Using Multiple Linear Regression Modelling to Predict Daily Average Solar Radiation Levels In Hawaii

Nathaniel Albano, Evan Lum, Kenneth Chow, Elise Pham, Jisu Chae, Mukhil Guna, Jasmine Chu

Stats 101A - Lecture 1

Dr. Maria Cha

9 March 2024

Section 1: Introduction

Given a meteorological dataset, are we able to predict the level of Solar Radiation based on Time, Radiation, Temperature, Pressure, Humidity, and Wind? Solar Radiation is an important measure to take into account when analyzing problems such as global warming and solar energy technology. To predict the level of Solar Radiation, a research station in Hawaii–the HI-SEAS weather station–collected meteorological data spanning two months in 2016, from September 29th to December 1st. The original data consists of 32,686 observations and 11 variables: UNIXTime, Data, Time, Radiation, Temperature, Pressure, Humidity, Wind Direction, Speed, TimeSunRise, and TimeSunSet. During the process of deciding a method to model the relationship, raw data appeared to deviate from the regression model. Therefore, our group chose a power-transformed multiple linear regression model to model it. This final report encompasses three different sections besides the introduction–data description, results and interpretation, and discussion. Data cleaning (removing redundant time columns), violation of assumptions, summary statistics, and graphs used to find the distribution of each variable and relationships are discussed in the data description. Following the data description, the best predictive model is determined through the use of box-cox transformation and the interpretation of R. After selecting and assessing the model using diagnostic tools, the discussion section addresses the limitations of the model and provides suggestions for future improvements by describing how some variables could affect our predictive model.

Section 2: Data Description

After removing unnecessary columns from the dataset (Table 2.1), we noticed that there were some issues with the time variable, and decided to modify it. The time goes on without measuring the day. Hence the data was modified so that each day starts at 0 minutes and ends at 1440 minutes (Figure 2.1)

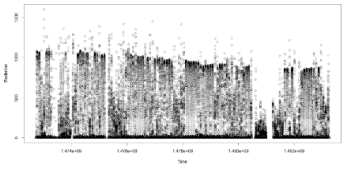
Table 2.1 - Data Cleaning

Unix Time (Seconds)	Data	Time	Radiation (W/m^2)	Temperature (F)	Pressure (Hg)	Humidity (%)	Wind Direction (Degree)	Speed (MPH)	Time Sun Rise

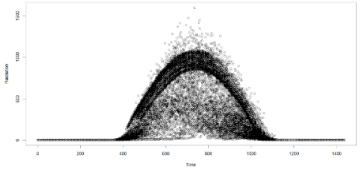
Cleaned Data

Time (Mins)	Radiation (W/m^2)	Temperature (F)	Pressure (Hg)	Humidity (%)	Wind Direction (Degree)	Speed (MPH)

Figure 2.1 - Modifying Time



Before



After

Initially our dataset had multiple time columns like *UNIXTime*, *Data*, and *TimeSunRise* . We eventually decided that just the *Time* column would be sufficient as all the previous columns explained the same measurement just in different terms. Each observation for the *Time* column was recorded in five minute intervals that spanned over the course of two months. However we encountered many problems when we ran our model and looked at the diagnostic plots.

Equation for summary table and diagnostic plot.

$$Radiation = (Temperature)X_1 + (Pressure)X_2 + (Humidity)X_3 + (Wind.Direction.Degree)X_4 + (Speed)X_5 + (Time)X_6$$

Table 2.2 - Summary Table

Coefficients	Estimate	Standard Error	T-Value	Pr(>|t|)	Significant
(Intercept)	$1.966 * 10^4$	$6.820 * 10^2$	28.830	$< 2 * 10^{-16}$	Yes
Temperature	$3.995 * 10^1$	$2.062 * 10^{-1}$	193.738	$< 2 * 10^{-16}$	Yes
Pressure	$-7.043 * 10^2$	$2.244 * 10^2$	-31.384	$< 2 * 10^{-16}$	Yes
Humidity	$-1.468 * 10^2$	$4.788 * 10^{-2}$	-0.307	0.759	No
Wind Direction (degree)	$-2.732 * 10^1$	$1.431 * 10^{-2}$	-19.098	$< 2 * 10^{-16}$	Yes
Speed	$7.694 * 10^0$	$3.343 * 10^{-1}$	23.012	$< 2 * 10^{-16}$	Yes
Time	$-1.064 * 10^2$	$2.803 * 10^{-3}$	-37.969	$< 2 * 10^{-16}$	Yes

