ST406 FINAL PROJECT REPORT
BAYESIAN vs FREQUENTIST LINEAR REGRESSION ANALYSIS OF METEOROLOGICAL DATA

Pruthuvi Kehelbedda
S/15/359
Department of Statistics and Computer Science
Faculty of Science
University of Peradeniya
Sri Lanka

2021-01-15

Contents

List of Figures

List of Tables

1 Introduction

Bayesian Analysis, within the last decade more practitioner, especially in some fields such as medicine and psychology, are turning towards Bayesian analysis since almost everything can be interpreted straightforwardly with a probabilistic manner. However, Bayesian analysis has also some drawbacks, like a subjective way to define the priors (which plays an important role in Bayesian analysis, to find posterior), or the problems that do not have conjugate priors, and not always the mcmc algorithms converge easily to the right values when we use complex data (P.murphy 2012).

In this article, we aim to do a Bayesian Regression Model fitting to predict a predictor variable in our data set. Which is important because it will lead to results which have been not studied before and to answer some major questions that have not been answered (ZARAH 2020).

When we discuss Bayesian analysis, frequentist statistic inference is a field that we can not ignore. There is a huge question is which statistical methodology is best, frequentist or Bayesian (Bowles 2020). In a frequentist model, it's only using the data from the current experiment when evaluating outcomes. When we apply frequentist statistics we will likely use the term called "p-value".

The simple evaluation is the smaller the p-value, the more statistically significant our results. Also, p-value means the probability of a false positive based on the data in the experiment. Usually, it does not tell us two things. First, the probability of a specific event actually happening and the second is the probability of a variant is better than the control.

In Bayesian statistics means "Probability is an orderly opinion, and that inference from data is nothing other than the revision of such opinion in the light of relevant new information" (NSS 2016). This believed by Thomas Bayes. The simple meaning is to get the prediction of the next experiment based on today results. So that why in this article we provide frequentist statistical modelling as well.

2 Methodology

First, we discuss exploratory data analysis, and then we go for the frequentist linear regression modelling part and finally, we do Bayesian linear regression modelling for the data. And from this, we will get some knowledge that you've got solar energy batteries, and you want to know will it be reasonable to use them in the future. We will discuss these methodologies when we go further with the article. Now let's discuss the data.

2.1 Description of Data

Figure 1: Hawaii Space Exploration Analog and Simulation

The data set is meteorological data from the HI-SEAS (Hawaii Space Exploration Analog and Simulation) (Hawai'i, n.d.) weather station from four months (September through December 2016) between Mission IV and Mission V from NASA. Our plan is to fit a prediction model using the knowledge of frequentist and Bayesian statistics. Meteorological data is facts that affect military operations such as wind direction, wind speed, air density and other phenomena.

Table 1 shows data set is containing those features, and the data set is already cleaned no need any data pre-processing to do. Our model response variable will be Radiation in data set.

Table 1: Data Set

Variable	Description
Date	Date in format of yyyy-mm-dd
Time	The local time in the format of hh:mm:ss 24-hr
Radiation	Solar radiation in watts per meter squared (1kg/sˆ3)
Temperature	Temperature in degrees fahrenheit (degree F)
Pressure	Barometric Pressure in Hg
Humidity	Humidity precent
WindDirection	Wind derection in degrees
WindSpeed	Wind speed in miles per hour (mph)
TimeSunRise	Hawaii time of Sun rise
TimeSunSet	Hawaii time of Sun set

Table 2: First five observation of the data set

Date	Time	Rad	Temp	Pressure	Humidity	WindDir	WindSpeed	TimeSR	TimeSS
2016-09-29	23:55:26	1.21	48	30.46	59	177.39	5.62	06:13:00	18:13:00
2016-09-29	23:50:23	1.21	48	30.46	58	176.78	3.37	06:13:00	18:13:00
2016-09-29	23:45:26	1.23	48	30.46	57	158.75	3.37	06:13:00	18:13:00
2016-09-29	23:40:21	1.21	48	30.46	60	137.71	3.37	06:13:00	18:13:00
2016-09-29	23:35:24	1.17	48	30.46	62	104.95	5.62	06:13:00	18:13:00

2.2 Exploratory Data Analysis

To understand the main characteristics of the data in the form of visual methods we use EDA (Exploratory Data Analysis). Here we mainly examine what the data can tell us beyond the formal modelling or hypothesis testing task. In our analysis, graphical techniques used for EDA (From Wikipedia 2020).

2.3 Linear Regression Analysis

A linear approach to modelling the relationship between a response and one or more explanatory variables (also known as dependent and independent variables). In LR, relationships are moded using a linear predictor function and that will estimate the unknown model parameters. Depending on the number of explanatory variables we use in the model classify LR to two types, as Simple Linear Regression and Multiple Linear Regression. Since our goal is prediction, Linear Regression can be used to fit a predictive model to our data in the form of frequentist statistical method (Prabhakaran 2019).

