Solar Radiation Prediction

Alireza Mohammadi (250406221)

November 4, 2017

1 Introduction

With the rapid depletion of natural resources and at the same time the air pollution at all-time high, more and more emphasis is shifting towards methods that can efficiently produce energy from environmentally friendly renewable resources such as sun, wind, ocean waves, etc. Aside from that, the main source of energy on the surface of Mars is the solar energy and therefore identifying and understanding effects of key variables on solar radiation are of paramount values as they can help with building a model that can accurately predict the amount of solar radiation as a function of some measurable predictors. Such model can equip the astronauts or colonists with crucial information about when and where to deploy the solar energy harvesting equipment for optimal performance. Building models like this is also very useful on Earth for the same reason.

In this project, we are interested to build a suitable model using multiple linear regression for prediction of the average solar irradiance received in a day as a function of basic meteorological variables of that day. Solar irradiance is power per unit area received from the Sun. The solar power is in the form of the electromagnetic radiation which is emitted from the Sun at various range of wavelengths (Incropera and DeWitt 2002, 700–787). As a result, the amount of solar power measured by a measuring instrument depends on the range of wavelengths that the device can actually detect. For convenience, we shall refer to solar irradiance as solar radiation or just radiation in the rest of this report.

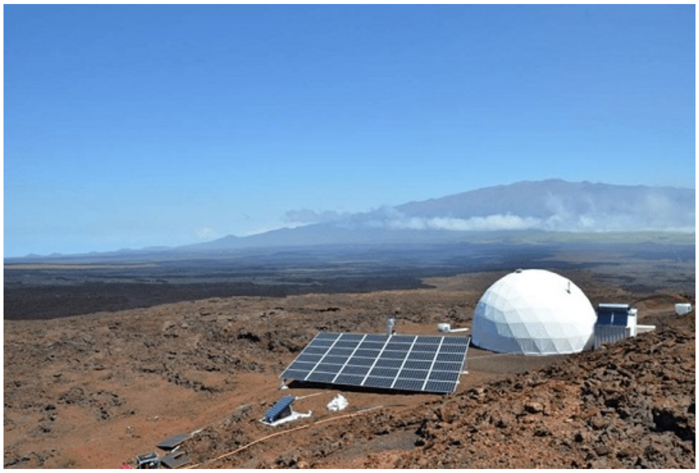

Figure 1: The HI-SEAS Habitat (image from HI-SEAS (2017)).

To build a simple model that can predict the solar radiation, we use the dataset that was collected by NASA HI-SEAS mission (Hawai'i Space Exploration Analog and Simulation, see HI-SEAS (2017)) and was provided on Kaggle. HI-SEAS is a Habitat located on the Big Island of Hawaii at approximately 2500 meters above the sea level (see Figure 1). This isolated and unique location was carefully chosen by NASA research team so that it resembles environment on a Mars site. In particular, this location has little variation in weather (with usually cool and dry climate) which makes it suitable for solar radiation study. This report has been prepared in **R** (R Core Team 2017).

2 Description of the Dataset

The dataset contains the measurements of the solar radiation along with meteorological measurements recorded daily at several time intervals from September 1st, 2016 until December 31st, 2016. In total there are 32,686 observations in the dataset. The raw data consists of UNIX time (indicating computer time in seconds measured from January 1st, 1970), date and time of observation, radiation (W/m^2), temperature (°F), pressure (in-Hg), wind direction (degree), wind speed (mph), time of sunrise and sunset. The UNIX time column is very helpful especially in sorting the data. The rest of columns were produced in MS Excel from these columns. In particular, *daylight* denotes the duration of daylight, that is the time between sunrise and sunset in seconds. The column with heading *dt* is the time interval between two consecutive measurements in seconds and the rest of columns is basically the numerical integration of each variables in each time intervals which will be used later. The goal of this project is to develop a model that can predict the radiation as a function of explanatory variables. A quick check of the data reveals that no observations were taken on these four days: September 30th, November 30th December 6th and 7th. The total number of days that data were collected is $n = 118$ days. Majority of data was recorded at approximately 5 minute time intervals. However, we notice that the time intervals are not the same for all observations. There are time intervals of approximately 10, 15, 20 minutes and even for a few cases of several hours. This is important especially in computation of daily average values. This means that we cannot simply use the sample mean (R function *mean()*) and we need to do a numerical integration in order to compute the correct daily average values. Before computing the integrals, we group the data conveniently by date. The daily average value then is defined by $\bar{f} = (\int_{t_1}^{t_2} f dt)/(\int_{t_1}^{t_2} dt)$ where f denotes any of the input/output variables and t_1 and t_2 denote the time of the first and last observations in each day, respectively. The time integrals are numerically computed using simple trapezoidal rule.

Table 1: Dataframe Summary.

Statistic	N	Mean	St. Dev.	Min	Max
R	118	210.458	76.103	27.112	331.543
T	118	51.153	3.458	43.583	58.158
P	118	30.423	0.050	30.231	30.511
H	118	75.317	20.780	23.440	101.184
WD	118	142.829	39.590	103.318	279.135
WS	118	6.211	1.548	2.461	13.805
DayRatio	118	0.480	0.021	0.456	0.522
TimeObsRatio	118	0.988	0.062	0.535	1.001

In total, we can define up to six linear predictors in our model ($p = 6$). The predictors are daily average temperature (T), pressure (P), humidity (H), wind direction (WD), wind speed (WS) and day ratio (defined as DayRatio $= \frac{\text{daylight}}{24\times3600}$). The response variable is the daily average solar radiation (R). All the input/output variables in this analysis are the daily average values and, for convenience, we may not explicitly mention **daily average** everywhere we want to refer to them throughout the rest of this report. The statistical summary of the data is given in Table 1. The last row (TimeObsRatio) indicates the ratio of each day that data was recorded defined by $(\int_{t_1}^{t_2} dt)/(24 \times 3600)$. This variable is not an explanatory variable but rather is defined to help with interpretation of other variables. The maximum value of this ratio is slightly above 1 and the reason is that for some cases the last point of integration was about a couple of seconds into the next day. Moreover, we notice that the duration of data collection in a day can be as low as 53.47% of a day (look at the minimum value of TimeObsRatio). It is worthwhile to note that the daily average and instantaneous pressure fluctuations are very small and essentially negligible. One may quickly expect that pressure should not play a significant role in the model.

We will perform a systematic analysis in a quest to find the best model for prediction of the solar radiation using available explanatory variables.

3 Preliminary Analysis of Dataset

3.1 Scatterplot Matrix

As the first step of exploration of the dataset, we use the scatterplot matrix shown in Figure 2.

