ISYE6501 HW3

June 3, 2018

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 (the first smoothing parameter) to be closer to 0 or 1, and why?

Response: Medical data was an example in the lectures, but I would think it would be an applicable way to analyze the stock market for possible trends. The data could be time-series values of a stock price at regular intervals such as daily closing value. The smoothing parameter is going to be closer to 0, because stock prices have historically had a large degree of volatility in them.

Question 7.2

We read in the data, convert it to a time series, and run Holt-Winters on the resulting TS with multiplicative seasonality:

We obtain smoothing parameters $\alpha = 0.615003$, $\beta = 0$, and $\gamma = 0.5495256$. The value of α that the model gives us tells us that there is a nice balance between randomness and the previous observation in the data, but it weighs a bit more with regards to the previous observation. This makes sense because daily high temperatures would have a reasonably high correlation with the temperature of the day before it, generally speaking.

Now we are going to look at seasonality factors to see whether the end of summer date is changing. We convert the data back into a matrix, and then write it to an excel file:

Now we apply the same CUSUM analysis as in homework 2 to this workbook with the seasonality factors - please reference the included xlsx file for the implementation.. We need to choose smaller values of C and T to correspond with the smaller values in this seasonality data, so we choose $C = 0.05$ and $T = 0.5$. While I used OpenOffice instead of excel and as a result some of the formulas were not working properly, the important results are still evident. Our results show that summer is actually ending earlier, not later, in most years (see Figure 1).

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.

One standard example would be to predict housing prices based on factors such as the square footage of the house, number of bedrooms, number of bathrooms, and acreage of the property.

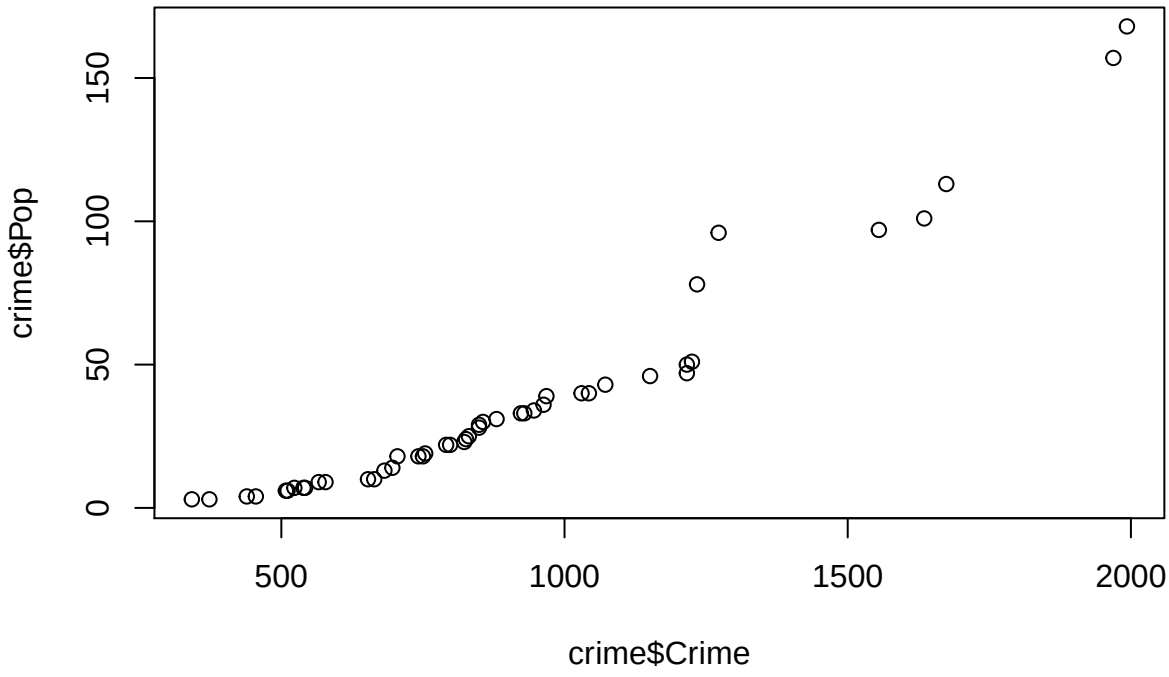
Year	1997	1998	1999	2000	2001	2002	2003	2004	2005
End of Summ.	28-Sept	28-Sept	27-Sept	24-Sept	23-Sept	24-Sept	23-Sept	23-Sept	22-Sept

2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
24-Sept	22-Sept	21-Sept	20-Sept	17-Sept	13-Sept	6-Sept	3-Sept	17-Aug	30-Aug

Figure 1: Yearly results via seasonality

Question 8.2

We read in the dataset, create a dataframe for the new data point, and create a QQplot for Crime vs Pop. Then we check Bartlett's test for unequal variance in these variables. Lastly, we create a simple linear model with Crime against all other variables as predictors, and then predict the new Crime value in a naive way:




```
##
##     Bartlett's Test of Homogenity of Variances
## ------------------------------------------------
## Ho: Variances are equal across groups
## Ha: Variances are unequal for atleast two groups
##
##         Data
##  --------------------
##  Variables: Crime Pop
##
##          Test Summary
##  -------------------------------
##  DF            =    1
##  Chi2          =    148.7898
##  Prob > Chi2   =    3.187733e-34
```

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

```
##        1
## 155.4349
```

Based on the QQ plot for Crime and Population, we question running OLS on these variables since nonlinearity is apparent. This is also reinforced by Bartlett's test, with a very close-to-0 p-value; heteroscedasticity is a problem in the dataset, at least for the variables we have looked at. Regardless of this we can easily compute the desired information. We attempt to use an automated selection process from the olsrr package to determine a subset of variables to use:

```
## [1] 16384
```

```
## # A tibble: 1 x 6
##   Index     N Predictors          `R-Square` `Adj. R-Square` `Mallow's Cp`
## * <int> <int> <chr>                    <dbl>           <dbl>         <dbl>
## 1 16384     8 M Ed Po1 M.F U1 U2~      0.789           0.744          4.24
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob,
##     data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -444.70 -111.07    3.03  122.15  483.30
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6426.10    1194.61  -5.379 4.04e-06 ***
```

```
## M              93.32      33.50   2.786  0.00828 **
## Ed            180.12      52.75   3.414  0.00153 **
## Po1           102.65      15.52   6.613 8.26e-08 ***
## M.F            22.34      13.60   1.642  0.10874
## U1          -6086.63    3339.27  -1.823  0.07622 .
## U2            187.35      72.48   2.585  0.01371 *
## Ineq           61.33      13.96   4.394 8.63e-05 ***
## Prob        -3796.03    1490.65  -2.547  0.01505 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 195.5 on 38 degrees of freedom
## Multiple R-squared:  0.7888, Adjusted R-squared:  0.7444
## F-statistic: 17.74 on 8 and 38 DF,  p-value: 1.159e-10
```

We test our two models for comparison using cross-validation:

```
## [1] 0.523
```

```
## [1] 0.693
```

```
## [1] 0.292
```

```
## [1] 0.628
```

We see that the true model quality in terms of the cross-validated Adjusted R^2 value is much higher for the second model with 8 predictors (0.628) instead of the first one with all of them (0.292).