Predicting Solar Radiation

MATH 261A

Chris Gerlach, Prathana Phukon
Qui Le, Chun For Tam, Wilson Strasilla

September 4, 2024

Contents

1 Introduction

The NASA-funded project HI-SEAS V, or Hawaii Space Exploration Analog and Simulation, is a habitat that acts as a training ground and testing site to explore Mars. This site is located on the Mauna Loa Big Island of Hawaii in order to simulate the isolated environment that Mars has. As the habitat is isolated, it relies on solar batteries to power the daily activities of 6 crew members, from conducting experiments to cooking and exercising. Therefore, there is a strict energy consumption plan in place to ensure that vital equipment stays online, and key experiments can be run [SG17].

1.1 Background

Our first goal would be to use the meteorological data collected on-site to build a prediction model for solar radiation, which would help the crew to better understand how much energy they will have to work with when planning their activities. However, considering Mars has a very different atmosphere than Earth, the meteorological data gathered on Earth may not be in the same domain as Mars, and precise prediction can be difficult to achieve. Therefore, our secondary goal would be to identify which kind of atmospheric characteristics contribute significantly to solar radiation generation.

1.2 Dataset Description

The original data set collected the following variables:

UNIX Time	Seconds since Jan 1, 1970
Data	UNIX time in yyyy-mm-dd hh:mm:ss format
Time	UNIX time in local time, hh:mm:ss format
Time Sun Rise	Local time when the sun rose in hh:mm:ss format
Time Sun Set	Local time when the sun set in hh:mm:ss format
Radiation	amount of Radiation in watts per square meter
Temperature	Fahrenheit
Pressure	Hg
Humidity	Percentage
Wind Direction	Degrees counterclockwise from panel face
Speed	Wind speed in miles per hour

Table 1: Original Dataset [AND17]

2 Data Preprocessing

2.1 Data Formatting

Before a model can be fitted to the current data, some verifications and modifications need to be done in preparation for model building and model adequacy investigation. With some necessary data preprocessing steps taken in advance, model performance can be enhanced in terms of the accuracy of future predictions about solar radiation given new observations for the current independent variables of the data set. First, a data set may contain invalid or missing data points, and not addressing this issue beforehand is an extremely bad decision. There are a few scenarios to consider. If a lot of observations are invalid or missing, then data augmentation is needed. However, if a small percentage of data is missing, a removal of those data points is acceptable. In this data set, no invalid or missing data are found and thus no action is taken on that front. Second, an assessment is done to discover any highly influential data points, and no influential points are found after removing the night-time data points. Further discussion about data removal will be included later. Then, a few predictor variables need their data to be broken into smaller bits of information or converted into a format that is suitable for model building.

The Data predictor variable has information about the date in which observations were collected. Two new variables are created, Month and Day, from the Data variable with Month becoming a new categorical variable. Since data was collected from September 2016 to December 2016, our new categorical variable has four levels. The Time predictor variable currently has its data in the hours, minutes, and seconds format. Its data is then converted into value in seconds only, and a new variable, Time of Day, is created to capture that change. Similarly, data for variables, Time Sunrise and Time Sunset, are converted and saved into new variables of the same name. With data from Time of Day and Time Sunrise, a new variable is created to capture the time since sunrise, and it is called Time Since Sunrise. Also, a new categorical variable, Day/Night, is created from Time of Day, Time Sunrise, and Time Sunset. After data conversion, we checked for correlation between each possible pair of variables, and found that there are significant correlations between variables such as UNIXTime, Time Since Sunrise, Time Sunrise, and Time Sunset. Consequently, the original time variables are dropped in favor of Time Since Sunrise, and we are left with the following variables:

1. Radiation
2. Temperature
3. Pressure
4. Humidity
5. Wind Direction
6. Speed
7. Time Since Sunrise (TSS)

8. Month

With this new data set of eight predictor variables, we use the hold-out method to split our data using the 80:20 ratio, where 80% is used to train our model and 20% will be used to test the performance of our model.