First, we fit a simple linear regression model to our data and validate the results and identify the best-fitted model. Then we fit a multiple linear regression model to identify the best multiple linear model.

Consider our data is $\{y_i, x_{i1}, x_{i2}, ..., x_{ip}\}$ where $i = 1, ..., n$. Thus the model is in the form,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i$$

$$\Rightarrow y_i = x_i^T \beta + \epsilon_i \; ; where \; i = 1, 2, ..., n$$

linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors x is linear ("Simple Linear Regression Numerical Example," n.d.).

- Assumption for the linear models never perfectly met,
- But we must check if they are reasonable enough to work with.
- Assumptions,
  - 1. Y-values (or the error e) are independent.

- 2. Y-values (or the error e) are normally distributed.
- 3. Y-values can express by a linear function of X-variable.
- 4. Homoscedasticity: variation of observation around the regression line is constant.

Based on these we fit a simple linear regression model in R using lm function. Then using forward selection we got the best simple linear regression model. Which done by based on the AIC (Akaike information criterion) value (Douglas Montgomery and Vining, n.d.). And also evaluate the regression assumption by plotting the residual vs fitted, normal Q-Q plots. Similarly, we perform multiple linear regression. And choose the best-fitted model and clarify the assumptions.

2.4 Bayesian Regression Analysis

$$\overbrace{p(\theta/D)}^{Posterior} = \frac{\overbrace{p(D/\theta)}^{Likelihood} . \overbrace{p(\theta)}^{Prior}}{\underbrace{p(D)}_{Evidence}}$$

Above equation called Baye's rule When we know prior information about data from that we make current information more valuable, which is called Posterior (Kruschke 2010).

We have used two approaches to perform bayesian regression in R. Which are stan_glm and bas.lm. These are functions that in rstanarm and BAS packages respectively (Streftaris and Gibson 2004). The best fitted model is chooses by identifying minimum BIC value.

Let's discuss about the prior. If we have the domain knowledge and the previous any guess about the model parameters, we can include these information in to our model. Unlike in frequentist, assume we about every information on parameters comes from the data. If we do not have any information about prior, we can use normal distribution like non-informative priors to the model.

If we talk about the posterior, results The result of performing Bayesian Linear Regression is a distribution of possible model parameters based on the data and the prior.

3 Results and discussion

3.1 EDA

First, let's discuss the Exploratory part. In that, we tried to give a visual representation to the data.

Radiation	Temperature	Pressure	Humidity	WindSpeed
Min. : 1.11	Min. :34.0	Min. :30.19	Min. : 8.00	Min. : 0.000
1st Qu.: 1.23	1st Qu.:46.0	1st Qu.:30.40	1st Qu.: 56.00	1st Qu.: 3.370
Median : 2.66	Median :50.0	Median :30.43	Median : 85.00	Median : 5.620
Mean : 207.12	Mean :51.1	Mean :30.42	Mean : 75.02	Mean : 6.244
3rd Qu.: 354.24	3rd Qu.:55.0	3rd Qu.:30.46	3rd Qu.: 97.00	3rd Qu.: 7.870
Max. :1601.26	Max. :71.0	Max. :30.56	Max. :103.00	Max. :40.500

Figure 2: R Summary Description of Numerical data

We can see that Solar Radiation level is in a range of $1.11 kg/s^3$ to $1601.26 kg/s^3$. Also Temperature is in the range of $34°F$ to $71°F$. Pressure is almost similar within all the data points and Humidity is also having a range between 8% and 103%. We can see wind speed is oscillation between 0mph and 40.5mph.

Since our response variable is Radiation we will look at it.

We can see from the Figure 3 Radiation has a minimum $1.11 kg/s^3$ value and a maximum of $1601.26 kg/s^3$. Radiation got a bell shape curve with respec to Time variable. And Radiation got it's highest value when the date is 2016-09-04, time is 12:15:04 PM and at that time outer temperature is $61°F$, Humidity is 93%, Barometric Pressure is 30.47Hg, wind direction is clockwise from 3.56 degrees North, and wind speed is 9mph.

Figure 4 explains the box plot (Team 2018) of daily radiation level by each month. Each month median radiation levels are different. Median radiation is highest in October, and It's increasing toward from September to October, but then it's decreasing from October to December. The daily radiation level is getting lower from September to December. In September and December, we can see a similar high dispersion of daily radiation that in October and November. The range of daily radiation is high to low in September, December, October, and November respectively.

From Figure 5 we can get an idea about radiation level and how radiation behaves with respecting to months and day time.