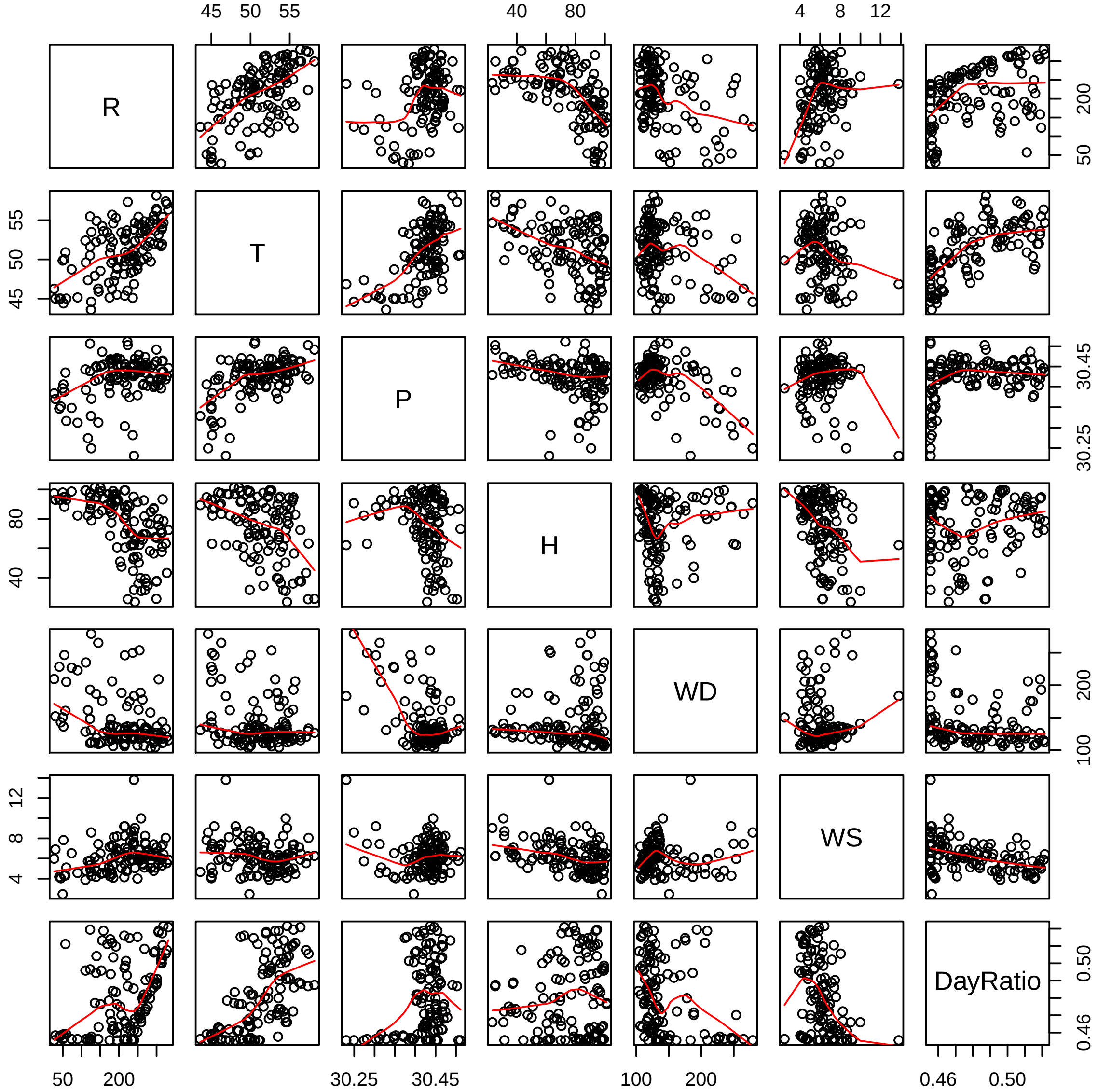

Figure 2: Scatterplot Matrix.

The main conclusions from the above scatterplot matrix can be summarized as:

- Reading across the top row for **radiation**, it seems that **humidity** is the most important predictor as the points are more tightly clustered around the loess. We further notice that **humidity** is negatively correlated with **radiation** and thus the lower the humidity, the higher is the amount of radiation received on the surface of the Earth. This physically makes sense as in this case higher portion of solar radiation is either absorbed by the water molecules in the air or by the clouds in the case of cloudy sky or is just simply reflected back into the sky and away from the Earth. **Pressure** and **wind direction** show the weakest influence on **radiation**. There is a moderate positive association between **wind speed** and **radiation**. **Temperature** and **day ratio** are both positively correlated with **radiation**.

- Reading across the second row from the top for **temperature**, we note that except the wind variables (wind speed and wind direction), **temperature** is associated with all other variables. **Temperature** is positively associated with **radiation**, **pressure** and **day ratio**, and is negatively associated with **humidity**. Positive association of **temperature** with **pressure** may remind us of the ideal gas law $PV = n\bar{R}T$, where P is pressure, V is volume of gas, n is the number of kmol of gas, $\bar{R}$ is the universal gas constant and T is the absolute temperature (Sonntag, Borgnakke, and Van Wylen 2003, 61–66).
- **Pressure** (third row from the top), except for the positive association with **temperature** mentioned above, does not illustrate any noticeable association with any other variables.
- Reading across the fourth row from the top for **humidity**, except for the negative association with **radiation** and **temperature** discussed above, we do not see any other clear association.
- Wind variables (**wind direction** and **wind speed**, rows five and six, respectively) demonstrate the weakest associations with any other variables.
- **Day ratio** is clearly positively associated with **radiation** and **temperature** (see the last row). These associations are not surprising since the length of a day is longer during warmer seasons and thus both average **radiation** and **temperature** are higher during warmer seasons.

In summary, the scatterplot matrix (Fig. 2) suggests that: (i) radiation is most closely predicted by humidity, and (ii) pressure and wind direction are the least important explanatory variables.

3.2 Variance Inflation Factor (VIF)

There is multicollinearity in the explanatory variables when one of them can be linearly predicted with high accuracy using other explanatory variables. The variance inflation factor (VIF) for the design matrix indicates which variables contribute to multicollinearity. The VIF is shown in Figure 3 using bar chart. As can be seen, there is no significant multicollinearity in the dataset as all VIF values are much smaller than the empirical threshold VIF = 10 (marked by red dotted line in Fig.3).

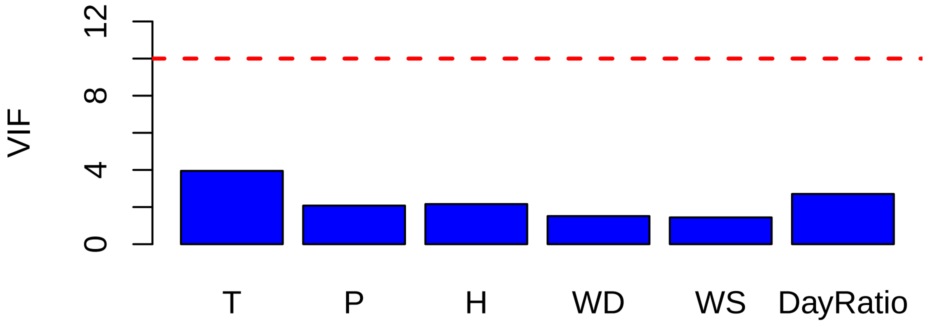

Figure 3: VIF for solar radiation predictors.

Correlation matrix is shown in Table 2 which provides another useful tool for detecting possible multicollinearity. A close inspection of the correlation matrix reveals that **temperature** and **day ratio** have a correlation 0.6318 which is considered a moderate correlation since the coefficient of determination for the regression of **temperature** on **day ratio** is only 39.92%.

Table 2: Correlation Matrix.

	T	P	H	WD	WS	DayRatio
T	1	0.578	-0.440	-0.283	-0.161	0.632
P	0.578	1	-0.268	-0.537	-0.115	0.292
H	-0.440	-0.268	1	0.078	-0.382	0.118
WD	-0.283	-0.537	0.078	1	0.023	-0.283
WS	-0.161	-0.115	-0.382	0.023	1	-0.394
DayRatio	0.632	0.292	0.118	-0.283	-0.394	1

It may be helpful to also visualize the correlation matrix as illustrated in Figure 4. The moderate positive correlations between **temperature** and **pressure**, and also **temperature** and **day ratio** are easily noticeable in this figure.

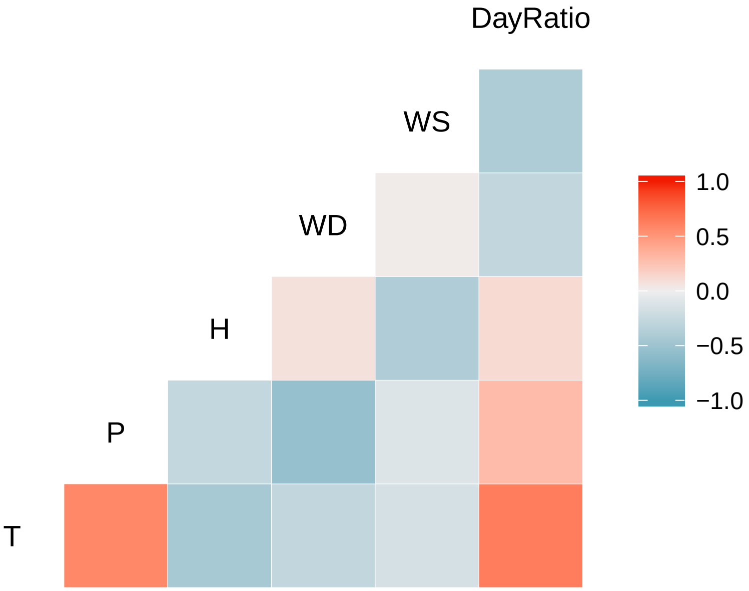

Figure 4: Visualization of the Correlation Matrix.

4 Model Building

4.1 Initial Model

As the first step towards constructing our model, we divide the dataset into two parts: train and test datasets. The training data is used for model building. The test data is used for model validation. We randomly pick $\frac{2}{3}$ of the data as training set, i.e. number of observations in the training set is $n_{tr} = 79$. The remaining observations are stored in the test set, $n_{te} = 39$. It is suggested that when $p > 5$, the best subset method can be used for model selection (see A. I. McLeod 2017, 9). For this purpose, we use function *bestglm::bestglm()* developed by A. McLeod and Xu (2017) with BIC as criterion. The best subset method suggests the following model:

$$R = \beta_0 + \beta_1\text{H} + \beta_2(\text{WD}) + \beta_3(\text{WS}) + \beta_4(\text{DayRatio}) + error. \tag{1}$$

Following this suggestion, we fit a linear model with predictors as described in Eq.(1). Table 3 summarizes the fitted linear regression model R ~ H + WD + WS + DayRatio. It shows that all predictors are significant at less than 1%. However $R^2 = 61.3\%$ is not that impressive and hence the proportion of variability explained by the model is low. We use basic model diagnostic checks to see how we may improve this model.

Table 3: Summary of the fitted linear regression model R $\sim$ H + WD + WS + DayRatio.

