HW-3

Responses & Summary

Question 7.1

Question: 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?

Answer:

I will consider Electricity consumption by weather change for this scenario.

Required inputs for that will be : Consumption per Kwh and Outside Temperature.

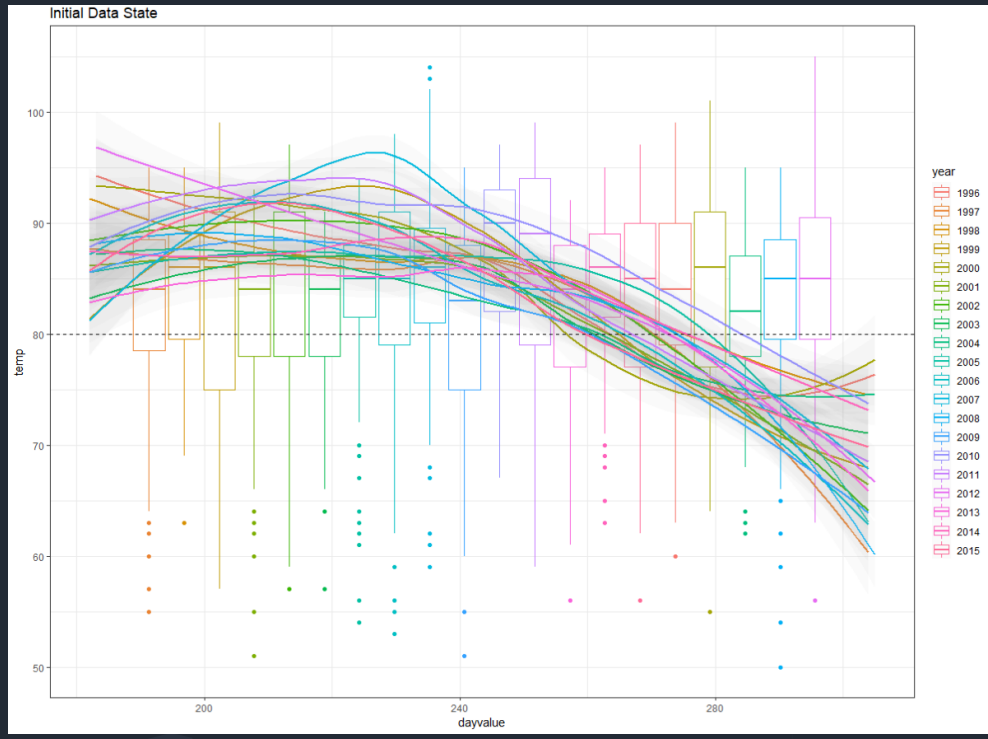
α will be set to 0.3, as there will still be some appliances drawing electricity irrespective of outside temperature. Hence not expecting much variation.