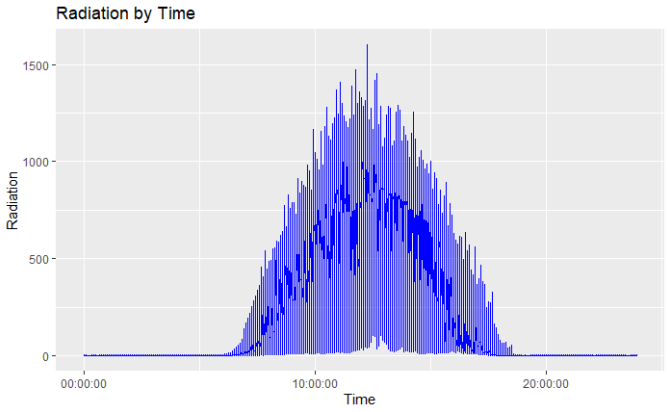


Figure 3: Radiation by Time

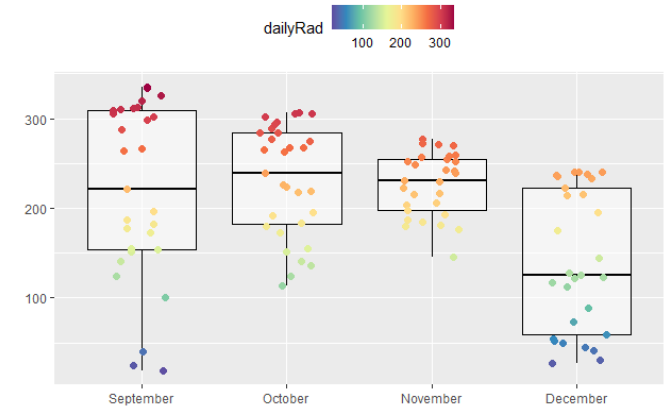


Figure 4: Box Plot of Daily Radiation level by Months

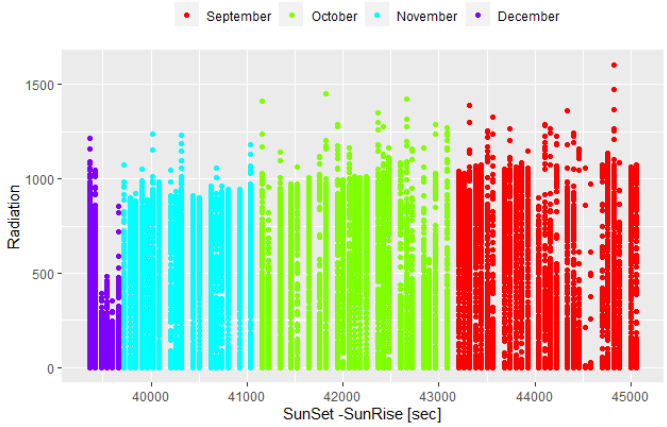


Figure 5: Radiation in a Day time w.r.t each month

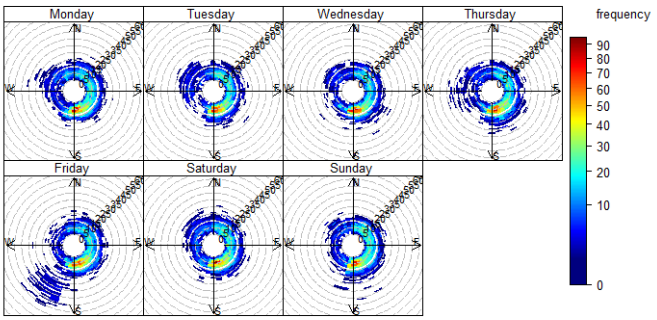


Figure 6: Polar frequency plot of Wind direction & speed on week days

Figure 6 shows us how the Polar frequency plot (Brandon Feenstra 2020) of Wind direction & speed change in weekdays. Most of the time higher frequency wind is directed towards SE.

3.1.1 Correlation

Our main purpose of the study is fitting a model to predict the radiation level. So to select variables who are really correlated with radiation we plot a correlation plot.

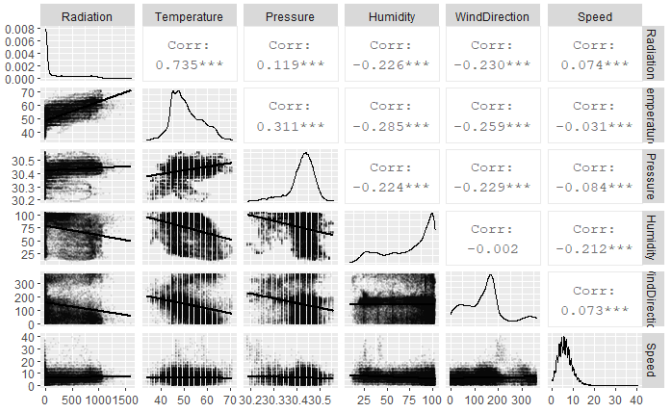


Figure 7: Correlation plot

From the above figures we can see, high positive correlation between Radiation & Temperature. Pressure & Temperature has positive low correlation. Humidity with Radiation, Temperature, and Pressure and WindDirection with Radiation, Temperature, and Pressure and Speed with Humidity are all having a same negative lower correlation.