2.2 Night-Time Data Removal

Given the Day/Night categorical variable, a useful plot of Radiation against Day/Night is produced to have a general view of the full data set in Figure 1. What is observed is that a little over half of the data set is recorded during the night time when there is little to no sunlight, and thus there is no radiation. As there is no radiation during this time and we are only interested in modeling times when it is possible to generate energy, we removed all of the data recorded between sunset and sunrise, while noting that the weather makes little to no difference in solar radiation if the sun is not out. After all night-time observations were removed, the highest Cook's distance value is 0.001168884 which is small enough to confirm that no influential data points exists in our updated data set.

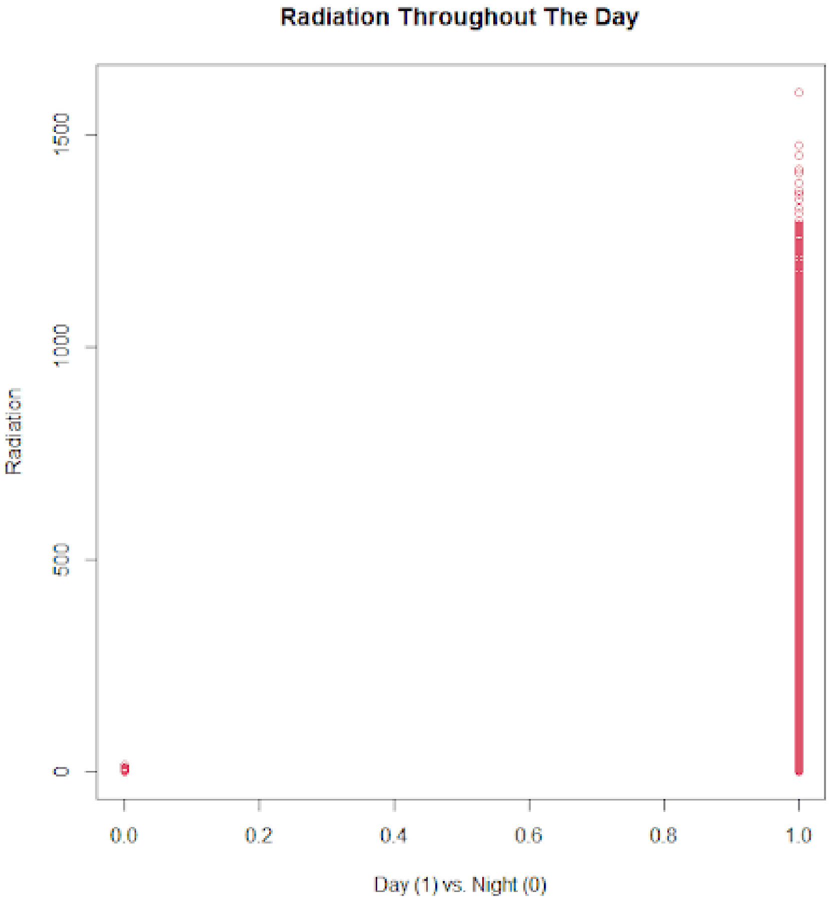


Figure 1: Plot of Day/Night Against Radiation

3 Model Assumptions

While we have eliminated the uninteresting night-time data, and dealt with our highly correlated time variables by aggregating them into one attribute in TIME SINCE SUNRISE, our data is still lacking in a few areas. For one, radiation does not have constant variance with respect to some variables such as time.