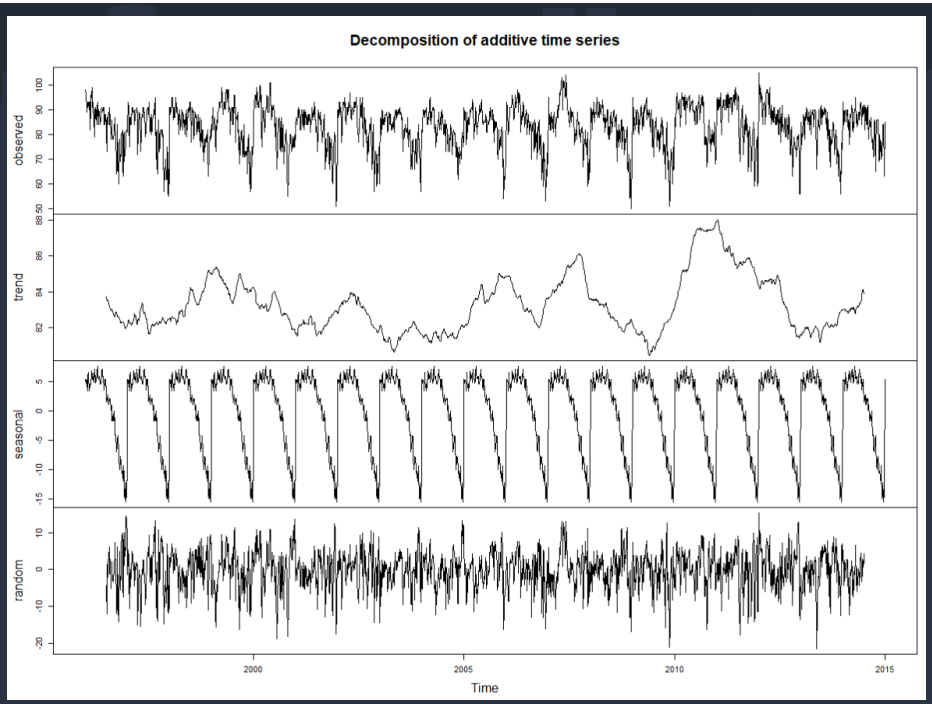
Question 7.2

- Reshaped data by melting in Year, Value of day and Temp value
  - Considered Day Value, as had issues with time series conversion with normal date.
  - Time series converted the date into a big number
- Analyzed the initial state of data (snapshot top right), and it was clear that there is some variation in summer end date with 80deg F as threshold.
- Converted data to time series and decomposed to check for trends, data shows regular seasonality.
- In term's of trend, this goes with last assignment findings that the summers are getting hotter.

Complete Repository available on GitHub:

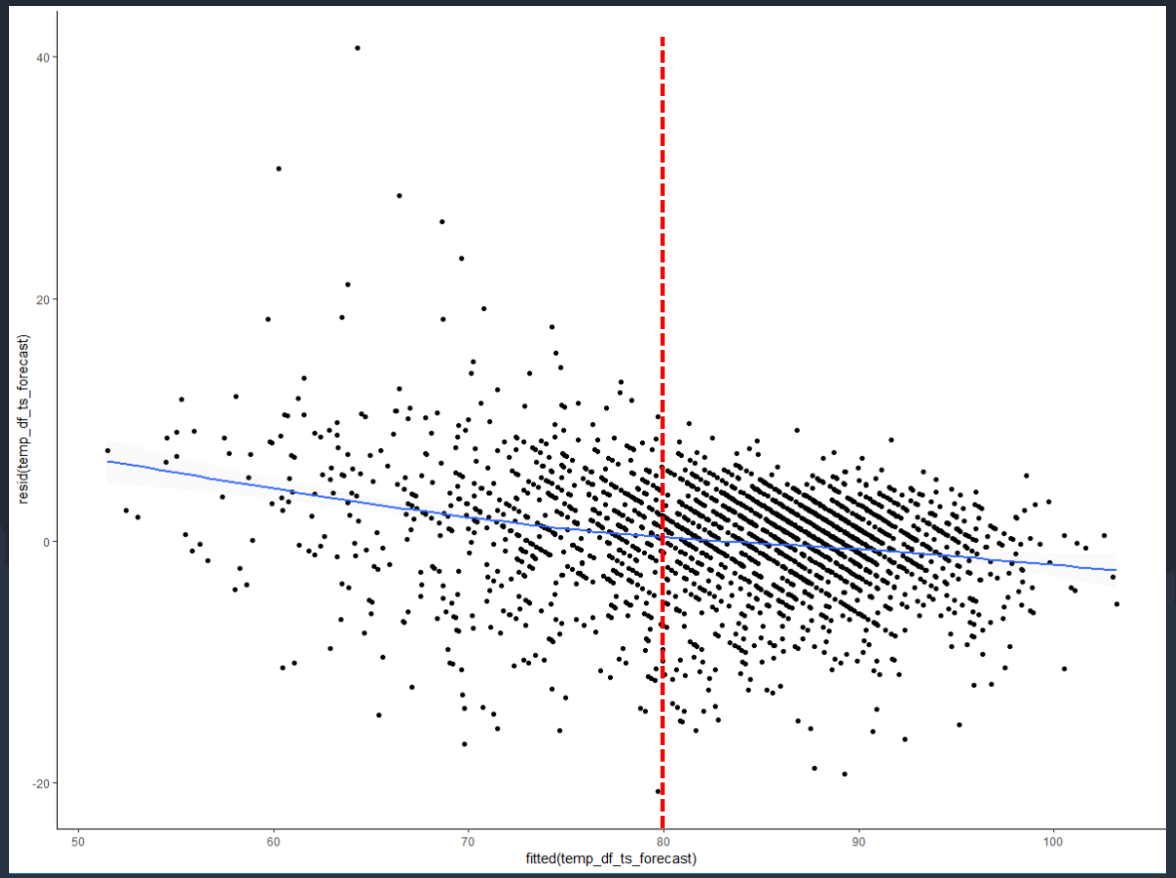
https://github.com/Hizzyth/GTX_Introduction-to-Analytics-Modelling