3.2 Regression Model

We fitted a simple linear regression model. And depending on the AIC value we have chosen the best model as

$$\hat{Radiation} = (-1706.2876) + (37.4421) * Temperature + \epsilon$$

Above summary output gives us these results,

- Looking at the residuals we can say it is do not appear to be strongly symmetrical.
- From the coefficient estimates 1 $^{\circ}F$ increment of Temperature, will result 37.4421 kg/s^3(watts per meter squared) increment of Radiation.
- Depending on the p-values we can say variables are statistically significant.
- Temperature can explain 54.02% of the variation in Solar Radiation.
- Considering finally F-stat and p-value, the model is significant.

For the above model almost the all assumptions are valid.

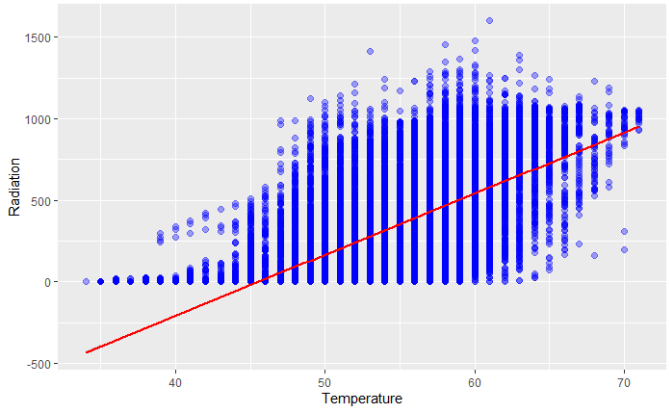


Figure 8: Radiation vs Temperature

```
Call:
lm(formula = Radiation ~ Temperature, data = numData)

Residuals:
    Min      1Q  Median      3Q     Max
-720.40 -130.80  -14.85  104.55 1132.38

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1706.2876     9.8369  -173.5   <2e-16 ***
Temperature    37.4421     0.1911   195.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 214.2 on 32684 degrees of freedom
Multiple R-squared:  0.5402,    Adjusted R-squared:  0.5401
F-statistic: 3.839e+04 on 1 and 32684 DF,  p-value: < 2.2e-16
```

Figure 9: R Summary for the model

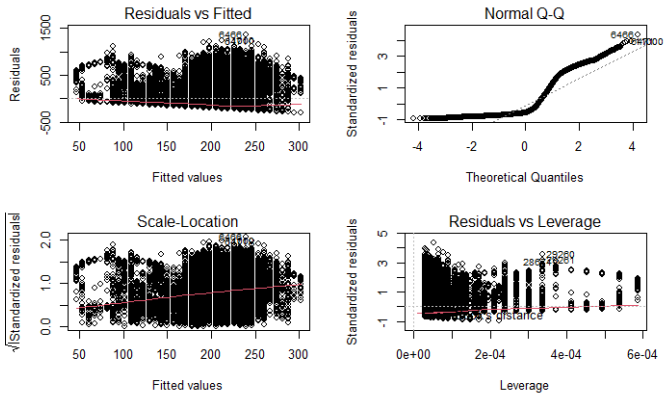


Figure 10: Residual plots

```
Call:
lm(formula = Radiation ~ ., data = numData)

Residuals:
    Min      1Q  Median      3Q     Max
-752.37 -132.65  -20.19  102.98 1121.15

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.098e+04  6.959e+02  30.153  < 2e-16 ***
Temperature    3.837e+01  2.064e-01 185.930  < 2e-16 ***
Pressure      -7.471e+02  2.290e+01 -32.622  < 2e-16 ***
Humidity      -2.691e-01  4.844e-02  -5.555 2.79e-08 ***
WindDirection -2.694e-01  1.462e-02 -18.432  < 2e-16 ***
Speed          7.875e+00  3.416e-01  23.053  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 208.2 on 32680 degrees of freedom
Multiple R-squared:  0.5659,    Adjusted R-squared:  0.5659
F-statistic:  8522 on 5 and 32680 DF,  p-value: < 2.2e-16
```

Figure 11: R Summary for the Full Model

Then we fit a multiple linear model for the data. Similar way we got the full model as the best model from minimum AIC value (PorrasJuly 2018).