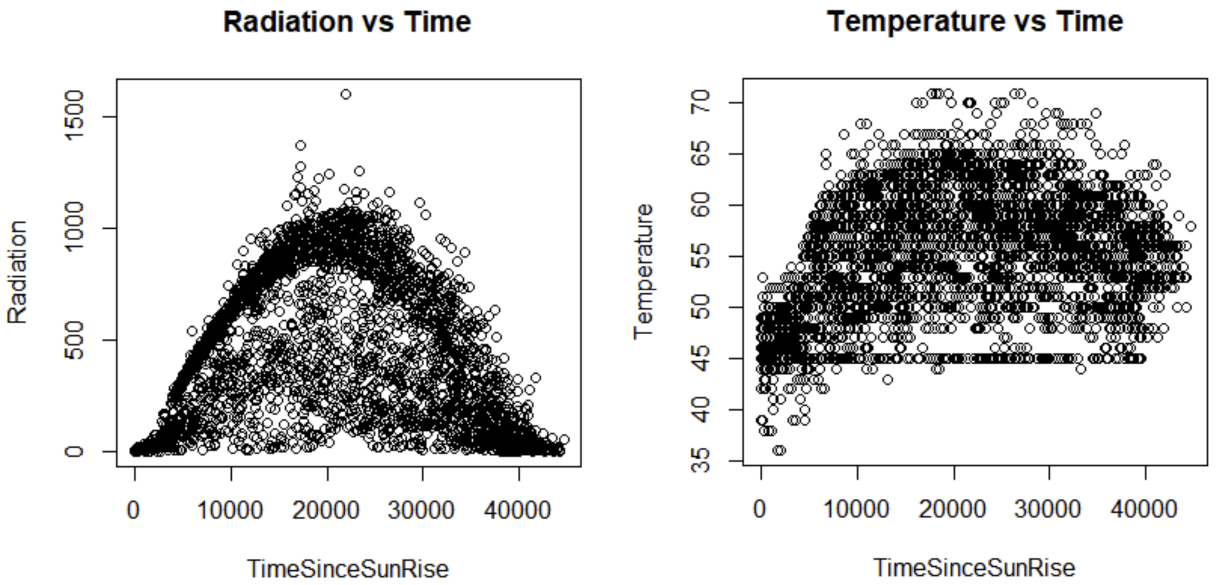


Figure 2: Scatter Plot of Radiation and Temperature against Time

We can see in Figure 2 that around sunrise, and late in the day the radiation tends to be near zero, while varying greatly during other times of the day. Our early models did an extremely poor job of replicating this, sometimes predicting high magnitude negative values which is impossible for solar radiation. This issue is caused by other predictors having high variance regardless of what time it is. We can see in Figure 2 that temperature can vary by nearly twenty degrees in the early morning. As our models are regression models which assign a weight to each variable, this variance in temperature at times around sunrise will lead to very different predictions for radiation at this time, despite our intuitive knowledge that radiation will be negligible.

After fitting the processed data in a model, another issue surfaces as the residual assumptions were not met. Looking at the residual plot (Figure 3), we can see that the residual variance is not constant, and that there is a clear pattern. The red line presents the negative forty-five degree line, and there are no points below this. This is due to the fact that any point below the line would be an observation with negative radiation which is impossible. Any points that are along this line are observations with radiation equal to or very near to zero, showing just how poorly this model performs at predicting observations with zero radiation.

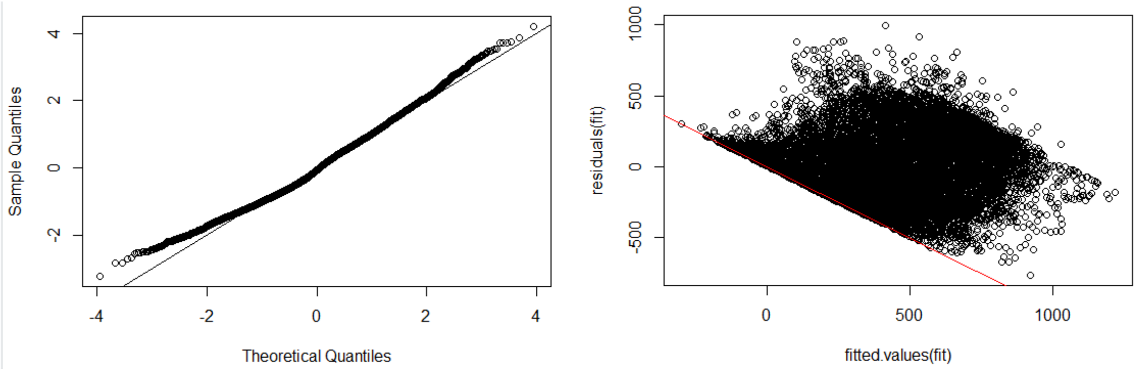


Figure 3: (Left) A Q-Q plot showing how normally distributed our residuals are. Ideally points will lie on the line with no other clear patterns. (Right) A residual plot for our original model. There are numerous issues on display here.

Moreover, the residuals should not depend on each other, have a functional relation, or show non-constant variance relative to what they are plotted against. The residual plot displayed above violates each of these model assumptions by demonstrating strong autocorrelation, a decreasing linear relation, and non-constant variance. Therefore, the data is not suitable for regression in its current state.