	Dependent variable:
	R
H	-1.770^{***}
	(0.280)
WD	-0.515^{***}
	(0.157)
WS	14.895^{***}
	(3.952)
DayRatio	$1{,}803.694^{***}$
	(271.570)
Constant	-542.852^{***}
	(150.261)
Observations	79
R^2	0.613
Adjusted R^2	0.592
Residual Std. Error	46.657 (df = 74)
F Statistic	29.348^{***} (df = 4; 74)
Note:	*p<0.1; **p<0.05; ***p<0.01

4.1.1 Basic Diagnostic Checks

The basic diagnostic checks for the model R ~ H + WD + WS + DayRatio are shown in Figure 5.

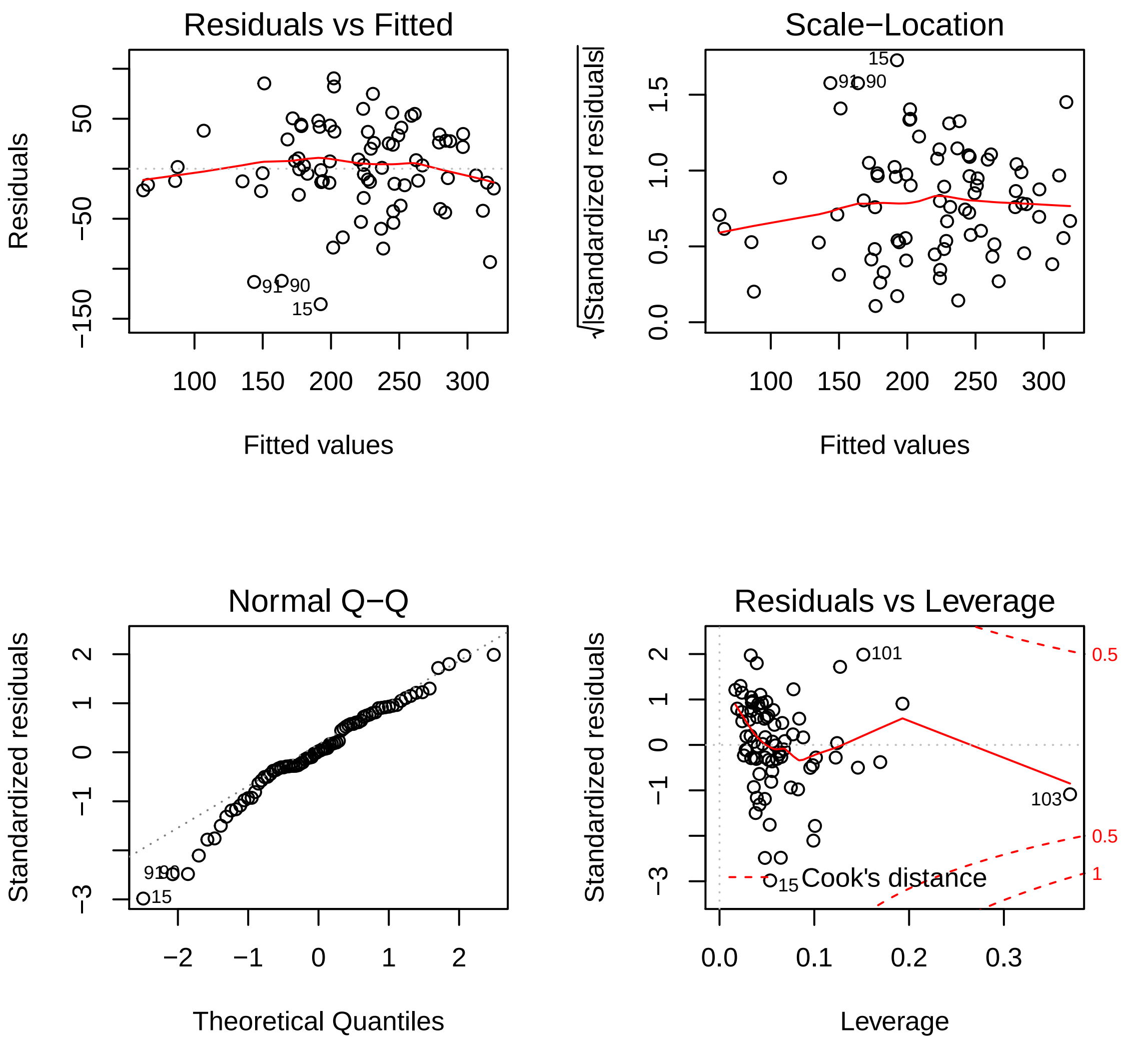

Figure 5: Basic regression diagnostic plots for model R ~ H + WD + WS + DayRatio.

A problem is indicated in the *Residuals vs Fitted* diagnostic plot since the loess trend is not flat. The curve suggests possible nonlinearity due to interaction and/or some nonlinearity present in the inputs. Also, the *Normal Q-Q* plot suggests that the distribution of the residuals does not follow a normal distribution. However, the *Residuals vs leverage* plot shows that none of the observations have a strong influence on the fit.

4.1.2 Residual Dependency Checks

The residual dependency plots shown in Figure 6 indicates that indeed **humidity** as well as **day ratio** and possibly **wind direction** exhibit nonlinear effects.

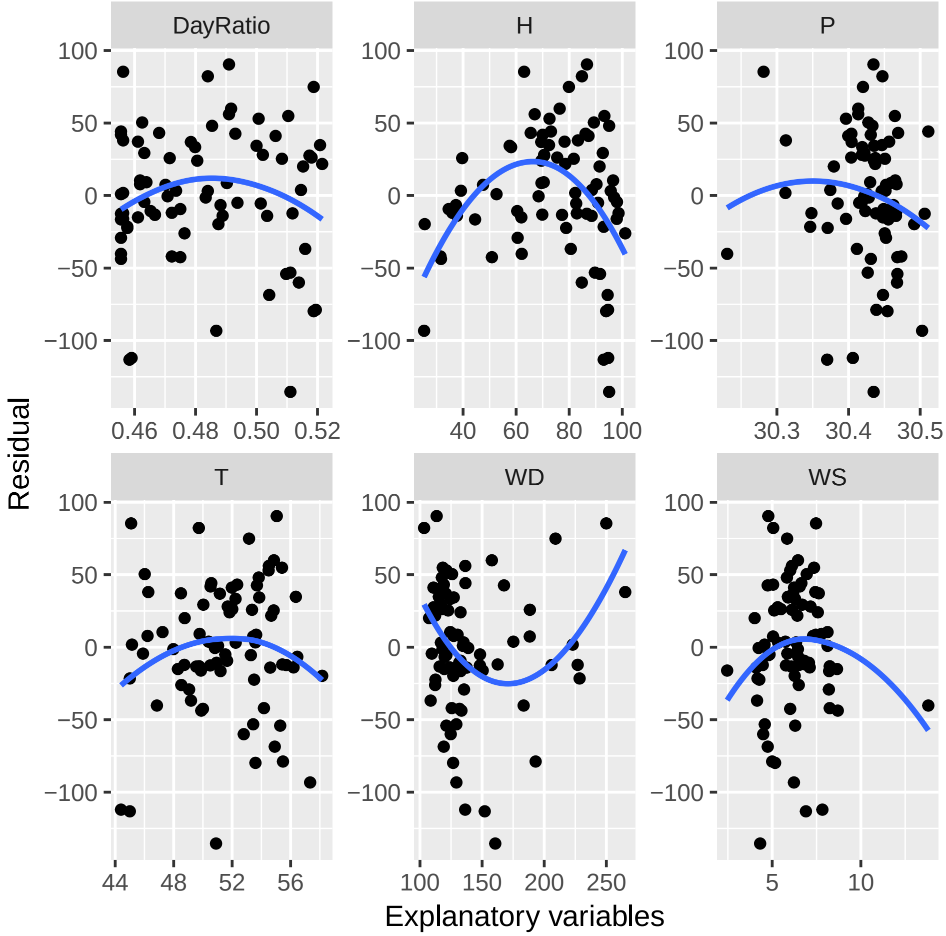

Figure 6: Residual dependency plots for model R ∼ H + WD + WS + DayRatio.

4.2 Revised Model 1

The diagnostic checks suggest that the initial model may be improved by including a quadratic term with humidity so the regression model can be represented as

$$R = \beta_0 + \beta_1 \mathrm{H} + \beta_2 \mathrm{H}^2 + \beta_3(\mathrm{WD}) + \beta_4(\mathrm{WS}) + \beta_5(\mathrm{DayRatio}) + \mathit{error}. \tag{2}$$

Table 4 summarizes the fitted linear regression model 1 (see Eq.(2)). We see that the quadratic term is significant at less than 1%. The R^2 has increased by 8.2% from the previous model to 69.5% and the residual sum squares (RSS) has dropped to RSS = 126943.4 from RSS = 161087.8 for the previous model. The analysis of variance (ANOVA) lack-of-fit test comparing this model with the initial model gives F-statistic $F = 19.64$ on (1, 73) DF and has a two-sided p-value 0.003238% which shows that the null-hypothesis (H_0: the initial model is true) is rejected at level less than 0.1%. Thus far this model has shown some improvement compared with the previous model. Next, we will look at the basic diagnostic plots for this model.

