ISYE6501 Week 3 Questions 7 and 8

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?

I have remembered that at my previous job as a production planner, exponential smoothing is used in order to determine the following criteria in the products, and the data is used is the production of daily inventory:

- Whether or not the product would yield 90% of the produced values given from the ordered parts
- Double Exponential smoothing has been used to determine which products will be manufactured
- When the products are having an alpha value closer to zero, the product would be scrapped at any given time.
- Monthly, Daily, and Weekly production levels to determine if there is enough to justify a forecast.

Question 7.2

Loading the Libraries

```
library(smooth)
library(stats)
library(forecast)
```

Reading the Dataset

```
atl_temperature <- read.table("temps.txt")
head(atl_temperature)

##       V1   V2   V3   V4   V5   V6   V7   V8   V9  V10  V11  V12  V13
V14
## 1   DAY 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007
2008
## 2 1-Jul   98   86   91   84   89   84   90   73   82   91   93   95
85
## 3 2-Jul   97   90   88   82   91   87   90   81   81   89   93   85
87
## 4 3-Jul   97   93   91   87   93   87   87   87   86   86   93   82
91
```

```
## 5 4-Jul   90   91   91   88   95   84   89   86   88   86   91   86
90
## 6 5-Jul   89   84   91   90   96   86   93   80   90   89   90   88
88
##    V15  V16  V17  V18  V19  V20  V21
## 1 2009 2010 2011 2012 2013 2014 2015
## 2   95   87   92  105   82   90   85
## 3   90   84   94   93   85   93   87
## 4   89   83   95   99   76   87   79
## 5   91   85   92   98   77   84   85
## 6   80   88   90  100   83   86   84
```

Calculating the Average Daily Temperature

```
atl_avg  <- colMeans(atl_temperature[,2:21], na.rm = TRUE)
atl_avg
```

```
##        V2        V3        V4        V5        V6        V7
V8
##  99.13710  97.12097  99.69355  98.80645  99.48387  97.03226
99.05645
##        V9       V10       V11       V12       V13       V14
V15
##  96.97581  97.26613  98.85484  98.55645 100.89516  98.04032
96.54032
##       V16       V17       V18       V19       V20       V21
## 102.71774 100.80645 100.19355  97.24194  99.50806  98.87903
```

Converting to Time-Series Data

```
atl_ts <- ts(atl_avg)
atl_ts
```

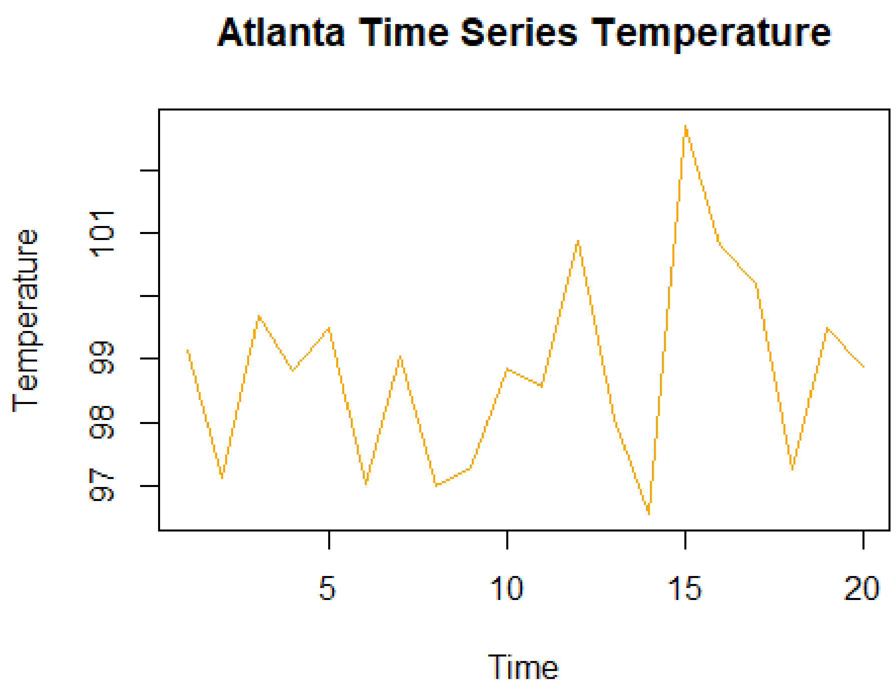
```
## Time Series:
## Start = 1
## End = 20
## Frequency = 1
##        V2        V3        V4        V5        V6        V7
V8
##  99.13710  97.12097  99.69355  98.80645  99.48387  97.03226
99.05645
##        V9       V10       V11       V12       V13       V14
V15
##  96.97581  97.26613  98.85484  98.55645 100.89516  98.04032
96.54032
##       V16       V17       V18       V19       V20       V21
## 102.71774 100.80645 100.19355  97.24194  99.50806  98.87903
```