Using the equation (above), we can analyze the summary table (Table 2.2). We take particular notice of humidity as it is the only predictor that is not significant with a p-value of 0.759. Furthermore, plotting the diagnostic plots (Figure 2.2), it is apparent that the assumptions for a linear model are violated. This includes: linearity (Residual vs Fitted), normality (Q-Q Residuals), constant variance (Scale-Location). Also note that there are a lot of leverage points (Residuals vs Leverage). After doing some analysis of the model, we decided that it was impossible to fit it into a linear regression model, without some changes to the data. Upon further inspection of the data, we realized that each data point being recorded in five minute intervals will lead to autocorrelation. A data point at 2:05 pm will be heavily correlated with the following data point at 2:10 pm, and 2:15 pm, and so on and so forth. Hence, we decided to separate the data points by day and average each measure so we can get the daily average of each variable (with the exception of the wind variable, which is recorded in degrees and had its average calculated using circular mean). This allows us to create a linear regression that models the prediction of average radiation of a day as our linearity assumptions are no longer being violated.

Figure 2.2 - Diagnostic Plots

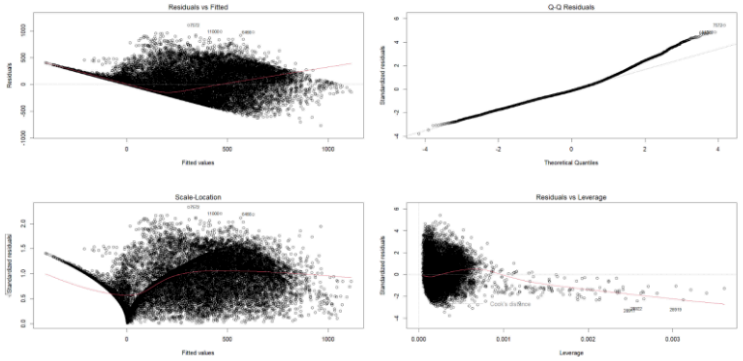
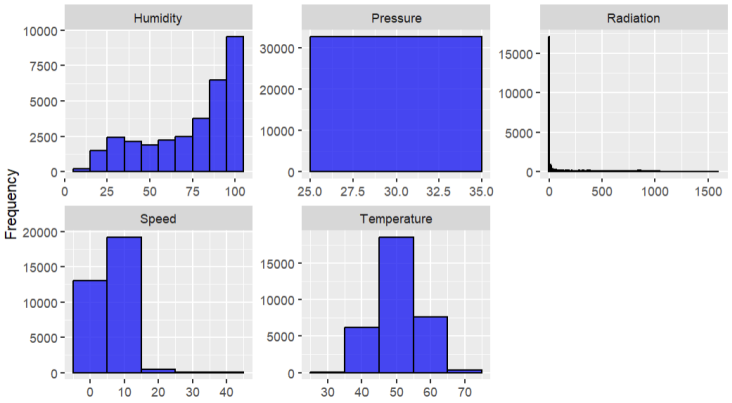


