Data Analysis Project

Code

Dhirendra Khanka

19-June-2019

INTRODUCTION

This is a simple attempt in creating a Linear Regression Model to predict Temperature based on some explainatory Variables. Consider the following dataset.
Daily readings of the following air quality values for May 1, 1973 (a Tuesday) to September 30, 1973.
Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
Solar.R: Solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park
Wind: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport
Temp: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.

As you can see some values are missing from the dataset and it needs to be omitted.

	Ozone <int>	Solar.R <int>	Wind <dbl>	Temp <int>	Month <int>	Day <int>
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6

6 rows

Lets check to fit a linear model for tempature as predictor and Ozone, Solar and Wind as explainatory variables ### Omit Missing Records

	Ozone <int>	Solar.R <int>	Wind <dbl>	Temp <int>	Month <int>	Day <int>
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8

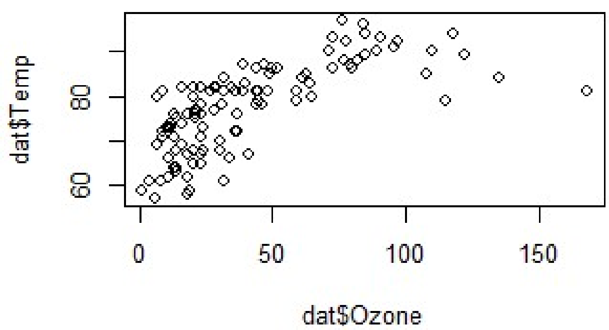
6 rows

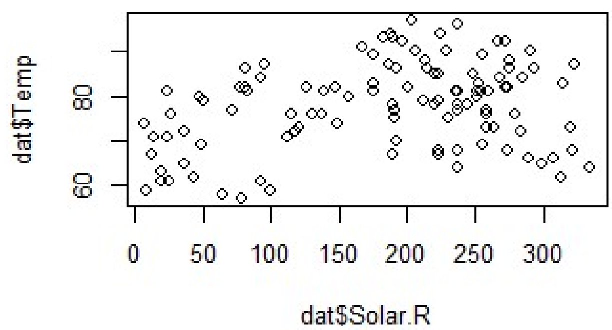
Data Plots of Explainatory Variables vs Predictor Variable

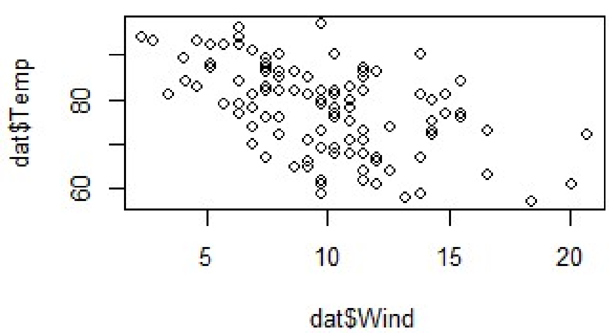
Plot of Ozone level vs Temparature. - Looks like positive correlation between the two.

Plot of Wind vs Temparture - There seems to be a somewhat negative correlation between Temparature and Wind.

Plot of Solar Radiation vs Temperature - There does not seem to be much co-relation between Solar Radiation and Temparture. The points are well scattered. We will include it for now and later compare it with another model wherein Solar Radiation is not included using the DIC.







JAGS MODEL 1

Lets build below Linear Model considering following explanatory variables affecting temparture.
Temperature ~ Ozone + Solar.R + Wind

Consider following JAGS model

temp ~iid N(mu, sig)
mu = Beta1 + Beta2 * Ozone + Beta3 * Solar.R + Beta4 * Wind
beta ~ N(1, 1e6)
sig ~ IG(1, 5)