Table 4: Summary of the fitted linear regression for model R ∼ poly(H,2) + WD + WS + DayRatio.

	Dependent variable:
	R
poly(H, 2)1	-323.587^{***} (44.857)
poly(H, 2)2	-186.524^{***} (42.094)
WD	-0.529^{***} (0.141)
WS	12.775^{***} (3.565)
DayRatio	$1{,}722.992^{***}$ (243.405)
Constant	-619.582^{***} (133.955)
Observations	79
R^2	0.695
Adjusted R^2	0.674
Residual Std. Error	41.701 (df = 73)
F Statistic	33.317^{***} (df = 5; 73)
Note:	*p<0.1; **p<0.05; ***p<0.01

4.2.1 Basic Diagnostic Checks for Revised Model 1

The basic regression plots for model R ~ poly(H, 2) + WD + WS + DayRatio clearly show patters in *Residuals vs Fitted* as well as *Scale-Location* diagnostic plots (see Figure 7). It seems that there is still some nonlinearity that has not been adequately accounted for by the model. The *Normal Q-Q* plot has improved and suggests that the distribution of the residuals for this model is closer to a normal distribution compared with the initial model. The *Residuals vs leverage* plot shows that none of the observations have a strong influence on the fit.

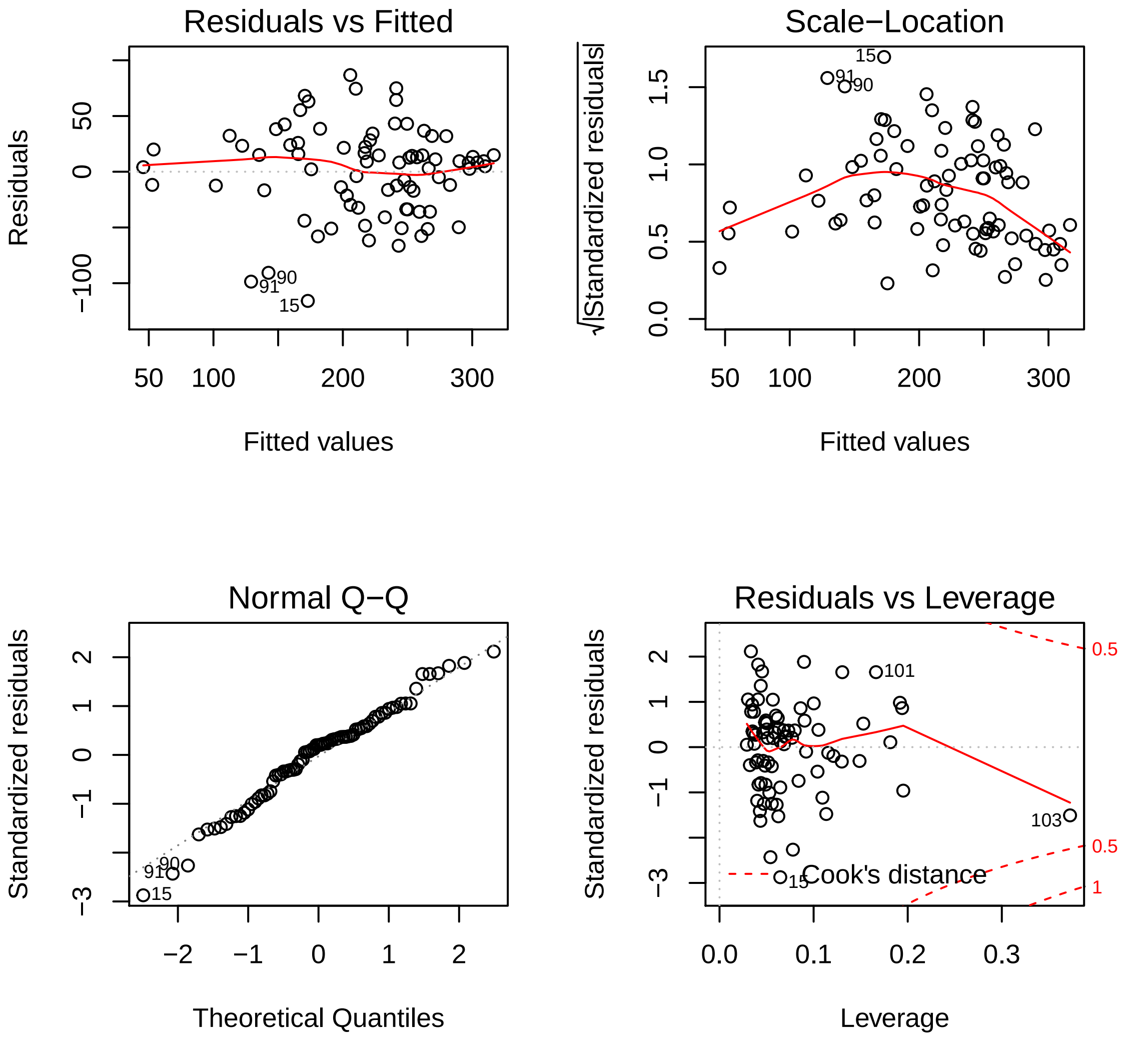

Figure 7: Basic regression diagnostic plots for model R ~ poly(H,2) + WD + WS + DayRatio.

4.2.2 Residual Dependency Checks for Revised Model 1

To further analyze which variables may still have nonlinear behaviour, we employ the residual dependency plots shown in Figure 8. We can see from the residual dependency plots for **day ratio** and **wind direction** that they both display possible nonlinear behaviour. We will add a quadratic term for day ratio as the next step which is discussed in the next section.

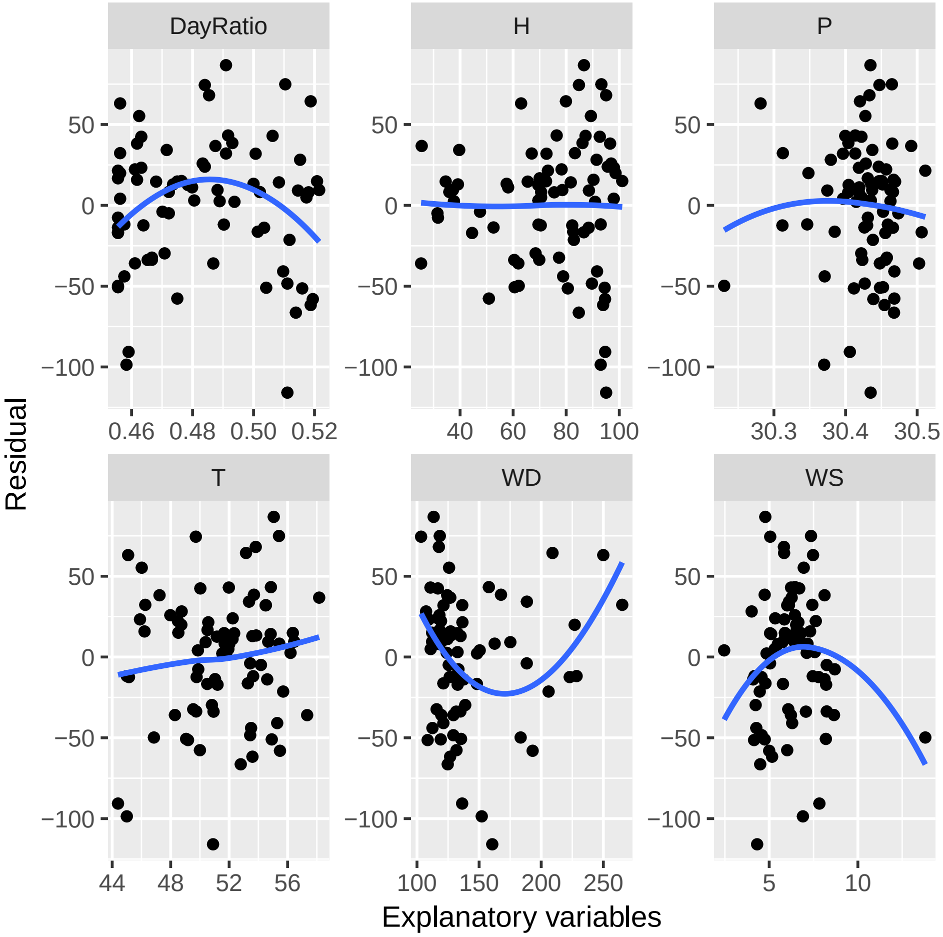

Figure 8: Residual dependency plots for model R ∼ poly(H,2) + WD + WS + DayRatio

4.3 Revised Model 2

As the next step in our model building procedure, we add a quadratic term for day ratio. Thus, the revised model takes the form

$$R = \beta_0 + \beta_1 \mathrm{H} + \beta_2 \mathrm{H}^2 + \beta_3(\mathrm{WD}) + \beta_4(\mathrm{WS}) + \beta_5(\mathrm{DayRatio}) + \beta_6(\mathrm{DayRatio})^2 + error. \tag{3}$$

The summary of the fitted model based on Eq.(3) is provided in Table 5. Note that all regression coefficients are significant at less than 1%. The R^2 has increased by 2.7% from the previous model to 72.2% and the residual sum squares has dropped to RSS = 115666.7 from RSS = 126943.4 for the previous model. The ANOVA lack-of-fit test comparing this model with the revised model 1 gives F-statistic $F = 7.02$ on (1, 72) DF and has a two-sided p-value 0.9904% which shows that the null-hypothesis (H_0: the revised model 1 is true) is rejected at level less than 1%. Thus, this model is preferred compared with the previous model. The next logical step is to analyze the basic diagnostic plots for this model.