Table 2.3 - Mean and Standard Deviation

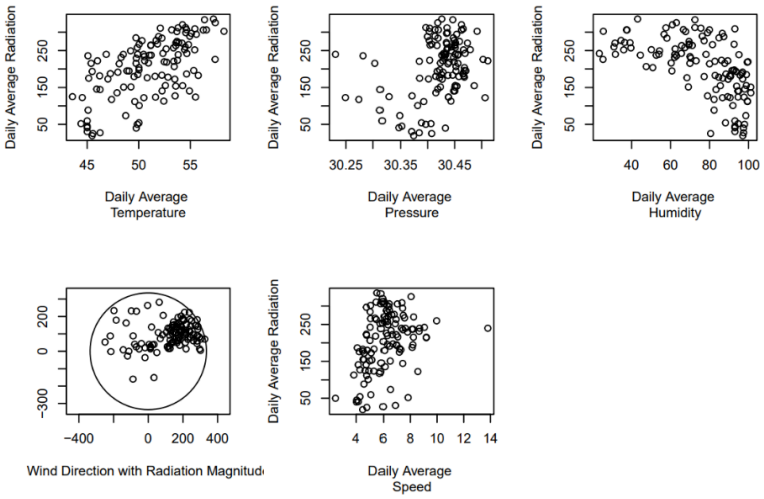
	Mean	Standard Deviation
Average Radiation	284.430	80.071
Average Temperature	51.070	3.548
Average Pressure	30.422	0.050
Average Humidity	75.396	20.862
Average Wind Speed	6.204	1.543

Figure 2.3 - Distribution of each variable



After finalizing the data cleaning, we get the following distribution as shown in, Figure 2.3.

Figure 2.5 - Daily Averaged Predictors vs Daily Averaged Responses



Within our predictor vs residuals plots, we can see many visible trends (the most obvious being the Temperature and Humidity variables). It should be noted that the Wind Direction variable is displayed as a directional scatter plot. The points are orientated around the center similar to the face of a compass, with the top-most point of the circle representing North, and the right-most point of the circle representing East. Subsequently, each point is a single day, with the distances from the center representing the average level of radiation, with the center being 0 and increasing the farther you get. By examining the plot, we can see a majority of the points being plotted in the north-east direction, with a random spread of points elsewhere. Hawaii winds generally point to the north-east due to trade winds [discussed in length within the real-world impact section], and because we do not see any visible patterns or trends for solar radiation outside of the north-east direction, we can conclude that we should treat wind-direction as a constant and remove it from our model. We then continue with further analyzing the other variables.

Section 3: Results and Interpretation

Figure 3.1 - Full Daily Average Linear Model Summary

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1672.9885  3540.0128  -0.473   0.6374
avg_temp       12.9978     1.8582   6.995 1.99e-10 ***
avg_pres       37.6968   117.7470   0.320   0.7494
avg_humid      -0.6817     0.2965  -2.299   0.0234 *
avg_speed      19.0520     3.6304   5.248 7.28e-07 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 51.74 on 113 degrees of freedom
Multiple R-squared:  0.5967, Adjusted R-squared:  0.5824
F-statistic: 41.79 on 4 and 113 DF,  p-value: < 2.2e-16
```

Figure 3.1 shows the summary of our full multiple linear regression model. Our R^2 value is 0.5967 with an overall F-statistic of less than 2.2e-16. However, examining the p-values of each coefficient shows that the pressure variable has a large p-value of 0.7494.

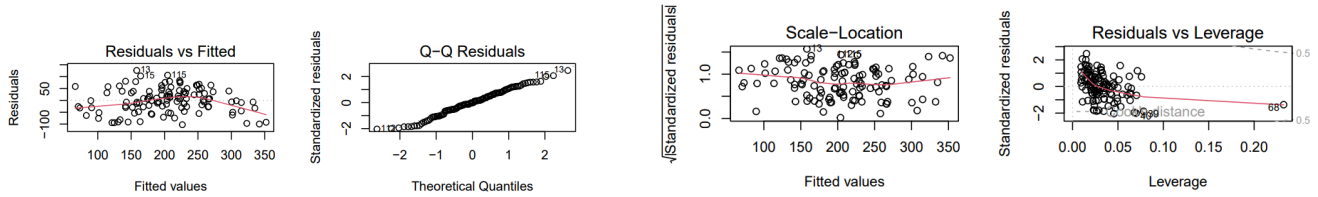
Figure 3.2 - Partial F-Test Summary [Pressure]

Figure 3.2 shows the summary of our partial F-test of our model reduced by the pressure variable against our full model, which shows a high p-value of 0.7494. This means we cannot reject the null hypothesis that our full model with the pressure variable makes a significant difference, and can use the reduced model