Summary of Model1

Hide

```
summary(mod_sim)
```

```
Iterations = 1001:6000
Thinning interval = 1
Number of chains = 3
Sample size per chain = 5000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean       SD  Naive SE Time-series SE
b[1] 72.292704 3.213370 0.0262371      0.1556705
b[2]  0.172554 0.026577 0.0002170      0.0008177
b[3]  0.007419 0.007519 0.0000614      0.0001716
b[4] -0.315815 0.235626 0.0019239      0.0104468
sig   6.830316 0.474078 0.0038708      0.0043172

2. Quantiles for each variable:

          2.5%       25%       50%      75%    97.5%
b[1] 65.822261 70.133926 72.315869 74.56104 78.38286
b[2]  0.120561  0.154428  0.172755  0.19038  0.22463
b[3] -0.007646  0.002388  0.007524  0.01239  0.02211
b[4] -0.771567 -0.478918 -0.319962 -0.15524  0.15148
sig   5.978155  6.498992  6.799089  7.13193  7.83064
```

From summary, we find that the co-efficient for Solar Radiation is close to Zero.

DIC of Model1

Deviance Information Criterion

```
|
|                                                  |   0%
|
|*                                                 |   2%
|
|**                                                |   4%
|
|***                                               |   6%
|
|****                                              |   8%
|
|*****                                             |  10%
|
|******                                            |  12%
|
|*******                                           |  14%
|
|********                                          |  16%
|
|*********                                         |  18%
|
|**********                                        |  20%
|
|***********                                       |  22%
|
|************                                      |  24%
|
|*************                                     |  26%
|
|**************                                    |  28%
|
|***************                                   |  30%
|
|****************                                  |  32%
|
|*****************                                 |  34%
|
|******************                                |  36%
|
|*******************                               |  38%
|
|********************                              |  40%
|
|*********************                             |  42%
|
|**********************                            |  44%
|
|***********************                           |  46%
|
|************************                          |  48%
|
|*************************                         |  50%
```

```
  |
  |**************************                        |  52%
  |
  |***************************                       |  54%
  |
  |****************************                      |  56%
  |
  |*****************************                     |  58%
  |
  |******************************                    |  60%
  |
  |*******************************                   |  62%
  |
  |********************************                  |  64%
  |
  |*********************************                 |  66%
  |
  |**********************************                |  68%
  |
  |***********************************               |  70%
  |
  |************************************              |  72%
  |
  |*************************************             |  74%
  |
  |**************************************            |  76%
  |
  |***************************************           |  78%
  |
  |****************************************          |  80%
  |
  |*****************************************         |  82%
  |
  |******************************************        |  84%
  |
  |*******************************************       |  86%
  |
  |********************************************      |  88%
  |
  |*********************************************     |  90%
  |
  |**********************************************    |  92%
  |
  |***********************************************   |  94%
  |
  |************************************************  |  96%
  |
  |************************************************* |  98%
  |
  |**************************************************| 100%
Mean deviance:  742.6
penalty 4.954
Penalized deviance: 747.6
```