Random Number Generator

```
set.seed(1609)
```

Plot the Average Time-Series Graph

```
plot(atl_ts, xlab="Time", ylab="Temperature",
     main = "Atlanta Time Series Temperature", col = "orange")
```



By looking at the average temperatures in Atlanta with the time-series data, it has seem that summer is ending later than expected. However, by looking at the graph, there has been some sharp dips in the later periods, indicating that temperature is cooling off, but it is not.

Exponential Smoothing

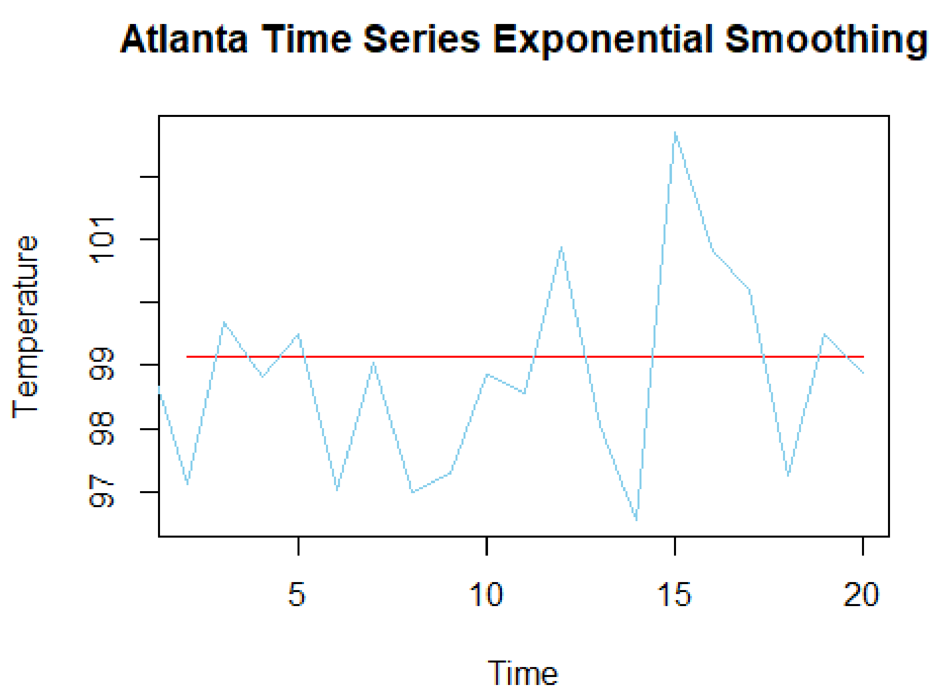
```
atl_ts_es <- HoltWinters(atl_ts, beta = FALSE, gamma = FALSE)
atl_ts_es

## Holt-Winters exponential smoothing without trend and without
seasonal component.
##
## Call:
## HoltWinters(x = atl_ts, beta = FALSE, gamma = FALSE)
##
## Smoothing parameters:
##  alpha: 6.610696e-05
##  beta : FALSE
##  gamma: FALSE
##
## Coefficients:
```

```
##      [,1]
## a 99.1367
```

Exponential Smoothing Plot

```
plot(atl_ts_es, xlab="Time",
     ylab = "Temperature",
     main = "Atlanta Time Series Exponential Smoothing",
     col = "skyblue")
```



By including the line to determine the average temperature in Atlanta, it has been given the clearer picture to understand where does the average occurs at using the exponential smoothing method. Although, starting from the 20th day, the predicted values indicate that it will start cooling down.

