Linear Regression JAGS model for predicting temperature

Author: Dhirendra Khanka
Date: 19-Jun-2019

Summary

This is a simple attempt in creating a JAGS Linear Regression Model to predict Temperature based on some explanatory Variables. We are going to use the training dataset only and no separate test set for this assignment. Using residuals, we will get an idea on how good the model is fitting the data set.

Consider the following dataset. Daily readings of the following air quality values for May 1, 1973 (a Tuesday) to September 30, 1973. Records containing missing values are omitted.

- Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
- Solar.R: Solar radiation 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.

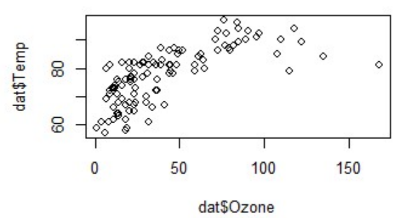
	Ozone <int>	Solar.R <int>	Wind <dbl>	Temp <int>	Month <int>	Day <int>
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8

6 rows

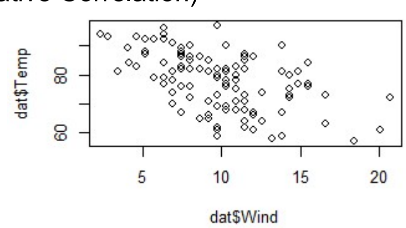
Plotting Data

Data Plots of Explanatory Variables vs Predictor Variable

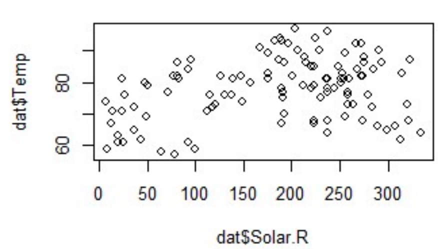
- Plot of Ozone level vs Temperature. (Positive Correlation)



- Plot of Wind vs Temperature (Negative Correlation)



- Plot of Solar Radiation vs Temperature (No Correlation)



JAGS MODEL 1

Let's build below Linear Model considering following explanatory variables affecting temperature.

$Temparature \sim Ozone + Solar + Wind$

Consider following JAGS model

$$\boldsymbol{Temp \ \widetilde{\iota\iota d} \ Normal(\mu, \sigma^2)}$$
$$\boldsymbol{\mu = \beta_1 + \beta_2 * Ozone + \beta_3 * Solar + \beta_4 * Wind}$$
$$\boldsymbol{\beta \sim Normal(1, 1e^6)}$$
$$\boldsymbol{\sigma^2 \sim InverseGamma(5/2, 25)}$$

A non-informative prior with mean 0 and variance 1e6 for the Beta Co-efficients. Lets assume prior effective sample size of 5 and Prior point estimate for sig2 as 20, which gives the first parameter as 5/2 and 2nd parameters as 5*20/2 = 25

The model was run for an initial burn in period of 1000 samples and then updated for further 5000 samples. The model reported following DIC.

```
----------------------------
Mean deviance:  742.7
penalty 5.064
Penalized deviance: 747.7
----------------------------
```

The model Summary is as below

```
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] 72.481373 3.252614 2.656e-02      0.1628912
b[2]  0.171436 0.026033 2.126e-04      0.0008837
b[3]  0.007234 0.007686 6.276e-05      0.0001789
b[4] -0.325910 0.234577 1.915e-03      0.0107327
sig   6.755065 0.456674 3.729e-03      0.0038624
```

CONCLUSION for MODEL1

We did not explore the convergence diagnostics or the residual analysis yet for this model. The 3rd Beta Co-efficient is close to 0, which is highly suggestive that Solar Radiation has insignificant effect on the temperature. We now build another iteration for the model by simply taking out this explanatory variable from the model.