```
## Analysis of Variance Table
##
## Model 1: avg_radi ~ avg_temp + avg_humid + avg_speed
## Model 2: avg_radi ~ avg_temp + avg_pres + avg_humid + avg_speed
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    114 302831
## 2    113 302556  1    274.43 0.1025 0.7494
```

going forward.

Figure 3.3 - Daily Average Linear Model Diagnostic Plots



From viewing Figure 3.3, although there is some slight curvature within the red line of the 'Residuals vs Fitted' plot and 'Scale-Location' plot, we can see that none of our assumptions of linear regression are being violated. Within the 'Residuals vs Leverage' plot, there is a data point [Point 68 recorded on December 16, 2016] that is noticeably farther from the rest of the cluster. Closer analysis of data point 68 shows that it was recorded in the middle of a winter storm (Phys.org) during a particularly windy day for the Mauna Loa (Big Island Now), where the HI-SEAS weather station is located. Despite that, point 68 does not fall outside the standardized residual range of -2 and +2 which means we can generally classify it as a good leverage point, and we can continue with our model.

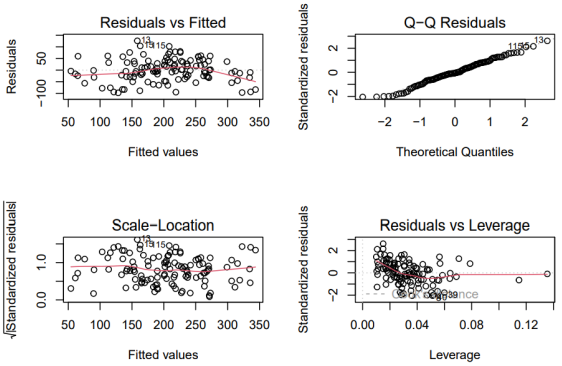
Examining the Response vs Predictors plot, we can see that the humidity and speed plots, although showing a visible trend, are curved and not linear. Therefore, we apply a box-cox power transformation.

Figure 3.4 - Power Transformation Summary

```
## bcPower Transformations to Multinormality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## avg_radi     1.2362           1       0.9434       1.5290
## avg_temp     1.8545           1      -0.3985       4.1075
## avg_humid    2.2234           2       1.6001       2.8467
## avg_speed   -0.0448           0      -0.4903       0.4006
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                  LRT df       pval
## LR test, lambda = (0 0 0 0) 153.3553  4 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                LRT df       pval
## LR test, lambda = (1 1 1 1) 41.0033  4 2.6836e-08
```

Figure 3.5 displays our transformed model summary. Each variable is statistically significant, and our F-statistic of the whole model is statistically significant. It should be noted that our R^2 value has improved from 0.5967 to 0.6404, showing that our transformed model better fits the data.

Figure 3.6 - Transformed Model Diagnostic Plots



The power transformation summary (Figure 3.4) supports our visual estimations and recommends that we keep the radiation and temperature untransformed, while squaring the humidity variable and logging the speed variable.

Figure 3.5 - Transformed Model Summary

```
Call:
lm(formula = t_radi ~ t_temp + t_humid + t_speed)

Residuals:
    Min      1Q  Median      3Q     Max
-97.907 -29.381  -3.102  33.728 126.060

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.361e+02  1.051e+02  -6.051 1.88e-08 ***
t_temp       1.274e+01  1.496e+00   8.516 7.64e-14 ***
t_humid     -6.161e-03  2.043e-03  -3.016  0.00316 **
t_speed      1.266e+02  2.161e+01   5.858 4.62e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 48.65 on 114 degrees of freedom
Multiple R-squared:  0.6404, Adjusted R-squared:  0.6309
F-statistic: 67.66 on 3 and 114 DF,  p-value: < 2.2e-16
```

Figure 3.6 shows our transformed model diagnostic plots. We can notice improvement in the curvature for the 'Residuals vs Fitted' and 'Scale-Location' plots, with the red line better conforming to a straight horizontal line. Our 'Residuals vs Leverage' plot also shows that our previous leverage points are not as severe. Therefore, transforming our model better fits

the assumptions for linear regression.

Our F-test shows very high significance as expected since all predictors show to be significant. Nonetheless, we test for multicollinearity.

Figure 3.7 - Correlation Matrix