Fitted Data Information

```
atl_ts_es$fitted

## Time Series:
## Start = 2
## End = 20
## Frequency = 1
##        xhat    level
##  2 99.13710 99.13710
##  3 99.13696 99.13696
##  4 99.13700 99.13700
```

```
##  5 99.13698 99.13698
##  6 99.13700 99.13700
##  7 99.13686 99.13686
##  8 99.13686 99.13686
##  9 99.13671 99.13671
## 10 99.13659 99.13659
## 11 99.13657 99.13657
## 12 99.13653 99.13653
## 13 99.13665 99.13665
## 14 99.13658 99.13658
## 15 99.13641 99.13641
## 16 99.13664 99.13664
## 17 99.13675 99.13675
## 18 99.13682 99.13682
## 19 99.13670 99.13670
## 20 99.13672 99.13672
```

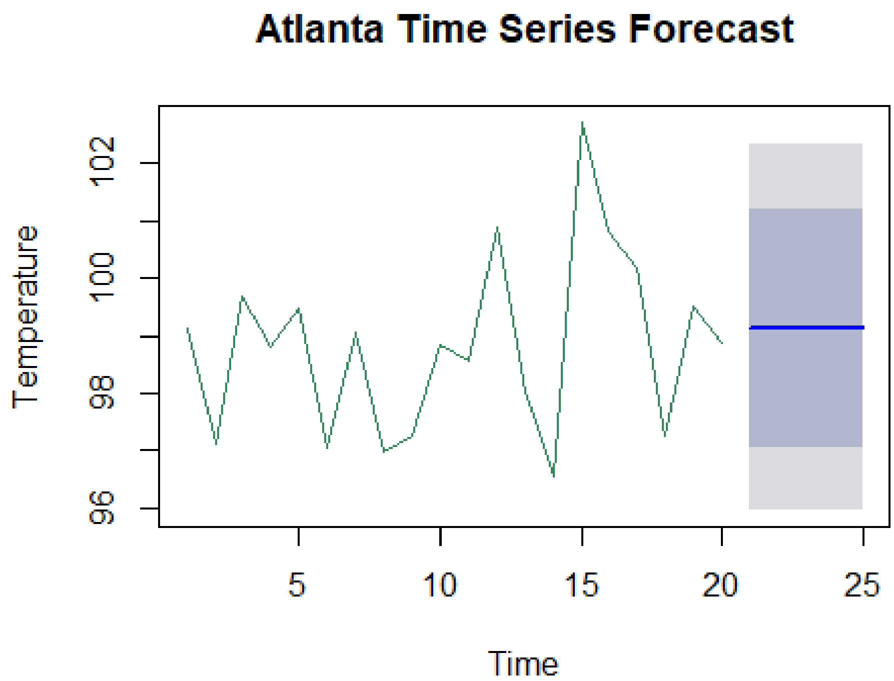
Forecast

```
atl_ts_forecast <- forecast:::forecast.HoltWinters(atl_ts_es, h=5)
atl_ts_forecast
```

```
##    Point Forecast    Lo 80   Hi 80    Lo 95    Hi 95
## 21        99.1367 97.05845 101.215 95.95829 102.3151
## 22        99.1367 97.05845 101.215 95.95829 102.3151
## 23        99.1367 97.05845 101.215 95.95829 102.3151
## 24        99.1367 97.05845 101.215 95.95829 102.3151
## 25        99.1367 97.05845 101.215 95.95829 102.3151
```

Forecast Plot

```
plot(atl_ts_forecast, xlab="Time",
     ylab="Temperature",
     main = "Atlanta Time Series Forecast",
     col = "seagreen")
```



By looking at the final plot in the forecasted version of the time-series data, it has been determined that when the Holt-Winters method was used, the temperatures do fall within the range of the original information. The values also do seem consistent as it is projected for the next period.

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.

I have done models in linear regression in relation to housing prices around the Bay Area based on the trends from the current events. There are many predictors that I have used to do the analyses, only to list a few. The list I have compiled follows:

- Sale price
- Square Feet Size
- City
- Bedrooms/Bathrooms
- Quality of Schools

Question 8.2

Reading the Crime Data

```
uscrime <- read.table("uscrime.txt", stringsAsFactors = FALSE, header =
TRUE)
head(uscrime)
```

```
##      M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq
## 1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1
## 2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4
## 3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0
## 4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7
## 5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4
## 6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6
##       Prob    Time Crime
## 1 0.084602 26.2011   791
## 2 0.029599 25.2999  1635
## 3 0.083401 24.3006   578
## 4 0.015801 29.9012  1969
## 5 0.041399 21.2998  1234
## 6 0.034201 20.9995   682
```