JAGS Model 2

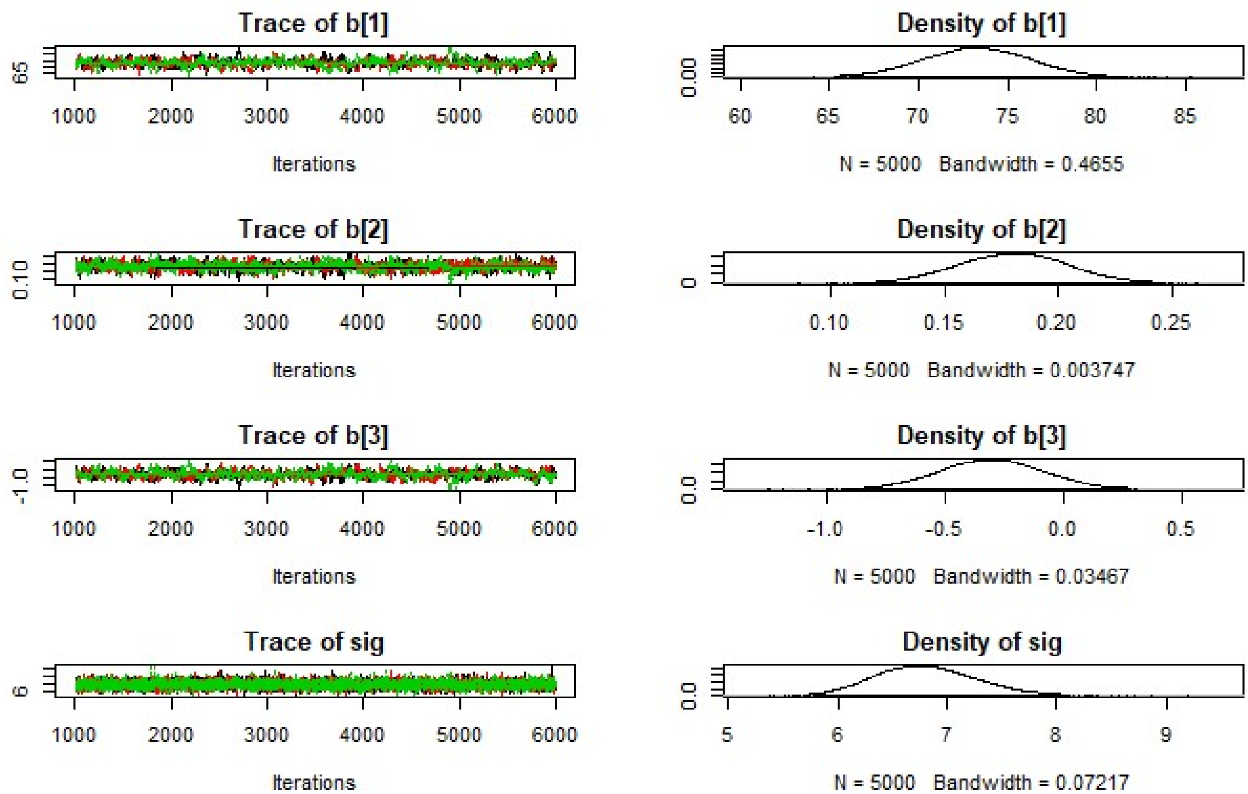
Lets now build 2nd Model with following

temp ~iid N(mu, sig)
mu = Beta1 + Beta2 * Ozone + Beta3* Wind
beta ~ N(1, 1e6)
sig ~ IG(1, 5)

Convergence Diagnostic Plots

Hide

```
plot(mod2_sim)
```



Hide

```
gelman.diag(mod2_sim)
```

```
Potential scale reduction factors:

     Point est. Upper C.I.
b[1]       1.01       1.01
b[2]       1.00       1.00
b[3]       1.01       1.01
sig        1.00       1.00

Multivariate psrf

1
```

Summary of Model2

Hide

```
summary(mod2_sim)
```

```


Iterations = 1001:6000
Thinning interval = 1
Number of chains = 3
Sample size per chain = 5000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

        Mean      SD  Naive SE Time-series SE
b[1] 73.2408 3.03131 0.0247506      0.1558508
b[2]  0.1798 0.02419 0.0001975      0.0008953
b[3] -0.3038 0.22551 0.0018412      0.0107987
sig   6.8162 0.46587 0.0038038      0.0039236

2. Quantiles for each variable:

        2.5%     25%     50%     75%   97.5%
b[1] 67.2079 71.2394 73.2514 75.2662 79.1237
b[2]  0.1326  0.1634  0.1800  0.1962  0.2274
b[3] -0.7370 -0.4542 -0.3025 -0.1542  0.1430
sig   5.9829  6.4874  6.7920  7.1135  7.7974
```

DIC of model2

Hide

```
dic.samples(mod2, n.iter = 1e3)
```

```
|
|                                                  |   0%
|
|*                                                 |   2%
|
|**                                                |   4%
|
|***                                               |   6%
|
|****                                              |   8%
|
|*****                                             |  10%
|
|******                                            |  12%
|
|*******                                           |  14%
|
|********                                          |  16%
|
|*********                                         |  18%
|
|**********                                        |  20%
|
|***********                                       |  22%
|
|************                                      |  24%
|
|*************                                     |  26%
|
|**************                                    |  28%
|
|***************                                   |  30%
|
|****************                                  |  32%
|
|*****************                                 |  34%
|
|******************                                |  36%
|
|*******************                               |  38%
|
|********************                              |  40%
|
|*********************                             |  42%
|
|**********************                            |  44%
|
|***********************                           |  46%
|
|************************                          |  48%
|
|*************************                         |  50%
```