The final assumption we make is that the residuals are normally distributed. Fortunately, Figure 3 shows that this is not as much of a concern with our data. In this graph, seeing the majority of the points on the forty five degree line would mean that the residuals have a normal distribution. While we can see a bit of curvature, it was something we were able to address in later models.

4 Model Building

In order to meet the model assumptions and fix the issues highlighted in the previous section, several different transformations were tried on the response and predictor variables.

4.1 Transformation On Response

Using the Box-Cox algorithm to find the best transformation for the response, we found that the best transformation with an intuitive interpretation is a square root transformation:

$$\text{Radiation}' = \sqrt{\text{Radiation}}$$

We then fit a model with this transformed response and found that our residual assumptions were still not met, but normality was somewhat improved as we see the points to not start curving away from the 45-degree line until the tail in Figure 4.

We did, however, see a slight improvement in model accuracy, as R^2 increased. As our residual assumptions are still not met with this model, we attempted several other transformations on the response but were unable to find a better result, so we moved forward with the root transformation and attempted to find improvements through the transformation of our predictors. The improvements of R^2 can be seen in the table. Note that, since we transformed the response, MS_{Res} between the models are no longer comparable.

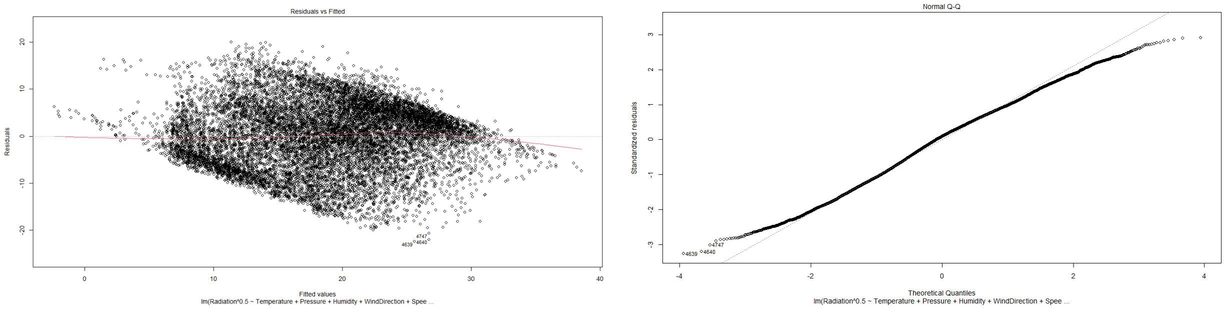


Figure 4: Residual Plot and Q-Q Plot after Transformation on Response

Metric	Original Model	Model with Transformed Response
R^2	0.4377	0.4466
MS_{Res}	63201.96	47.513449
Residual Plot	Almost the same	Almost the same
Q-Q Plot	Better	Worse

