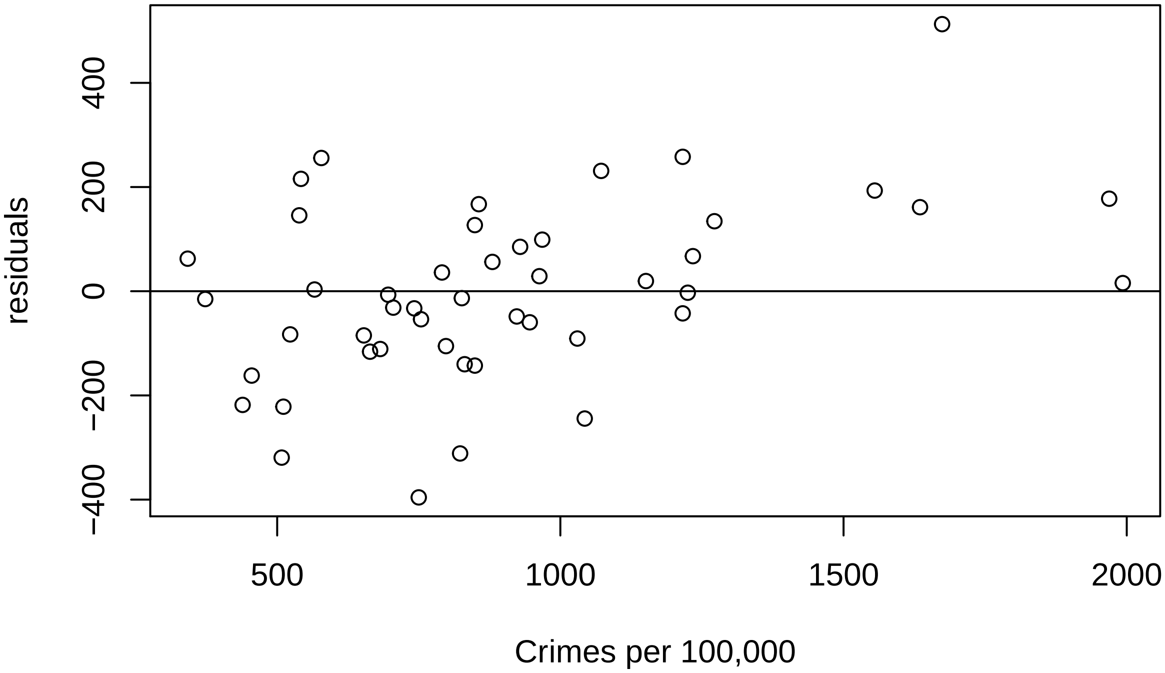
```
crimes = read.csv('uscrime.txt', header=T, sep='\t')
model1 = lm(Crime ~ ., data=crimes)
summary(model1)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = crimes)
##
## 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
```

We can note the asteriks indicating statistical signficance. It appears that only a few independent variables may be significant, therefore we may be able to simplify our model.

We can plot the residuals to visualize any potential patterns. As shown in the plot, there does appear to be an obvious pattern. This indicates that a linear model may be reasonable.

```
model1.res = resid(model1)

plot(crimes$Crime, model1.res, ylab="residuals", xlab="Crimes per 100,000")
abline(0, 0)
```



And finally, we can provide a prediction using the values provided in the assignment.

```
sample = data.frame(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)

 p = predict(model1, sample)
```

This prediction indicates that there would be about `155.43` crimes per 100,000 people in the population.