Table 5: Summary of the fitted linear regression for model R ∼ poly(H,2) + WD + WS + poly(DayRatio,2).

	Dependent variable:
	R
poly(H, 2)1	-287.222^{***}
	(45.246)
poly(H, 2)2	-203.976^{***}
	(40.991)
WD	-0.418^{***}
	(0.141)
WS	14.287^{***}
	(3.474)
poly(DayRatio, 2)1	347.178^{***}
	(46.096)
poly(DayRatio, 2)2	-118.512^{***}
	(44.731)
Constant	187.260^{***}
	(31.175)
Observations	79
R^2	0.722
Adjusted R^2	0.699
Residual Std. Error	40.081 (df = 72)
F Statistic	31.224^{***} (df = 6; 72)
Note:	*p<0.1; **p<0.05; ***p<0.01

4.3.1 Basic Diagnostic Checks for Revised Model 2

The basic diagnostic plots for model R ~ poly(H, 2) + WD + WS + poly(DayRatio, 2) are shown in Figure 9. Patterns are still easily detectable especially in *Scale-Location* plot (see Figure 7 top right plot). This signals the possibility of presence of some other form of nonlinearity and/or interaction effects that have not been handled well by the model. The *Normal Q-Q* plot has improved from the previous model 1 and suggests that the distribution of the residuals of this model is close enough to a normal distribution. The *Residuals vs leverage* plot shows that none of the observations have a strong influence on the fit.

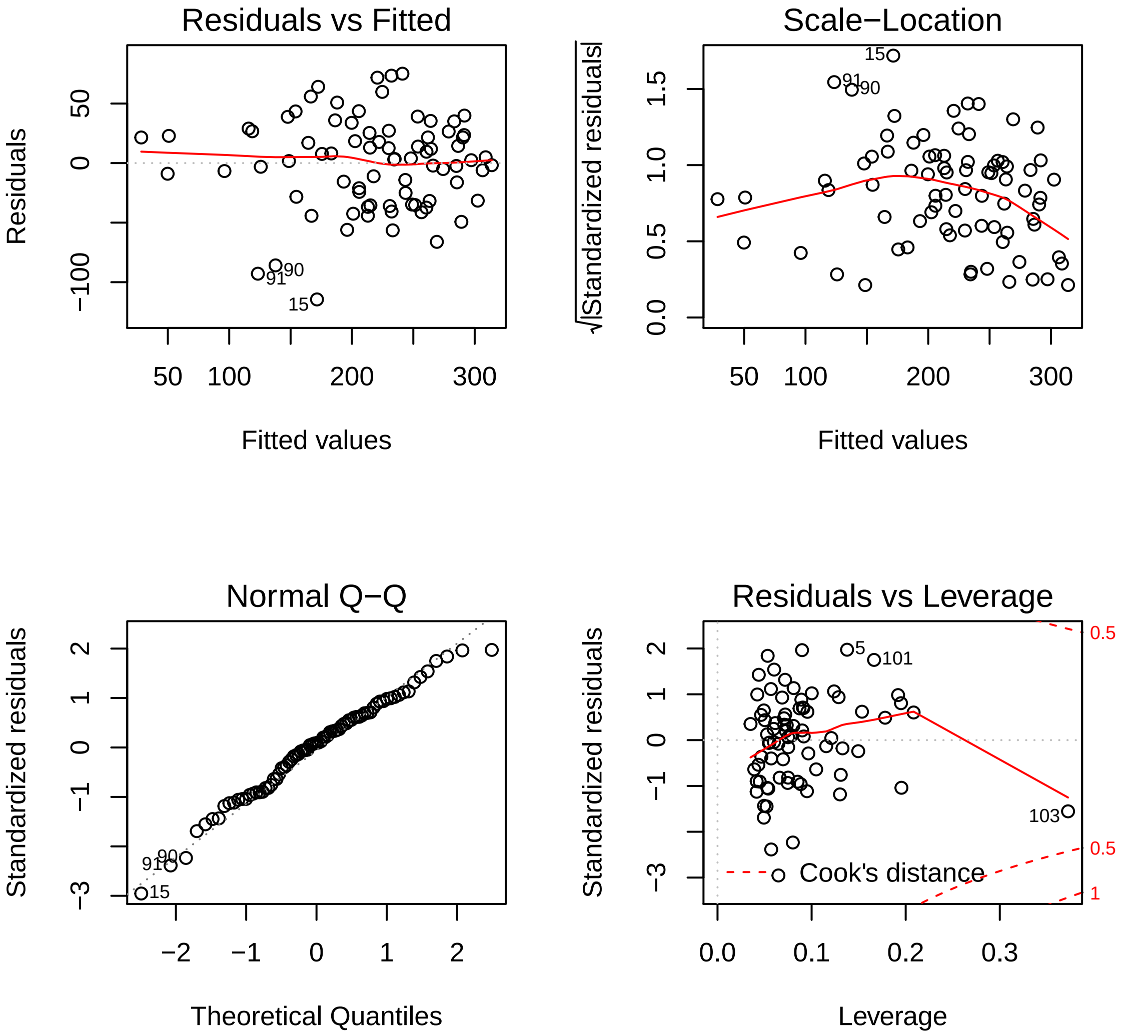

Figure 9: Basic regression diagnostic plots for model R ∼ poly(H,2) + WD + WS + poly(DayRatio,2).

4.3.2 Residual Dependency Checks for Revised Model 2

Residual dependence plots are shown for revised model 2 in Figure 10. The only predictor that still demonstrates some nonlinear behaviour is wind direction. In the next section we will add a quadratic term for wind direction.

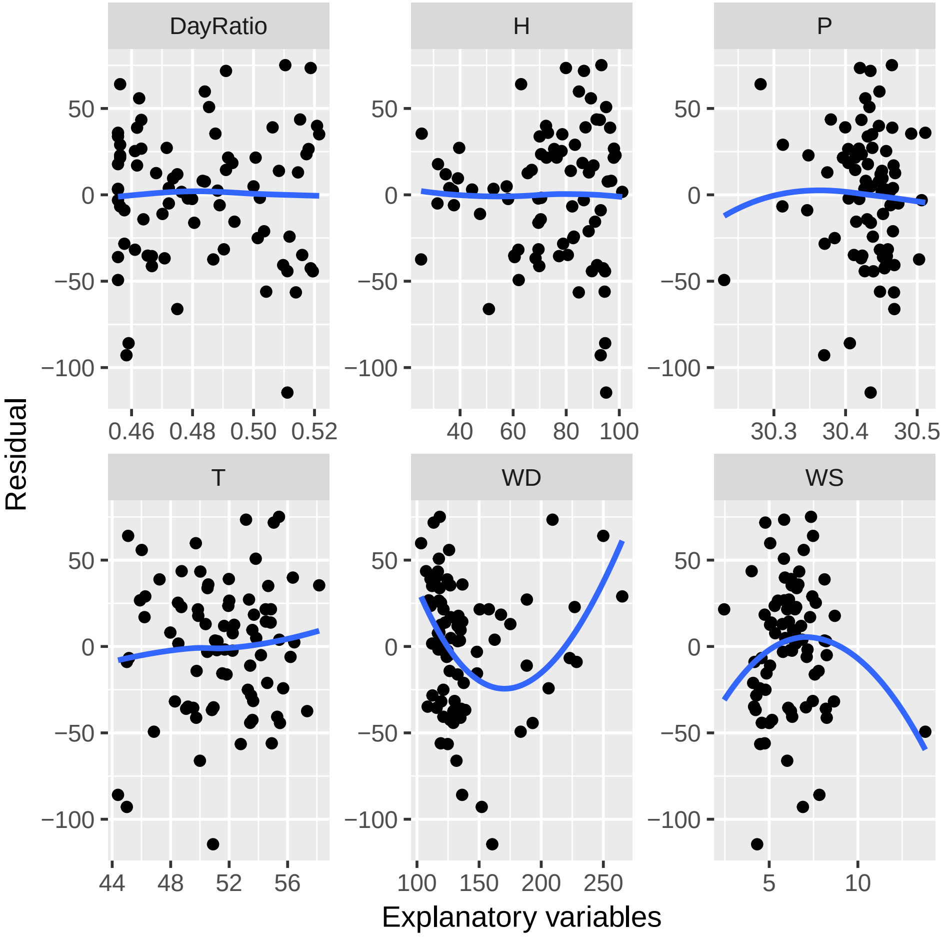

Figure 10: Residual dependency plots for model R ∼ poly(H,2) + WD + WS + poly(DayRatio,2)

4.4 Revised Model 3

Now we add a quadratic term for wind direction. The revised model 3 takes the form

$$R = \beta_0 + \beta_1 \mathrm{H} + \beta_2 \mathrm{H}^2 + \beta_3(\mathrm{WD}) + \beta_4(\mathrm{WD})^2 + \beta_5(\mathrm{WS}) + \beta_6(\mathrm{DayRatio}) + \beta_7(\mathrm{DayRatio})^2 + error. \quad (4)$$

Table 6: Summary of the fitted linear regression for model R ∼ poly(H,2) + poly(WD,2) + WS + poly(DayRatio,2).