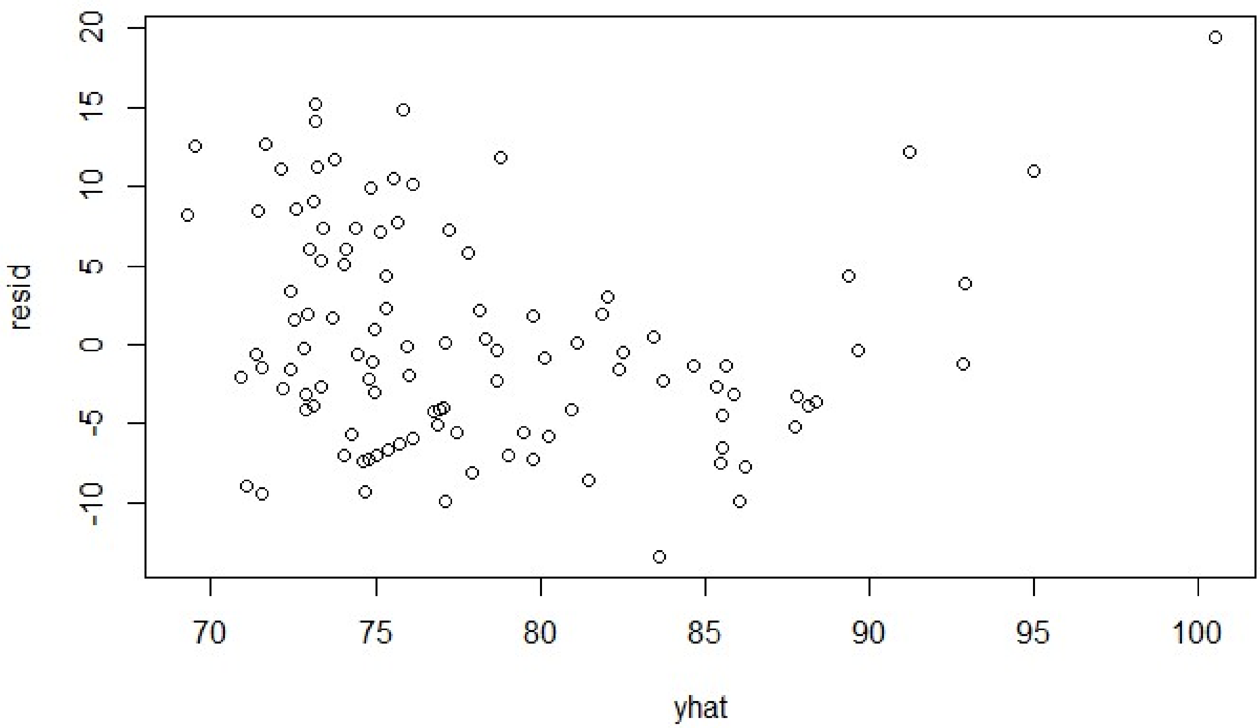
```
  |
  |**************************                        |  52%
  |
  |***************************                       |  54%
  |
  |****************************                      |  56%
  |
  |*****************************                     |  58%
  |
  |******************************                    |  60%
  |
  |*******************************                   |  62%
  |
  |********************************                  |  64%
  |
  |*********************************                 |  66%
  |
  |**********************************                |  68%
  |
  |***********************************               |  70%
  |
  |************************************              |  72%
  |
  |*************************************             |  74%
  |
  |**************************************            |  76%
  |
  |***************************************           |  78%
  |
  |****************************************          |  80%
  |
  |*****************************************         |  82%
  |
  |******************************************        |  84%
  |
  |*******************************************       |  86%
  |
  |********************************************      |  88%
  |
  |*********************************************     |  90%
  |
  |**********************************************    |  92%
  |
  |***********************************************   |  94%
  |
  |************************************************  |  96%
  |
  |************************************************* |  98%
  |
  |**************************************************| 100%
Mean deviance:  742.5 
penalty 4.203 
Penalized deviance: 746.7 
```

Residual and Q-Q Plot

Lets Plot residuals for the 2nd model with lower DIC. The residual plots does not convey any pattern and looks fine. Although there is visible variance throughout the plot, which means the predictions are not great from actual.