Question 7.2

- Using Holt winter method, model was created
- Use the model to forecast and check for variation between fitted value and residual)
- Residual difference on Threshold 80F shows that the summers are shifting.
- And combining the findings with previous graph, I conclude that summers are ending late.



https://github.com/Hizzyth/GTX_Introduction-to-Analytics-Modelling

QUESTION 8.1 : : Linear Regression

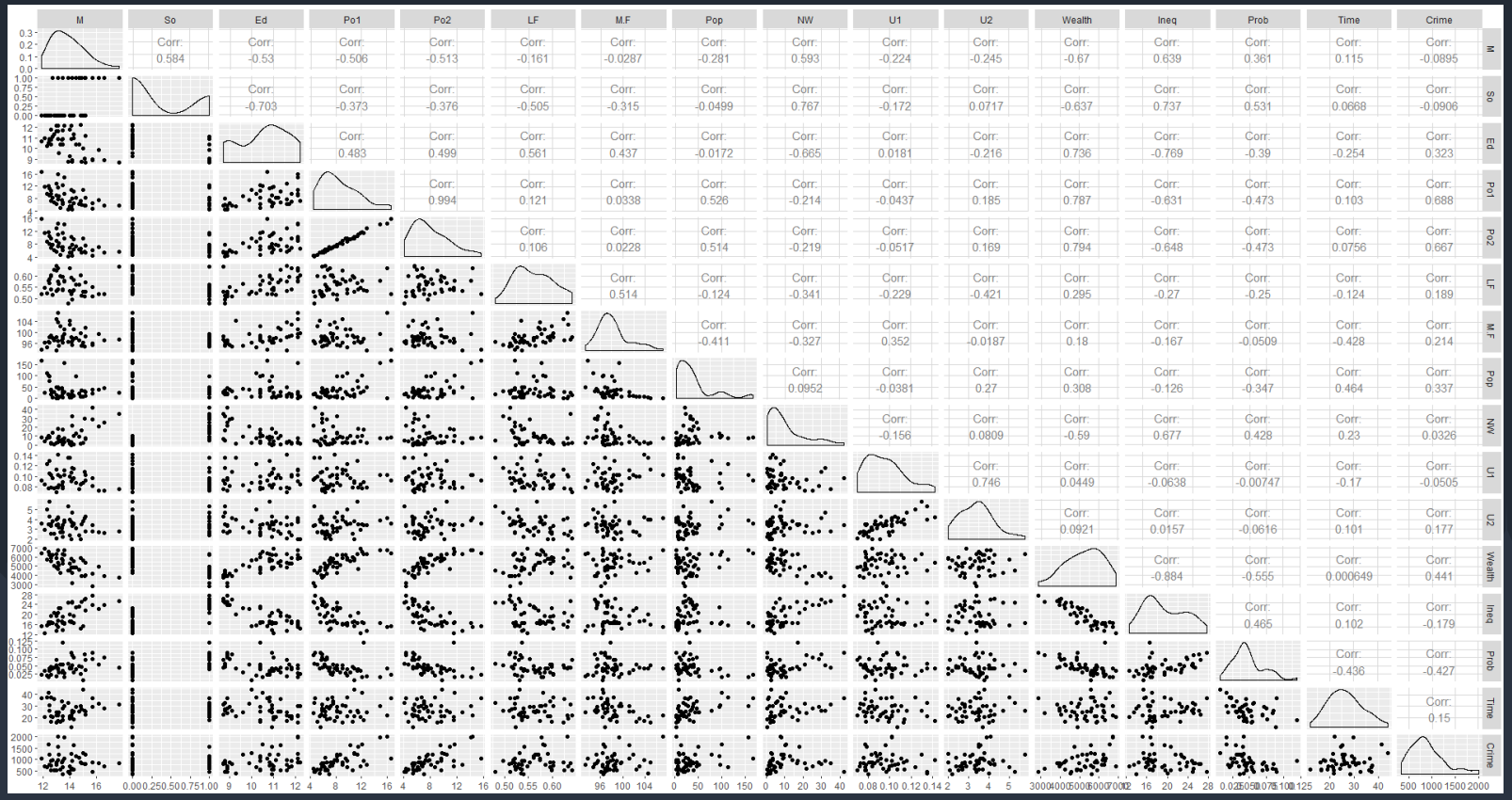
Question: Describe a situation or problem from your job, everyday life, current events, etc., for which a linear regression model would be appropriate. List some (up to 5) predictors that you might use.