In this outcome, it can be seen that the p-value of the F-statistic is < 2.2e-16, which is highly significant. This means that, at least, one of the predictor variables is significantly related to the outcome variable. It can be seen that, changing on Temperature, Pressure, Humidity, WindDirection, and WindSpeed variables are significantly associated to changes in Solar Radiation. For example, we can say depending on the summary detail that, 1 mph additional increment of wind speed leads to an increase of 7.875 kg/s^3 in solar radiation level. 1 $°F$ additional increment of temperature leads to an increase of 38.37 kg/s^3 in solar radiation level. 1 Hg additional increment of barometric pressure leads to a decrease of 747.1 kg/s^3 in solar radiation level. 1% additional increment of humidity leads to a decrease of 0.2691 kg/s^3 in solar radiation level. 1 degree additional increment of wind direction leads to a decrease of 0.2694 kg/s^3 in solar radiation level. So the final full model is,

$$\hat{Radiation} = (20980) + (38.37) * Temperature + (-747.1) * Pressure +$$
$$(-0.2691) * Humidity + (-0.2694) * WindDir + (7.875) * WindSpeed + \epsilon$$

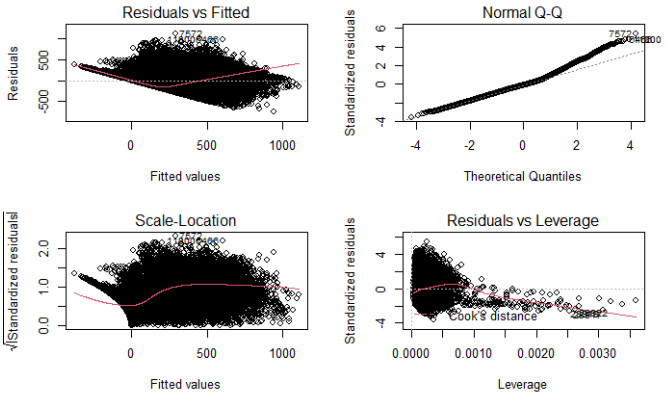


Figure 12: Residual Plots

Above Figure 12 will show us how the full model is validation regression assumptions.

3.3 Bayesian Regression Model

3.3.1 First Approach - stan_glm

We use two approaches to build a Bayesian Regression model. First one is using the function stan_glm from the rstanarm package. In this function, we apply family as default use gaussian distribution, prior as default normal prior. If we need a flat uniform prior we put it to be NULL, and prior_intercept normal, student_t, or cauchy. Here also if we need a flat uniform prior we put it to be NULL.

```
stan_glm
 family:       gaussian [identity]
 formula:      Radiation ~ .
 observations: 32686
 predictors:   6
------
             Median    MAD_SD
(Intercept)  20998.757  704.571
Temperature     38.373    0.211
Pressure      -747.457   23.291
Humidity        -0.269    0.050
WindDirection   -0.270    0.014
WindSpeed        7.863    0.338

Auxiliary parameter(s):
      Median  MAD_SD
sigma 208.152  0.788

------
* For help interpreting the printed output see ?print.stanreg
* For info on the priors used see ?prior_summary.stanreg
```

Figure 13: R Bayesian full model stan_glm output

We start with the full model and the outcome gives us,

From this output, we got median estimator as the median computed from the MCMC simulation. Also MAD_SD is median absolute deviation computed from the same simulation.