Setting a Random Number Generator

```
set.seed(1609)
```

Unscaled Regression Model

```
crime_lm <-lm(Crime~.,uscrime)
summary(crime_lm) # To take a look at the F-Statistic, R-Squared, and
P-Value
```

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

Testing the Given Information

```
uscrime_test <- 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 = 0.040,Time = 39.0)

uscrime_test

##    M So Ed Po1  Po2   LF M.F Pop  NW   U1  U2 Wealth Ineq Prob Time
## 1 14  0 10  12 15.5 0.64  94 150 1.1 0.12 3.6   3200 20.1 0.04   39

pred_uscrime_test <- predict(crime_lm, uscrime_test)
pred_uscrime_test

##        1
## 155.4349
```

As what it was said on the lecture, it turns out that having too many variables could lead to irrelevant factors to make a decision based on significance. The adjusted R-squared is at 0.8031, which turns out pretty good. However, there are only four factors that are significant though.

Concentrating on P-Values < 0.1

```
uscrime0.1 <-lm(Crime~M+Ed+Po1+U2+Ineq+Prob,uscrime)
summary(uscrime0.1)

##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data =
uscrime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -470.68  -78.41  -19.68  133.12  556.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -5040.50     899.84  -5.602 1.72e-06 ***
## M             105.02      33.30   3.154  0.00305 **
## Ed            196.47      44.75   4.390 8.07e-05 ***
## Po1           115.02      13.75   8.363 2.56e-10 ***
## U2             89.37      40.91   2.185  0.03483 *
## Ineq           67.65      13.94   4.855 1.88e-05 ***
## Prob        -3801.84    1528.10  -2.488  0.01711 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 200.7 on 40 degrees of freedom
## Multiple R-squared:  0.7659, Adjusted R-squared:  0.7307
## F-statistic: 21.81 on 6 and 40 DF,  p-value: 3.418e-11
```

For the 0.1 value, the R-squared model is at 0.7659, which is still considered good with a few variables removed. In addition, the variables are much more significant for the linear regression equation as it shows in the results.

Concentrating on P-Values < 0.05

```
uscrime0.05 <- lm(Crime~M+Ed+Ineq+Prob,uscrime)
summary(uscrime0.05)

##
## Call:
## lm(formula = Crime ~ M + Ed + Ineq + Prob, data = uscrime)
##
## 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
```

By filtering out the factors, the significance of variables have become much clearer to determine which ones are relevant and correlated to crime. However, with the R-squared value at 0.2629, the model seems to be a poor fit at the 0.05 level, and only two factors that are significant as well.

Scaling the Crime Regression Data

```
uscrime_scale <- uscrime

for (g in 1:15) {
  uscrime_scale[,g] <- (uscrime_scale[,g]-min(uscrime_scale[,g]))/
(max(uscrime_scale[,g])-

min(uscrime_scale[,g]))
  }
```

Running the Regression Model for the Scaled Crime Regression Data

```
crime_lm_scaled <- lm(Crime ~., uscrime_scale)
summary(crime_lm_scaled)

##
## Call:
## lm(formula = Crime ~ ., data = uscrime_scale)
##
## 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) -556.706    411.271  -1.354  0.18564
## M            509.415    241.940   2.106  0.04344 *
## So            -3.803    148.755  -0.026  0.97977
## Ed           659.135    217.309   3.033  0.00486 **
## Po1         2332.932   1283.927   1.817  0.07889 .
## Po2        -1269.294   1362.739  -0.931  0.35883
## LF          -106.876    236.626  -0.452  0.65465
## M.F          238.474    278.848   0.855  0.39900
## Pop         -120.946    212.777  -0.568  0.57385
## NW           177.008    272.846   0.649  0.52128
## U1          -419.551    303.141  -1.384  0.17624
## U2           637.639    312.877   2.038  0.05016 .
## Wealth       385.627    415.701   0.928  0.36075
## Ineq        1060.081    340.748   3.111  0.00398 **
## Prob        -548.179    256.560  -2.137  0.04063 *
## Time        -110.636    227.861  -0.486  0.63071
## ---
## 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
```

By looking at the scaled crime data regression equation, it has turned out that the equation is matching up to the original one that was given from the beginning of the analyses. The r-squared value is still the same as from the beginning. In addition, the p-values for all the regression analyses are all below 0.05.