	Dependent variable:	
	R	
	train dataset	complete dataset
	(1)	(2)
poly(H, 2)1	-311.334^{***}	-413.490^{***}
	(42.450)	(40.641)
poly(H, 2)2	-194.111^{***}	-259.527^{***}
	(38.073)	(36.105)
poly(WD, 2)1	-122.786^{***}	-147.307^{***}
	(40.030)	(38.951)
poly(WD, 2)2	136.706^{***}	160.461^{***}
	(38.065)	(36.349)
WS	14.426^{***}	14.605^{***}
	(3.218)	(2.519)
poly(DayRatio, 2)1	353.716^{***}	404.964^{***}
	(42.741)	(40.560)
poly(DayRatio, 2)2	-128.389^{***}	-138.061^{***}
	(41.529)	(39.094)
Constant	127.882^{***}	119.750^{***}
	(20.433)	(15.975)
Observations	79	118
R^2	0.765	0.799
Adjusted R^2	0.742	0.786
Residual Std. Error	37.130 (df = 71)	35.189 (df = 110)
F Statistic	33.028^{***} (df = 7; 71)	62.461^{***} (df = 7; 110)
Note:		*p<0.1; **p<0.05; ***p<0.01

The summary of the fitted model 3 described in Eq.(4) is provided in Table 6 (see the train dataset column). Note that all regression coefficients are significant at less than 1%. The coefficient of determination for this model is $R^2 = 76.5\%$ which shows an increase of 4.3% from the previous model. The residual sum squares has dropped to RSS = 97885 from RSS = 115666.7 for the previous model. The ANOVA lack-of-fit test comparing this model with the revised model 2 gives F-statistic $F = 12.9$ on (1, 71) DF and has a two-sided p-value 0.0602% which indicates that the null-hypothesis (H_0: the revised model 2 is true) is rejected at level less than 0.1%. As will be discussed later, this model provides the best performance compared with all other linear regression models we tested in this project. The results of the same model (Eq.(4)) fitted on the complete dataset is also provided in the second column of Table 6 for comparison purposes and as can be seen, for instance, $R^2 = 79.9\%$ for the complete dataset which shows that the proportion of variability explained by this model is very significant and therefore we may have found a useful model provided that all other diagnostic tests are OK. Next we will look at the basic diagnostic plots for this model.

4.4.1 Basic Diagnostic Checks for Revised Model 3

Basic diagnostic plots for model R ~ poly(H, 2) + poly(WD, 2) + WS + poly(DayRatio, 2) are shown in Figure 11. No noticeable patters or fan-shape behaviour are detected in *Residuals vs Fitted* and *Scale-Location* plots. Hence, there is no monotonic variance change. The *Normal Q-Q* plot demonstrates a satisfactory shape. The *Residuals vs leverage* plot shows that none of the observations have a strong influence on the fit.

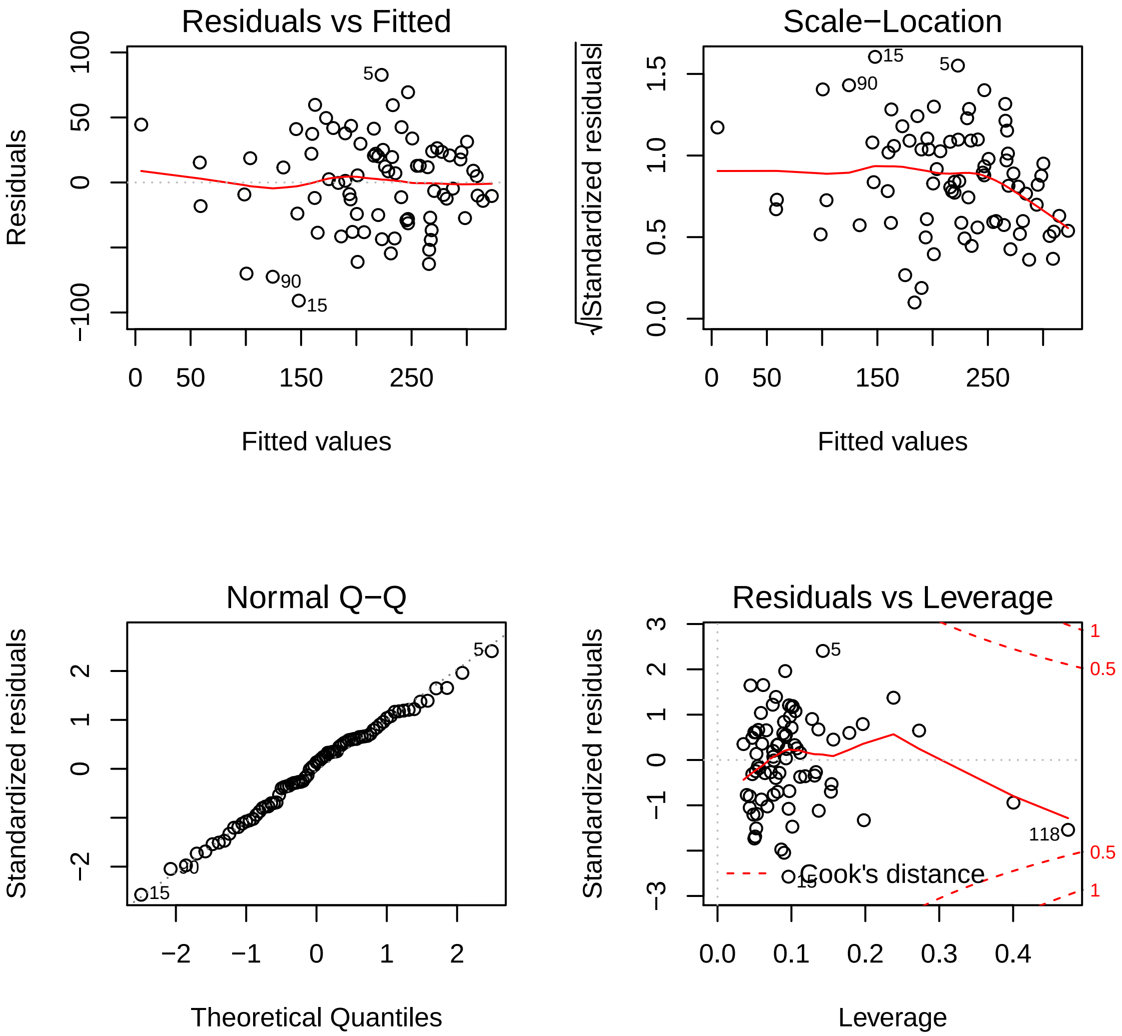

Figure 11: Basic regression diagnostic plots for model R ~ poly(H,2) + poly(WD,2) + WS + poly(DayRatio,2).

4.4.2 Residual Dependency Checks for Revised Model 3

Residual dependence plots are shown for the revised model 3 in Figure 12. The loess curves are approximately flat for the explanatory variables of this model (day ratio, humidity and wind direction) and hence the model has sufficiently handled the nonlinearity.