JAGS MODEL 2

Let's build 2nd iteration of the model by taking out the Solar Radiation Variable.

$$Temparature \sim Ozone + Wind$$

Consider following JAGS model

$$\boldsymbol{Temp} \ \ \widetilde{\boldsymbol{\iota\iota d}} \ \ \boldsymbol{Normal}(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$$
$$\boldsymbol{\mu} = \boldsymbol{\beta_1} + \boldsymbol{\beta_2} * \boldsymbol{Ozone} + \boldsymbol{\beta_3} * \boldsymbol{Wind}$$
$$\boldsymbol{\beta} \sim \boldsymbol{Normal}(\boldsymbol{1}, \boldsymbol{1e^6})$$
$$\boldsymbol{\sigma^2} \sim \boldsymbol{InverseGamma}(\boldsymbol{5/2}, \boldsymbol{25})$$

A non-informative prior with mean 0 and variance 1e6 for the Beta Co-efficients. Lets assume prior effective sample size of 5 and Prior point estimate for sig2 as 20, which gives the first parameter as 5/2 and 2nd parameters as 5*20/2 = 25

The model was run for an initial burn in period of 1000 samples and then updated for further 5000 samples.

DIC

```
--------------------------
Mean deviance:  742.4
penalty 3.891
Penalized deviance: 746.3
--------------------------
```

Summary

```
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] 73.2222 3.08362 0.0251776      0.1561256
b[2]  0.1802 0.02458 0.0002007      0.0008806
b[3] -0.3031 0.22970 0.0018755      0.0109432
sig   6.7641 0.45753 0.0037357      0.0038154
```

Let's have a look at the Convergence Diagnostics

Gelman Rubin Diag

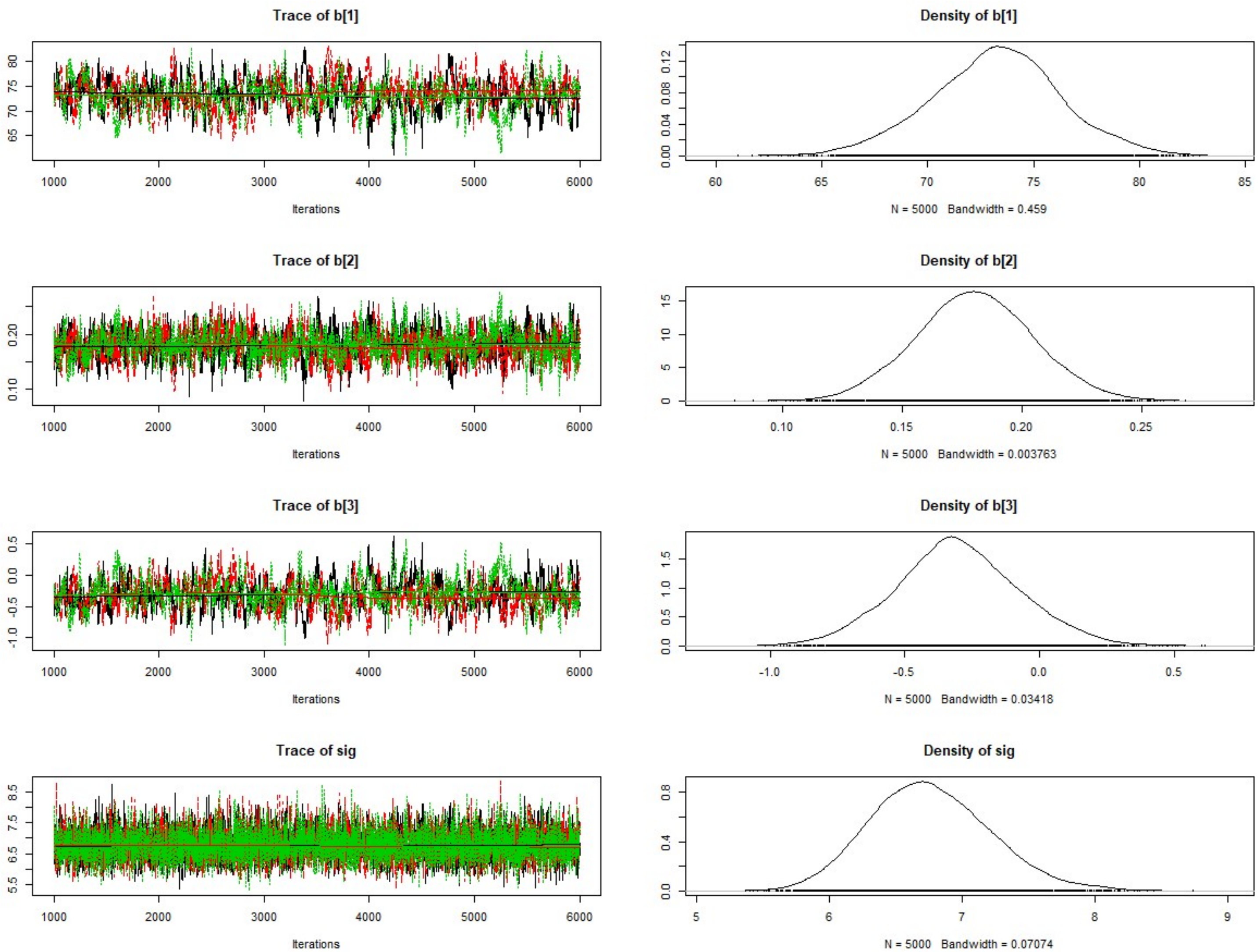
```
Potential scale reduction factors:

     Point est. Upper C.I.
b[1]       1.05       1.17
b[2]       1.03       1.09
b[3]       1.05       1.16
sig        1.00       1.00

Multivariate psrf

1.04
```