Answer:

Tracking my weight is a good user case for linear regression model. Predictors used for that:

1. Calories burnt in gym
2. Hours spent in Office Meetings
3. Total commuting time
4. Calories intake on daily basis
5. Hours slept on daily basis

QUESTION 8.2 : Linear Regression (1)



- Explored Data for correlation

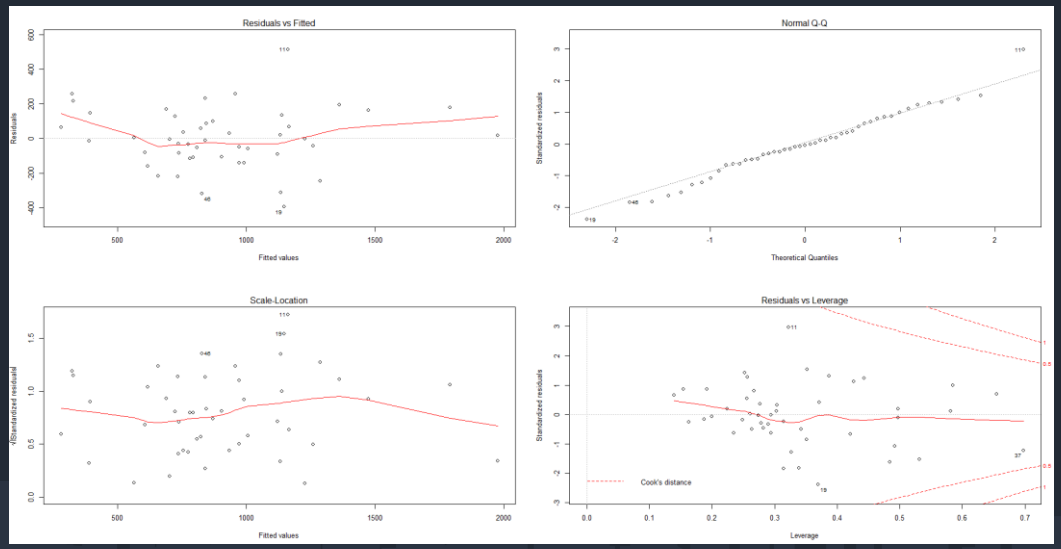
QUESTION 8.2 : Linear Regression (2)

```
lm(formula = Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop +
    NW + U1 + U2 + Wealth + Ineq + Prob + Time, data = crime_df)

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
```