Hide

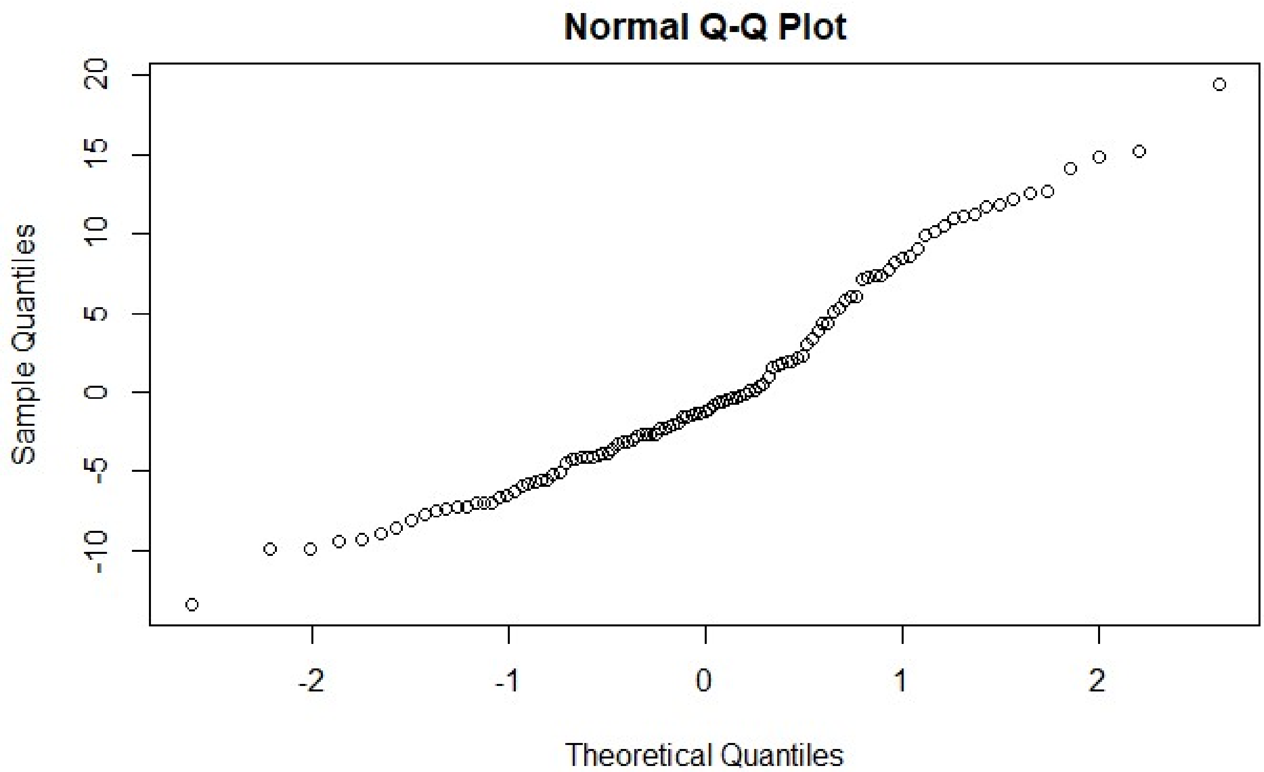
```
coeff = coefficients(mod2)
yhat = coeff$b[1] + coeff$b[2]* data2_jags$ozone + coeff$b[3]* data2_jags$wind
resid = yhat - data2_jags$y
plot(yhat, resid)
```



The QQ plot also looks okay except for 1 far off point.

Hide

```
qqnorm(resid)
```



CONCLUSION

From the two JAGS model which we build, the 2nd model is only marginally better than the first in terms of DIC. Overall the co-efficients remain largely unaffected after removing the Solar Radiation explainatory Variable. Due to simple order of the fit, the model is also modest in prediction as visible from the variation in residual Plot.