Table 2: Response Transformation

4.2 Variable Selection

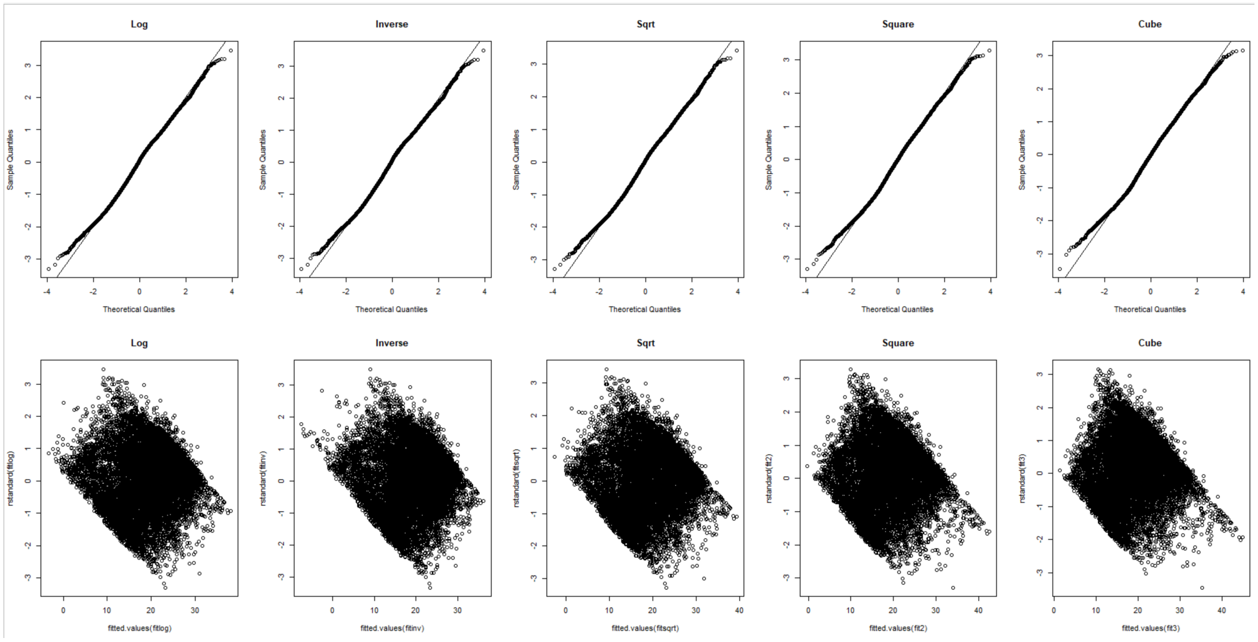
An exhaustive search was employed to find which of the transformed variables were to be included in the model. As seen in Figure 8 (see Appendix A.1), the most viable model contains all predictor variables. This can be seen by the relatively high Mallow's C_P statistic, high MS_{Res}, and low R^2 of the potential reduced models. It should be noted that using the threshold of requiring Mallow's C_P be less than or equal to p yields only one acceptable model: the full model. However, this is already guaranteed by Mallow's C_P test, and hence one can conclude that including all predictors reduces the mean square error more than any simpler model.

4.3 Transformations on Predictor Variables

Our first step in trying to find good transformations for our predictor variables was to look at scatter plots of radiation against each predictor to try to identify non-linear relations. As we fit more and more models, we found that our residual assumptions would not be easy to meet. With the more tactical approach failing, we decided to spread the net and fit models with each of the more common transformations of log, square, cube, root, and inverse to one predictor at a time, leaving the other predictors untransformed. By comparing R^2, MS_{Res}, and the residual plots of these five models for each predictor, we came to a decision on which transformations were best.

As an example, the comparisons between the transformations on Temperature can be seen in Figure 5.

Figure 5: Plots for Individual Transformations on Temperature



After considering the residual plots, Q-Q plots, R^2, and MS_{Res}, the best predictor transformations are as follows:

$$\begin{aligned}
\text{Temperature}' &= \log(\text{Temperature}) \\
\text{Pressure}' &= \text{Pressure} \qquad\qquad [\text{No Change}] \\
\text{Humidity}' &= \frac{1}{\text{Humidity}} \\
\text{Wind Direction}' &= \sqrt{\text{WindDirection}} \\
\text{Speed}' &= \sqrt{\text{Speed}} \\
\text{Time Since Sunrise (TSS)}' &= \text{TSS}^3
\end{aligned}$$

The residual plot is only marginally better than that of the original model. The autocorrelation is still present, contributing to the non-constant variance of the residuals. Hence, the same patterns seen within the very first models are still present.

On the other hand, the Q-Q plot looks close to perfect, suggesting that the residuals are normally distributed. There are also huge improvements in the prediction power of the model with R^2 drastically improving, and MS_{Res} decreasing.