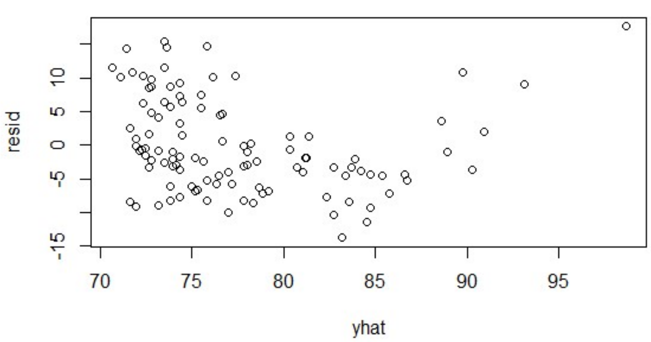
Trace Plots



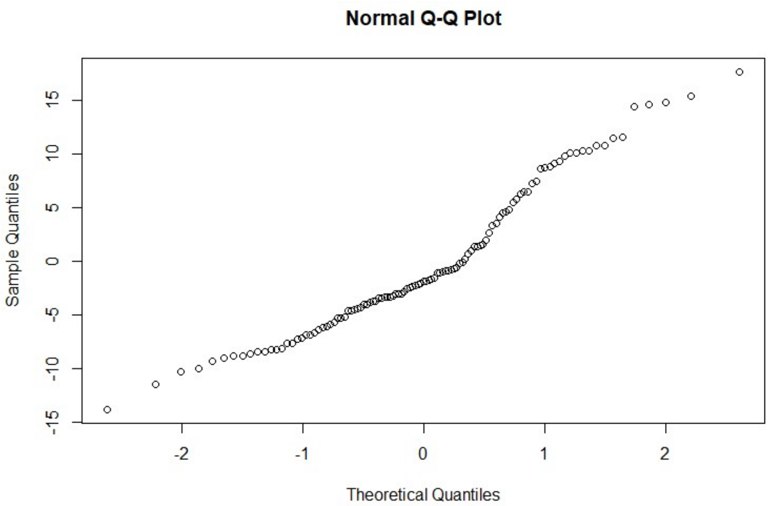
Residual Plots

Lets Plot residuals for the 2nd model with lower DIC.

```
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)
```



QQ Plot



CONCLUSION

The residual plots do not convey any pattern and looks fine. Although there is visible variance throughout the plot, which means the performance of the model is not that great.

The QQNorm plot is also acceptable as it is following a straight line though not perfect.

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 explanatory Variable. Due to simple order of the fit, the model is also modest in prediction as visible from the variation in residual Plot.