- Model 1: Used all the predictors and found that p-value for some of them was high
- Hence decided to drop them and refine the model

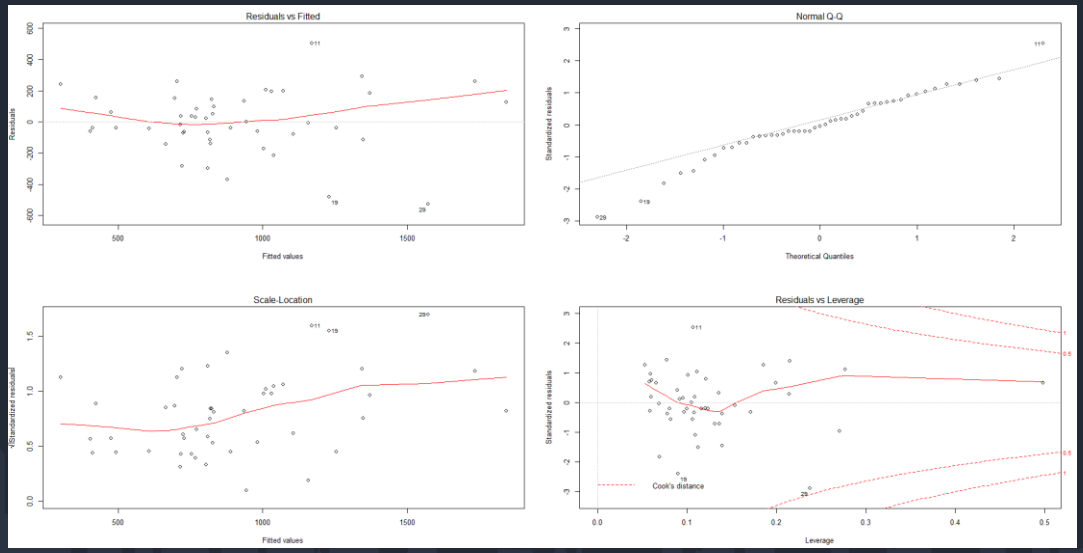
QUESTION 8.2 : Linear Regression (3)

```
Call:
lm(formula = Crime ~ M + Ed + Po1 + Ineq + Prob, data = crime_df)

Residuals:
   Min     1Q Median     3Q    Max
-528.2  -74.0   -7.0  139.8  503.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4064.57     816.28  -4.979 1.20e-05 ***
M              79.69      32.62   2.443 0.018964 *
Ed            160.15      43.42   3.688 0.000656 ***
Po1           121.23      14.06   8.621 9.47e-11 ***
Ineq           68.31      14.56   4.692 3.00e-05 ***
Prob        -3867.27    1596.55  -2.422 0.019930 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 209.7 on 41 degrees of freedom
Multiple R-squared:  0.7379,    Adjusted R-squared:  0.706
F-statistic: 23.09 on 5 and 41 DF,  p-value: 5.926e-11
```



- Model 2: Used only 5 predictors, and that provided a better correlation in terms of Residuals. However Q-Q plot was more aligned for Model 1.
- Further test set was run through both models

QUESTION 8.2 : Linear Regression (4)

- Crime Prediction result for:
  - Model 1 ~155
  - Model 2~1326

- Further used RSQ library, to calculate the R-Squared values and variability in data
- RSQ values for:
  - Model 1 ~ 0.803
  - Model 2 ~ 0.73

- Although both models shows good variability and a small random/other effects, model 2 looks to be better at predicting the crime rate. As value predicted by Model 1 is very small, and when we look at distribution of data the prediction of Model 2 >1000 looks more reasonable.