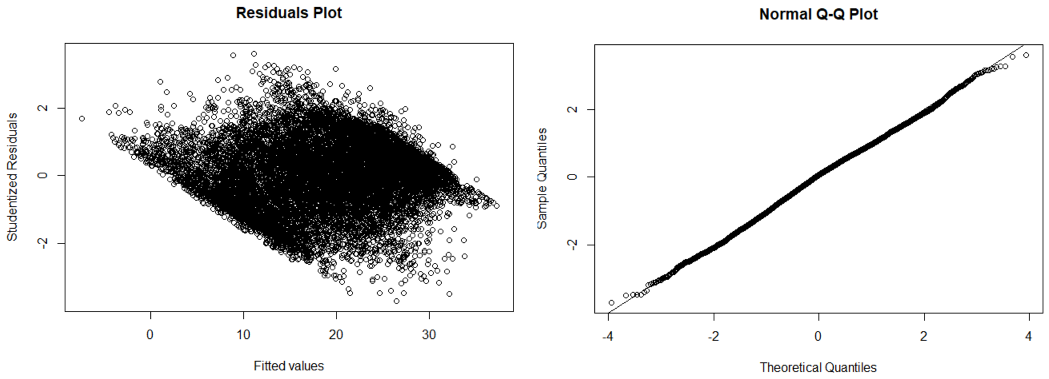


Figure 6: Residual Plot and Q-Q Plot after Transformation on Response

Metric	Transformed Response	Transformed Response and Predictors
R^2	0.4466	0.6176
MS_{Res}	47.513449	32.8329
Residual Plot	Better Than Original	Best
Q-Q Plot	Worse Than Original	Best

Table 3: Transformed Predictors v. Predictors

5 Results

The final model can be written as:

$$\sqrt{\text{Radiation}} = \begin{cases} 422.7 - 22.5\,(\text{Pressure}) + 70.5\log(\text{Temp}) - \frac{144.9}{\text{Humidity}} - .2\sqrt{\text{Wind Direct}} + .9\sqrt{\text{Speed}} - 2.04\text{e-}13\,(\text{TSSR})^3 & \text{Sept} \\ 425.2 - 22.5\,(\text{Pressure}) + 70.5\log(\text{Temp}) - \frac{144.9}{\text{Humidity}} - .2\sqrt{\text{Wind Direct}} + .9\sqrt{\text{Speed}} - 2.04\text{e-}13\,(\text{TSSR})^3 & \text{Oct} \\ 427.1 - 22.5\,(\text{Pressure}) + 70.5\log(\text{Temp}) - \frac{144.9}{\text{Humidity}} - .2\sqrt{\text{Wind Direct}} + .9\sqrt{\text{Speed}} - 2.04\text{e-}13\,(\text{TSSR})^3 & \text{Nov} \\ 426.0 - 22.5\,(\text{Pressure}) + 70.5\log(\text{Temp}) - \frac{144.9}{\text{Humidity}} - .2\sqrt{\text{Wind Direct}} + .9\sqrt{\text{Speed}} - 2.04\text{e-}13\,(\text{TSSR})^3 & \text{Dec} \end{cases}$$

Metric	Value
R^2	0.655
R^2_{adj}	0.644
MS_{Res}	30.5

Table 4: Final Model Summary Statistics

The summary statistics of this model can be seen in Table 4 above. It can be seen that every predictor variable is very significant with p-value less than 2.2e-16. These values, along with the respective residual and Q-Q plots, yield the best overall model.