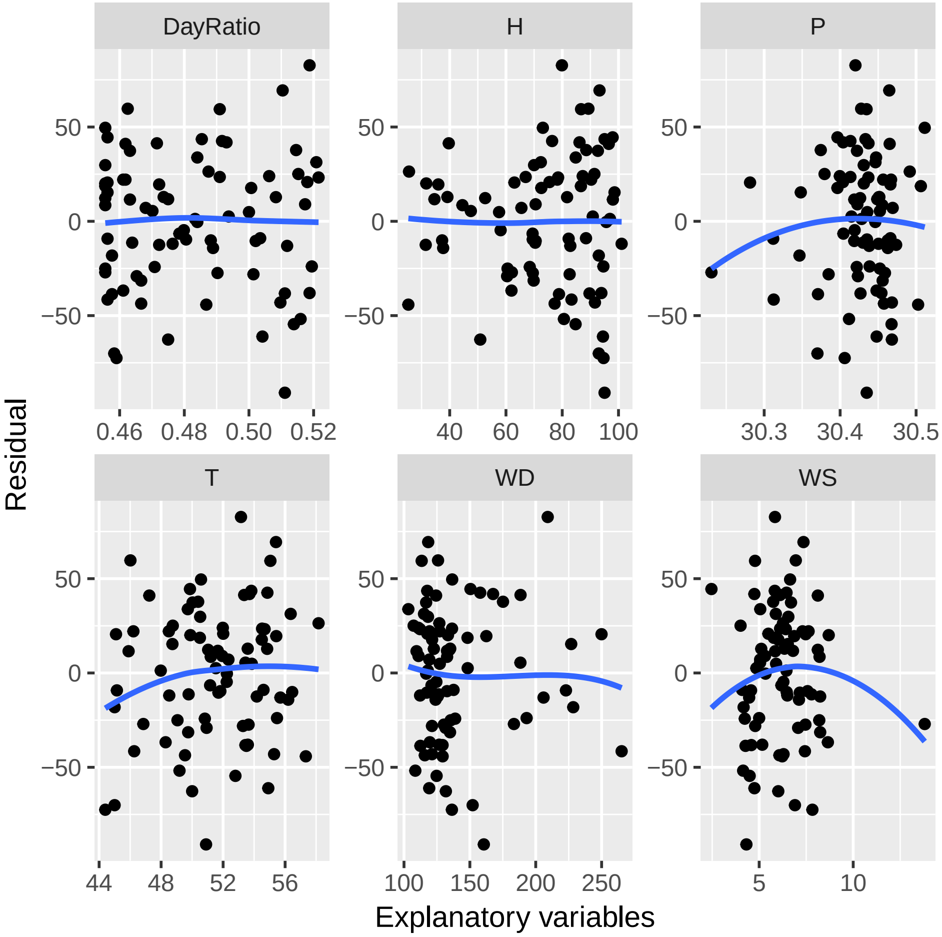

Figure 12: Residual dependency plots for model R ∼ poly(H,2) + poly(WD,2) + WS + poly(DayRatio,2)

Next, we compare the AIC and BIC values of all the above four models. Table 7 summarizes these results and shows that the revised model 3 has the lowest AIC and BIC values and are preferred based on these criteria. Backward stepwise/stagewise methods based on AIC and BIC criteria also suggests that the revised model 3 has the best performance (not shown).

Table 7: Summary of the AIC and BIC values for different models.

	initial model	revised model 1	revised model 2	revised model 3 (best model)
AIC	838.19	821.37	816.02	804.84
BIC	852.41	837.96	834.98	826.16

Shapiro-Wilk and Jarque-Bera normality tests give p-values of 83.94% and 63.03%, respectively, and therefore the null-hypothesis (H_0: data are from a normally distributed population) cannot be rejected. In other words, there is no evidence that the data are not from a normally distributed population as far as these tests are concerned.

Durbin-Watson test is useful to detect serial correlation. The null-hypothesis for this test is H_0: there is no correlation among residuals (no serial time dependence). We performed this test using both normal and bootstrap methods and the p-values are 67% and 71.4%, respectively. The D-W statistic is $d = 1.92$ and the sample autocorrelation of the residuals is $r = 0.03$. Thereby, the null-hypothesis cannot be rejected which means there is no evidence to support possible serial dependence in the residuals based on this test.

4.5 Overfitting Lack-of-Fit Test

In this section we overfit the best model (revised model 3) and perform various tests to judge whether the overfitted model is useful or not. To make the model overfit, let us add a quadratic term for wind speed. So the model can be represented as

$$R = \beta_0 + \beta_1 \mathrm{H} + \beta_2 \mathrm{H}^2 + \beta_3(\mathrm{WD}) + \beta_4(\mathrm{WD})^2 + \beta_5(\mathrm{WS}) + \beta_6(\mathrm{WS})^2 + \beta_7(\mathrm{DayRatio}) + \beta_8(\mathrm{DayRatio})^2 + error. \quad (5)$$

The summary of the fitted model described by Eq.(5) is provided in Table 8. We notice that based on the two-sided p-value for the t-test which is 21.1%, the quadratic term for wind speed is not significant at 10% and therefore should be dropped. The R^2 has increased by only 0.5% from the previous model to 77% and the residual sum squares has dropped to RSS = 95706 from RSS = 97885 for the revised model 3.

The ANOVA lack-of-fit test comparing this model with the revised model 3 gives F-statistic $F = 1.59$ on (1, 70) DF and has a two-sided p-value 21.1% (the same as p-value for the t-test discussed above) which implies that the null-hypothesis (H_0: the revised model 3 is true) cannot be rejected at 10%. The BIC values are $\mathrm{BIC}_3 = 826.16$ and $\mathrm{BIC}_{\mathrm{overfit}} = 828.75$ for the revised model 3 and the overfitted model, respectively, which further reinforces the conclusion obtained by the ANOVA test and suggests that the simpler model (revised model 3) is preferred. We have also tried other possibilities for the overfit model such as adding temperature or pressure and found out that adding these predictors substantially deteriorate the performance as compared with the revised model 3 and these overfitted models can be easily rejected using ANOVA or AIC/BIC tests (not shown).

Table 8: Summary of the fitted linear regression for model R ∼ poly(H,2) + poly(WD,2) + poly(WS,2) + poly(DayRatio,2).

	Dependent variable:
	R
poly(H, 2)1	-302.863^{***}
	(42.803)
poly(H, 2)2	-196.338^{***}
	(37.956)
poly(WD, 2)1	-116.810^{***}
	(40.144)
poly(WD, 2)2	124.587^{***}
	(39.104)
poly(WS, 2)1	201.173^{***}
	(44.244)
poly(WS, 2)2	-49.946
	(39.564)
poly(DayRatio, 2)1	354.764^{***}
	(42.572)
poly(DayRatio, 2)2	-120.891^{***}
	(41.781)
Constant	217.543^{***}
	(4.160)
Observations	79
R^2	0.770
Adjusted R^2	0.744
Residual Std. Error	36.976 (df = 70)
F Statistic	29.341^{***} (df = 8; 70)
Note:	*p<0.1; **p<0.05; ***p<0.01

5 Comparison with Other Methods

We showed in the previous sections, following a step by step approach, that the revised model 3 (best model) provides superior performance compared with all other linear regression models considered in this project. In this section, we will compare our best model with two penalized regression models as well as a Random Forest model.

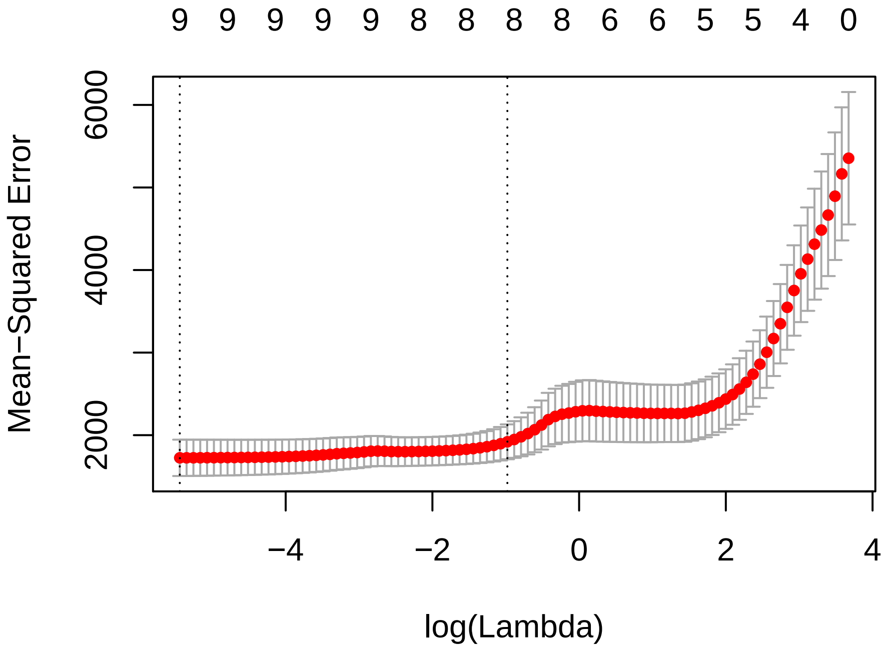

Figure 13: Variations of the mean-squared error as a function of log(λ) for L1-penalized (LASSO) regression. The selected tuning parameter is $\lambda = 0.38$.