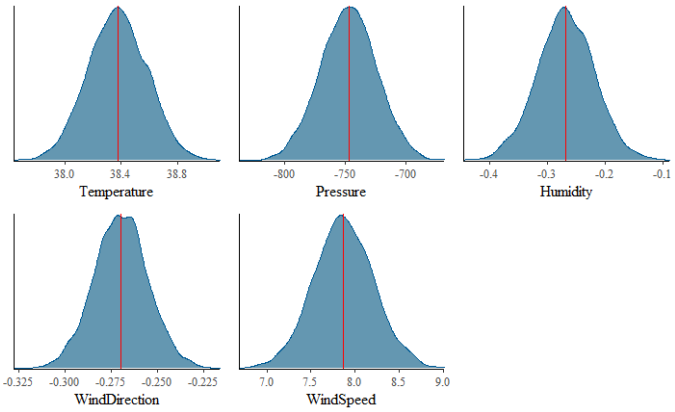


Figure 14: MCMC densities

Point estimates of variables falls on the median of this distribution. Then we need to evaluate the model parameters.

```
# Description of Posterior Distributions

Parameter     |    Median |                89% CI |     pd |          89% ROPE | % in ROPE |  Rhat |      ESS
-------------------------------------------------------------------------------------------------------------
(Intercept)   | 20998.757 | [19939.204, 22192.550] | 100.00% | [-31.592, 31.592] |         0 | 1.000 | 4587.185
Temperature   |    38.373 | [   38.040,    38.703] | 100.00% | [-31.592, 31.592] |         0 | 0.999 | 4383.438
Pressure      |  -747.457 | [ -786.907,  -712.738] | 100.00% | [-31.592, 31.592] |         0 | 1.000 | 4597.143
Humidity      |    -0.269 | [   -0.342,    -0.184] | 100.00% | [-31.592, 31.592] |       100 | 0.999 | 4391.914
WindDirection |    -0.270 | [   -0.292,    -0.246] | 100.00% | [-31.592, 31.592] |       100 | 1.000 | 4880.826
WindSpeed     |     7.863 | [    7.327,     8.405] | 100.00% | [-31.592, 31.592] |       100 | 1.000 | 5318.336
```

Figure 15: R Description of Posterior Distributions

In above description, Median is the Median estimator is the median computed from the MCMC simulation. 89% CI is a Credible Interval, used to quantify the uncertainty about the regression coefficients. With 89% probability (given the data) that a coefficient lies above the CI_low value and under CI_high value. pd is Probability of Direction, which is the probability that the effect goes to the positive or to the negative direction, and it is considered as the best equivalent for the p-value. 89% ROPE gives Region of Practice Equivalence. Rhat is the scale reduction factor $\hat{R}$. ESS gives effective sample size (Press, n.d.).

As we have seen from the above figure the values are closer to each other due to the like normality of the distribution of the posteriors where all the central statistics (mean, median, mode) are closer to each other. As expected they are approximately on top of each other. From these curves we can see the bell shape curves. That give us a hint that we used the prior's as Normal distribution.

Finally We need to check the significance of Bayesian Regression coefficients. That is done by checking whether

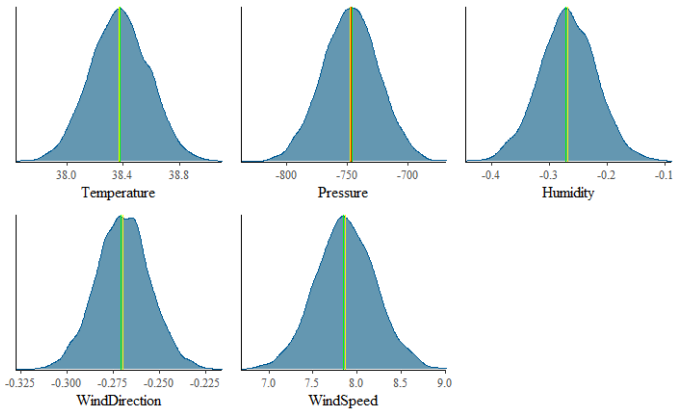


Figure 16: MCMC density plots

the corresponding credible interval contains zero or not, if no then this coefficient is significant. Now let's see significance of out model coefficients.

```
# Highest Density Interval

Parameter     |                89% HDI
-----------------------------------------
(Intercept)   | [19939.20, 22192.55]
Temperature   | [   38.04,    38.70]
Pressure      | [ -786.91,  -712.74]
Humidity      | [   -0.34,    -0.18]
WindDirection | [   -0.29,    -0.25]
WindSpeed     | [    7.33,     8.41]

# Equal-Tailed Interval

Parameter     |                89% ETI
-----------------------------------------
(Intercept)   | [19863.97, 22132.61]
Temperature   | [   38.04,    38.70]
Pressure      | [ -784.64,  -710.21]
Humidity      | [   -0.35,    -0.19]
WindDirection | [   -0.29,    -0.25]
WindSpeed     | [    7.33,     8.41]
```

Figure 17: R HD intervals & ET intervals

From the both result we can see all the coefficients are significant.Note that we got the satisfied results duo to the normal prior assumption. But in real world it is less often to be sure about the normality assumption.

3.3.2 Second Approach - BAS model

We fit a model taking model prior as bernoulli and prior as BIC (Bayesian Information Criterion) in bas.lm function in BAS package. We got the results as follows.

Looking at the summary table we have generated, we believe there is a 95% chance that the solar radiation increases by 37.97 to 38.78 kg/s^3 with one additional increase of the temperature. Similarly, there is a 95% chance that the solar radiation decreases by 702.18 to 791.96 with one additional increase of the Pressure. Again, there is a 95% chance that the solar radiation decreases by 0.17 to 0.36 with one additional increase of the wind direction.