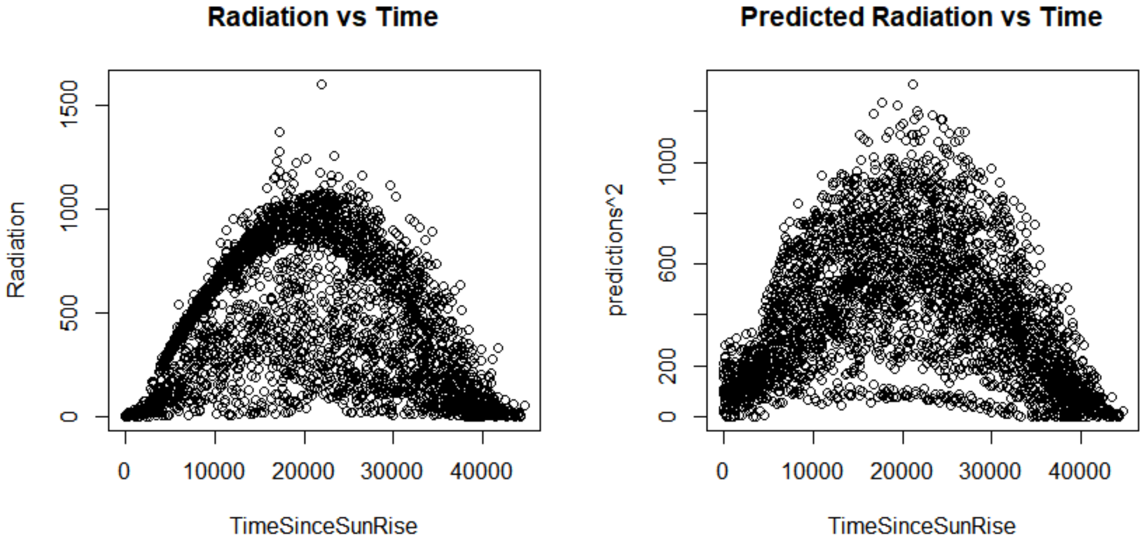


Figure 7: True Radiation v. Predicted Radiation

Now, the model can be validated by testing it on the test set which contains 20% of the data that was not used to fit the model. The model has a $R^2_{prediction}$ of 0.653, meaning that 65% of the variation in the test set is accounted for by the model. This is a relatively mediocre prediction. However, it is important to note that the prediction rate is higher for values away from sunrise or sunset and poorer for those close to sunrise or sunset. Given the similarity between R^2 and $R^2_{prediction}$, we note that there is negligible overfitting between the final model and the training set.

6 Conclusions and Future Scope

Recalling our project objectives, we first wanted to build a predictive linear regression model for solar radiation. There are a few key issues in our data set that make it difficult in developing such a model. First, the overall variation, explained through the model along with the mean square residual error, improved negligibly with the inclusion of additional predictor variables despite them being statistically significant, regardless of which transformation was used. This indicates that the system is much more complex than what the model can account for, so the inclusion of additional atmospheric variables such as cloud levels, gas composition, or ozone levels would potentially improve the model markedly.

Moreover, even after removing night-time data, there is still a large number of zero or near-zero response values immediately after sunrise or before sunset. Along with the variation amongst the other predictor variables, these time periods exhibit different behavior than the rest of the day time. Thus, a single model cannot accurately predict the behavior during every period. Lastly, solar radiation is time dependent which reduces the regression model adequacy. Since time is ordered, we can see the errors are dependent on the fitted values of the model and strong auto-correlation within the original model.

While we are able to achieve a $R^2_{prediction}$ of 0.653, which is greater than that of the model from the original author on Kaggle (0.55), we have large variances between the predictors and response, and our model does not uphold the constant variance assumption [AND17]. With that assessment, we conclude that this model will not be the best at predicting solar radiation.

However, we achieved our secondary goal which is that we have identified that Temperature, Pressure, Humidity, Wind Direction, Speed, Time Since Sunrise, and Month all contribute significantly to the prediction of solar radiation according to our summary statistic mentioned above. Therefore, for building future predictive models, these atmospheric characteristics should be taken into consideration.

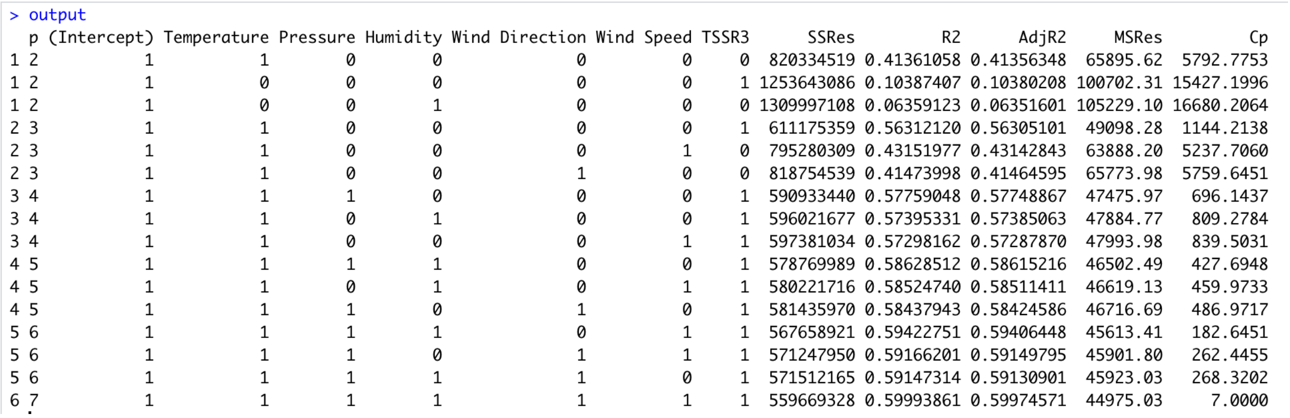
References

[AND17] ANDREY. Solar radiation prediction task from nasa hackathon. 2017.