In order to be able to make proper comparison between penalized regression models and the best model, we needed to redefine the design matrix to include the three quadratic variables for humidity, wind direction and day ratio. For penalized regressions, we used *glmnet* package which, by default, implements regularized 10-fold cross-validation to select the tuning parameter λ. Figure 13 shows the variation of mean-squared error as a function of λ for L1-penalized (LASSO) regression. The tuning parameter $\lambda = 0.38$ is chosen based on the one-standard-deviation rule, that is the simplest model with average mean-squared error within one standard deviation of the lowest mean-squared error is selected. Based on the one-standard-deviation rule, a model with eight predictors is chosen (quadratic term in day ratio is removed).

The results of L2-penalized (Ridge) regression is shown in Figure 14. The tuning parameter chosen based on the one-standard-deviation rule is $\lambda = 70.46$ which corresponds to a model with all the nine predictors (six linear terms plus three quadratic terms) retained in the model.

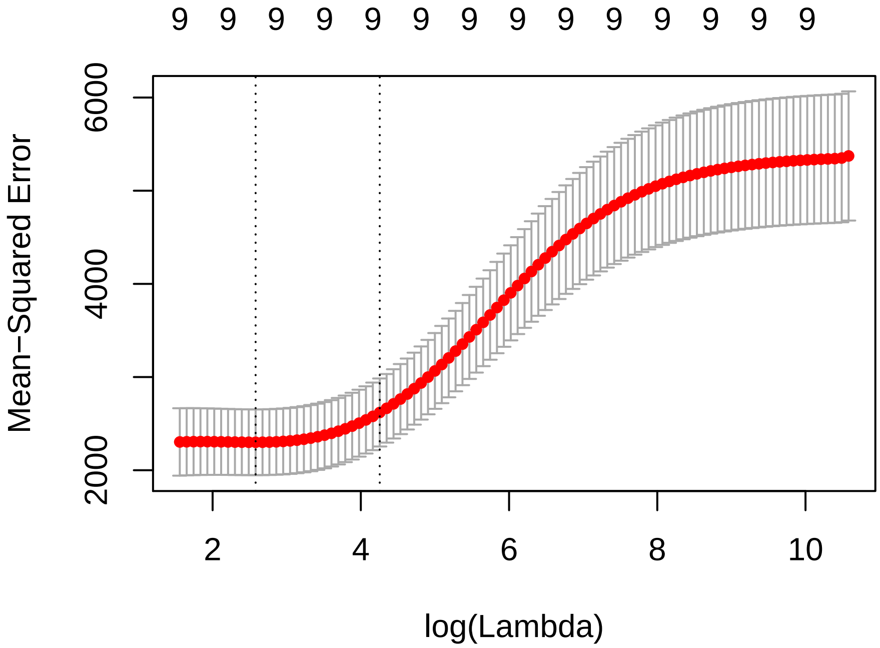

Figure 14: Variations of mean-squared errors as a function of $\log(\lambda)$ for L2-penalized (Ridge) regression. The selected tuning parameter is $\lambda = 70.46$.

As the last step, we use one of the machine learning technique, the Random Forest method. The results of the root-mean-square errors (RMSE) for all the above methods are summarized in Table 9. As can be seen, the best model (revised model 3) has the best performance on prediction of test dataset. However, as expected, Random Forest shows better performance on the train dataset. The best model also performs much better than both L1/L2 penalized regression models on both test and train datasets. The model selected by the backward stagewise method is identical to the best model and thus their performances are also identical.

Table 9: Summary of the root-mean-square errors.

	Best Model	BackStage	RidgeReg	LASSO	Random Forest
test	32.478	32.478	47.283	33.729	33.688
train	35.200	35.200	48.338	36.870	19.214

The last plot (Figure 15) shows the values predicted for the radiation by the best model and their corresponding actual values from the test dataset. The line $y = x$ is added to assist with analyzing the performance of the model. As can be seen, the points are clustered around the line $y = x$ which further shows that the best multiple linear regression model we built in this project is actually useful for prediction purposes on a new dataset (test dataset).

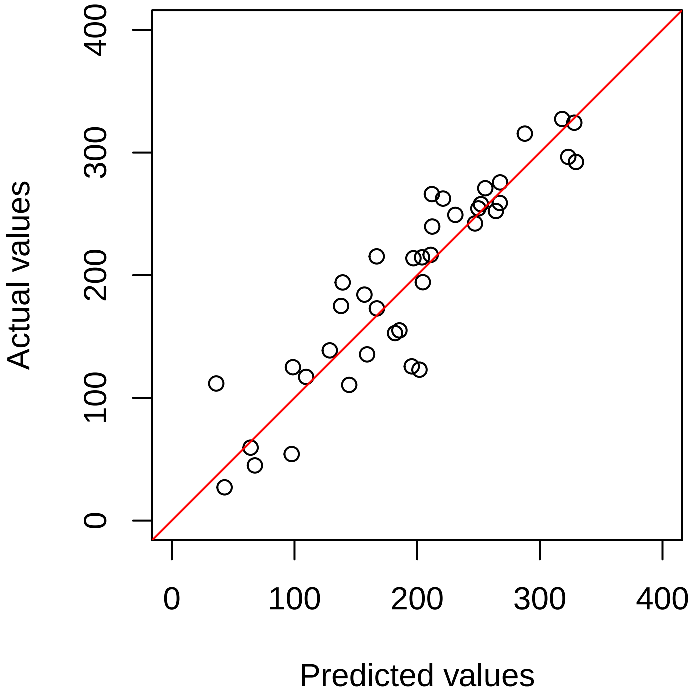

Figure 15: Actual response variable values of the test dataset versus their corresponding values predicted by the revised model 3 (the best model).

6 Conclusions

The main goal of this project was to build a model using multiple linear regression for prediction of solar radiation as a function of some basic meteorological variables as well as information about the length of a day. We used the dataset that was collected recently (less than a year ago) by NASA HI-SEAS mission in a period of four months on an isolated site located on the Big Island of Hawaii. The data was made available recently and thus is considered as a live dataset.

Through a rigorous analysis, we constructed a model using multiple linear regression. We discovered that temperature is a confounding variable as it is strongly correlated with the output variable (radiation) and all the predictors excluding wind variables. This causes a spurious association between temperature and radiation. This is in agreement with the findings of Sun et al. (2015) who suggested that day ratio is a much more important variable than temperature as one may intuitively consider it as an important variable. We found that the best model has the form R ~ poly(H, 2) + poly(WD, 2) + WS + poly(DayRatio, 2) with coefficient of determination $R^2 = 76.5\%$ and $R^2 = 79.9\%$ on train and complete datasets, respectively. The

R^2 of the best model is quite significant which shows that this model explains large proportion of variability of the data.

We performed various diagnostic tests to check the validity of the model as well as to verify whether the model obeys the standard assumptions of linear regression such as normality, constant variance and no serial correlation. In particular, we implemented Wilk-Shapiro and Jarque-Bera tests for normality as well as Durbin-Watson test for serial correlation. All of these tests showed that the dataset and the best model complied really well with the linear regression assumptions.

In order to arrive at the best model, we used many variable selection methods such as, iterative model building, stepwise/stagewise and best subset regression using AIC/BIC or repeated hold cross-validation, and L1/L2 penalized regression. In addition, we used Random Forest method as a benchmark for comparison purposes. We found that the best model has the superior prediction performance (lowest RMSE) on the test dataset against all other methods. However, as expected, the Random Forest has the best prediction performance on the train dataset. It was also shown that the best model performs better on both test and train datasets when compared with L1-penalized (LASSO) and L2-penalized (Ridge) regression models.

We had no reason to drop wind variables (wind direction and wind speed) in the best model from statistical point of view, however we suspect that these variables may be just manifestation of some lurking variables. It is hard to imagine that radiation and wind direction or wind speed are causally related. For instance, a high wind speed could be an indication of a sunny sky. If the dataset can be extended to at least one full year and/or more input variables such as cloud cover, precipitation and solar zenith angle of the sun can be considered, more accurate model may be constructed. It may be also useful to build a model that can provide accurate hourly average solar radiation prediction based on the explanatory variables using time series regression.

References

HI-SEAS. 2017. "Hawai'i Space Exploration Analog and Simulation." http://hi-seas.org/?page_id=5990.

Incropera, F. P., and D. P. DeWitt. 2002. *Fundamentals of Heat and Mass Transfer.* John Wiley & Sons.

McLeod, A. I. 2017. "Regression: Variable Selection." Lecture Note 05. London, Ontario: Western University.

McLeod, A.I., and Changjiang Xu. 2017. *Bestglm: Best Subset Glm.* https://CRAN.R-project.org/package=bestglm.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sonntag, R. E., C. Borgnakke, and G. J. Van Wylen. 2003. *Fundamentals of Thermodynamics.* John Wiley & Sons.

Sun, Huaiwei, Na Zhao, Xiaofan Zeng, and Dong Yan. 2015. "Study of Solar Radiation Prediction and Modeling of Relationships Between Solar Radiation and Meteorological Variables." *Energy Conversion and Management* 105 (Supplement C): 880–90. doi:https://doi.org/10.1016/j.enconman.2015.08.045.