We start with the full model, with all possible predictors. i.e Temperature, Pressure, Humidity, WindDirection, Speed. We drop one variable at a time and record BIC value. Then finally choose the best model having minimum

```
Marginal Posterior Summaries of Coefficients:

Using  BMA

Based on the top  1 models
               post mean   post SD     post p(B != 0)
Intercept       207.12470    1.15133     1.00000
Temperature      38.37214    0.20638     1.00000
Pressure       -747.06869   22.90106     1.00000
Humidity         -0.26910    0.04844     1.00000
WindDirection    -0.26944    0.01462     1.00000
Speed             7.87488    0.34160     1.00000
```

Figure 18: R Marginal Posterior Summaries of Coefficients

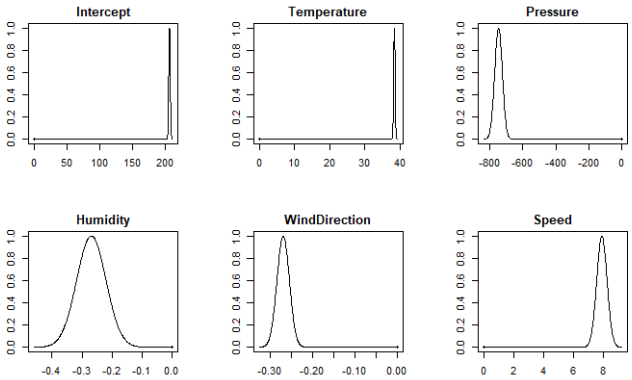


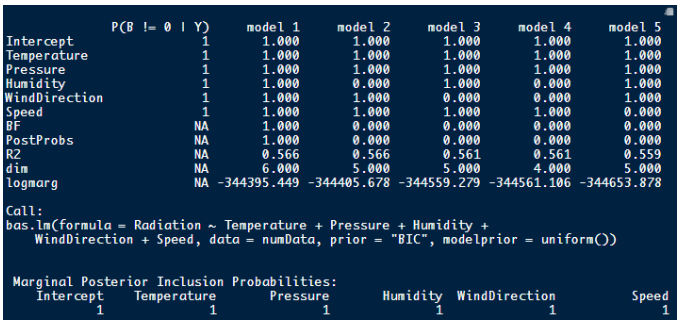
Figure 19: BAS Coefficients plot

```
Marginal Posterior Summaries of Coefficients:

Using  BMA

Based on the top  32 models
               post mean   post SD     post p(B != 0)
Intercept       207.12470    1.15133     1.00000
Temperature      38.37215    0.20639     1.00000
Pressure       -747.06784   22.90148     1.00000
Humidity         -0.26909    0.04847     0.99996
WindDirection    -0.26944    0.01462     1.00000
Speed             7.87490    0.34161     1.00000
              Posterior Mean Posterior SD        2.5%        97.5%
Intercept      207.1246974   1.15132973  204.8680490  209.3813458
Temperature     38.3721364   0.20637957   37.9676249   38.7766479
Pressure      -747.0686858  22.90105705 -791.9555954 -702.1817763
Humidity        -0.2691005   0.04844022   -0.3640451   -0.1741559
WindDirection   -0.2694445   0.01461813   -0.2980965   -0.2407924
Speed            7.8748844   0.34160218    7.2053317    8.5444372
```

Figure 20: R Posterior Infromation about model

BIC value.

Model	BIC value	BIC Difference from FM
FM	349029.7	0
FM-Temperature	372607.2	-23577.5
FM-Pressure	350066.7	-1037
FM-Humidity	349050.1	-20.4
FM-WindDirection	349357.3	-327.6
FM-Speed	349546.5	-516.8



```
                P(B != 0 | Y)     model 1     model 2     model 3     model 4     model 5
Intercept                   1       1.000       1.000       1.000       1.000       1.000
Temperature                 1       1.000       1.000       1.000       1.000       1.000
Pressure                    1       1.000       1.000       1.000       1.000       1.000
Humidity                    1       1.000       0.000       1.000       0.000       1.000
WindDirection               1       1.000       1.000       0.000       0.000       1.000
Speed                       1       1.000       1.000       1.000       1.000       0.000
BF                         NA       1.000       0.000       0.000       0.000       0.000
PostProbs                  NA       1.000       0.000       0.000       0.000       0.000
R2                         NA       0.566       0.566       0.561       0.561       0.559
dim                        NA       6.000       5.000       5.000       4.000       5.000
logmarg                    NA -344395.449 -344405.678 -344559.279 -344561.106 -344653.878

Call:
bas.lm(formula = Radiation ~ Temperature + Pressure + Humidity +
    WindDirection + Speed, data = numData, prior = "BIC", modelprior = uniform())


 Marginal Posterior Inclusion Probabilities:
     Intercept    Temperature       Pressure       Humidity  WindDirection          Speed
             1              1              1              1              1              1
```

Figure 21: R Marginal posterior inclusion probability

We got the almost similar models and results from both approaches. Which is the full model.