[SG17] S. Hemmings D. Ardalan L. Cater M. Scott N. K S. Gupta, B. Garcia. You are my sunshine. 2017.

A Appendix

A.1 Variable Selection: Exhaustive Search



```
> output
  p (Intercept) Temperature Pressure Humidity Wind Direction Wind Speed TSSR3      SSRes         R2      AdjR2     MSRes         Cp
1 2           1           1        0        0              0          0     0  820334519 0.41361058 0.41356348  65895.62  5792.7753
1 2           1           0        0        0              0          0     1 1253643086 0.10387407 0.10380208 100702.31 15427.1996
1 2           1           0        0        1              0          0     0 1309997108 0.06359123 0.06351601 105229.10 16680.2064
2 3           1           1        0        0              0          0     1  611175359 0.56312120 0.56305101  49098.28  1144.2138
2 3           1           1        0        0              0          1     0  795280309 0.43151977 0.43142843  63888.20  5237.7060
2 3           1           1        0        0              1          0     0  818754539 0.41473998 0.41464595  65773.98  5759.6451
3 4           1           1        1        0              0          0     1  590933440 0.57759048 0.57748867  47475.97   696.1437
3 4           1           1        0        1              0          0     1  596021677 0.57395331 0.57385063  47884.77   809.2784
3 4           1           1        0        0              0          1     1  597381034 0.57298162 0.57287870  47993.98   839.5031
4 5           1           1        1        1              0          0     1  578769989 0.58628512 0.58615216  46502.49   427.6948
4 5           1           1        0        1              0          1     1  580221716 0.58524740 0.58511411  46619.13   459.9733
4 5           1           1        1        0              1          0     1  581435970 0.58437943 0.58424586  46716.69   486.9717
5 6           1           1        1        1              0          1     1  567658921 0.59422751 0.59406448  45613.41   182.6451
5 6           1           1        1        0              1          1     1  571247950 0.59166201 0.59149795  45901.80   262.4455
5 6           1           1        1        1              1          0     1  571512165 0.59147314 0.59130901  45923.03   268.3202
6 7           1           1        1        1              1          1     1  559669328 0.59993861 0.59974571  44975.03     7.0000
```

Figure 8: Variable Selection: Exhaustive Search

A.2 Transformations On Predictors

The following table shows compares the R^2 and MS_{Res} of transformations on predictors.

Variable	Transformation	R^2	$\sqrt{MS_{Res}}$	MS_{Res}
Temperature	**Log**	**0.5176**	**6.436**	**41.422096**
	Square	0.5061	6.513	42.419169
	Cube	0.4936	6.595	43.494025
	Square Root	0.5166	6.443	41.512249
	Inverse	0.5156	6.449	41.589601
Pressure	Log	0.5143	6.458	41.705764
	Square	0.5143	6.458	41.705764
	Cube	0.5143	6.458	41.705764
	Square Root	0.5143	6.458	41.705764
	Inverse	0.5143	6.458	41.705764
Humidity	Log	0.5201	6.42	41.2164
	Square	0.51	6.487	42.081169
	Cube	0.508	6.5	42.25
	Square Root	0.5172	6.439	41.460721
	Inverse	**0.5231**	**6.4**	**40.96**
Wind Direction	Log	0.5129	6.468	41.835024
	Square	0.5107	6.482	42.016324
	Cube	0.5092	6.492	42.146064
	Square Root	**0.5158**	**6.449**	**41.589601**
	Inverse	0.5086	6.496	42.198016
Speed	Log	0.5086	6.496	42.198016
	Square	0.5083	6.498	42.224004
	Cube	0.505	6.52	42.5104
	Square Root	**0.5166**	**6.443**	**41.512249**
	Inverse	N/A	N/A	N/A
Time Since Sunrise	Log	0.4467	6.893	47.513449
	Square	0.5826	5.987	35.844169
	Cube	**0.6177**	**5.73**	**32.8329**
	Square Root	0.47	6.746	45.508516
	Inverse	0.4471	6.891	47.485881

Table 5: Predictor Transformations