4 Conclusion and recommendation

The goal is to build a model that assists statisticians in describing, controlling, and predicting the dependent variable based on the independent variable(s). Here we have explored simple and multiple linear regression and Bayesian linear regression (T. O'Hagan K. Cowles, n.d.). Ultimately, it is up to the statistician to choose which method he or she prefers to use based on any prior knowledge of the data. As we describe earlier, we assume that our bayesian family as gaussian distribution and the prior as default normal prior. Neither method is "better" than the other, it all depends on the prior knowledge of the data and the decision of the statistician as to which method he or she uses. Somehow we have talk about the Bayesian vs Frequentist debate as well.

The aim of Bayesian Linear Regression is not to find the single "best" value of the model parameters, but rather to determine the posterior distribution for the model parameters. The result of performing Bayesian Linear Regression is a distribution of possible model parameters based on the data and the prior. This allows us to quantify our uncertainty about the model: if we have fewer data points, the posterior distribution will be more spread out.

The formulation of model parameters as distributions encapsulates the Bayesian worldview: we start out with an initial estimate, our prior, and as we gather more evidence, our model becomes less wrong. Bayesian reasoning is a natural extension of our intuition. Often, we have an initial hypothesis, and as we collect data that either supports or disproves our ideas, we change our model of the world (ideally this is how we would reason) (Will 2018). For more further analysis we can use advanced tools like WinBugs.

Finally we can represent few regression model based on the data set. Which are,

- Simple Linear Regression
- Multiple Linear Regression
- Bayesian Linear Regression

5 Appendices

This article is done using Rmarkdown bookdown. You can find the all R codes for the analysis using these links,

- Data describe and Exploratory Analysis
- Regression Model fitting 1
- Regression Model fitting 2

Also information about data and details can get from github repository.

- GitHub repository

References

Bowles, Melanie. 2020. “Bayesian Vs. Frequentist Methodologies Explained in Five Minutes,” July. https://infotrust.com/articles/bayesian-vs-frequentist-methodologies-explained-in-five-minutes/.

Brandon Feenstra, Vasileios Papapostolou, Ashley Collier-Oxandale. 2020. “The Airsensor Open-Source R-Package and Dataviewer Web Application for Interpreting Community Data Collected by Low-Cost Sensor Networks.” *Environmental Modelling & Software* 134: 104832. https://doi.org/https://doi.org/10.1016/j.envsoft.2020.104832.

Douglas Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. n.d. *Introduction to Linear Regression Analysis*.

From Wikipedia, the free encyclopedia. 2020. “Exploratory Data Analysis.” https://en.wikipedia.org/wiki/Exploratory_data_analysis.

Hawai’i, University of. n.d. “Hawai’i Space Exploration Analog and Simulation.” https://hi-seas.org/.

Kruschke, J. K. 2010. *What to Believe: Bayesian Methods for Data Analysis. Trends in Cognitive Sciences*.

NSS. 2016. “Bayesian Statistics Explained to Beginners in Simple English.” *Analytic Vidhya*, June. https://www.analyticsvidhya.com/blog/2016/06/bayesian-statistics-beginners-simple-english/#:~:text=\T1\textquotedblleftBayesian%20statistics%20is%20a%20mathematical,the%20evidence%20of%20new%20data.\T1\textquotedblright.

P.murphy, Kevin. 2012. “Machine Learning: A Probabilistic Perspective,” 589. https://www.r-bloggers.com/2020/04/bayesian-linear-regression/amp/.

PorrasJuly, Eladio Montero. 2018. “Linear Regression in R.” *Data Camp Learn R*, July. https://www.datacamp.com/community/tutorials/linear-regression-R.

Prabhakaran, Selva. 2019. “Complete Introduction to Linear Regression in R.” *Machine Learning Data Science*, March. https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/#:~:text=Linear%20regression%20is%20one%20of,only%20the%20X%20is%20known.

Press, S. J. n.d. “Bayesian Statistics: Principles, Models, and Applications,” *JohnWiley and Sons*.

“Simple Linear Regression Numerical Example.” n.d. http://en.wikipedia.org/%0Awiki/Simple_linear_regression#Numerical_example.

Streftaris, G., and G. J. Gibson. 2004. “Bayesian Inference for Stochastic Epidemics in Closed Populations.” In, 63–75.

Team, BioTuring. 2018. “How to Compare Box Plots,” May. https://bioturing.medium.com/how-to-compare-box-plots-3da8e2adfa8f.

T. O’Hagan K. Cowles, R. Kass. n.d. “What ?,8 Bayesian Analysis?” http:%20/%20/%0Abayesian.org/Bayes-Explained.

Will, Koehrsen. 2018. “Introduction to Bayesian Linear Regression,” April. https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7.

ZARAH, LEANN. 2020. “7 Reasons Why Research Is Important.” https://owlcation.com/academia/Why-Research-is-Important-Within-and-Beyond-the-Academe.