```
##Check for correlation/ multicollinearity
cor(avg_solar[,-c(1,2,4,5)])
```

```
##             avg_temp    avg_wind   avg_speed
## avg_temp    1.0000000  0.14028065 -0.13910977
## avg_wind    0.1402806  1.00000000 -0.03840152
## avg_speed -0.1391098 -0.03840152  1.00000000
```

Figure 3.7 displays our correlation matrix. We see that the variables do not have high correlation. This is a great sign as it indicates that our coefficient estimates are reliable. Wind and Speed seem to have the smallest correlation while Wind and Temperature show to have the highest correlation.

Figure 3.8 - VIF's

Figure 3.8 shows us the Variance Inflation factor among the three of our variables. They all show to be < 5 indicating that our coefficients are sufficiently estimated and we do not have reason to suspect multicollinearity.

```
vif(t_model)
```

```
##   t_temp  t_humid  t_speed
## 1.393605 1.649894 1.358375
```

Figure 3.9 - Finding the best subset model

```
##      adj_r2      AIC     AICC      BIC
## 1 0.3755985 980.7768 981.3031 986.3182
## 2 0.6049255 927.7443 928.2706 936.0564
## 3 0.6309066 920.6868 921.2131 931.7696
```

The table in Figure 3.9 shows us that R^2adj, AIC, AICC, and BIC all indicate our best subset model would be with p = 3. With the highest adjusted R-squared and the lowest AIC, AICC, and BIC values. This means that this model would account for the greatest proportion of variance in Radiation.

Therefore, our final equation is:

$$Radiation_{avg} = -636.1 + 12.74 * Temperature_{avg} - .006161 * Humidity_{avg}^{2} + 126.6 * log(Wind\ Speed_{avg})$$

From this equation, we can see that for every degree increase in Temperature, we see an increase in Radiation levels of 12.74. For every increase in square root percent increase in Humidity, Radiation levels decrease by 0.00616. And finally, for every 10% increase of Wind Speed, Radiation levels will increase by 5.24 [log(1.1) * 126.6 = 5.24].

Section 4: Discussion

Summary

In conclusion, our analysis demonstrates the effectiveness of using a linear regression model to predict solar radiation levels based on variables: time, temperature, pressure, humidity, and wind speed. After a rigorous process of data cleaning and careful consideration of various predictive models, we uncovered some meaningful relationships between these environmental factors and solar radiation.

Our findings reveal that temperature, wind speed, and humidity significantly influence the intensity of

solar radiation in Hawaii. We observed positive relationships with temperature and wind speed and a negative one with humidity when we split the dataset into daily average radiations. This means that solar radiation is stronger when the temperature is higher, when wind is stronger, or when it is less humid. However, you may ask, why does all this matter?

Real-world Impact

After analyzing our results, we concluded that our research aligns with real-world natural phenomena. When the sun reaches its zenith during the day, radiation and temperature are also at their highest. Additionally, according to Falayi E.O. (2013), humidity can cause the formation of clouds, which block direct sunlight and reduce radiation toward the Earth's surface. Some variables also directly impacted our predictive model. For instance, we removed the wind direction variable from our model due to its lack of predictive power for radiation. This is largely due to Hawaii's trade winds – winds caused by the difference in air pressure between the equator and the upper atmosphere over the North Pacific (Pacific Disaster Center). These winds move northeast along the ocean's surface toward Hawaii and are supported by our data, as seen in Figure 2.3 in the "Wind Direction Radiation Magnitude" table. Therefore, we have valid reason to believe it should be treated as a constant.

One application of our project to current issues is in climate change research. Solar radiation plays a significant role in influencing the Earth's climate, making our project relevant to addressing the increasing threat of global warming. While solar radiation does not cause global warming, we can measure how it interacts with the main contributing factors (NASA). For instance, large aerosol particles in the Earth's atmosphere tend to absorb radiation, causing it to feel warmer. Therefore, we could track temperature data, analyze the impact of solar radiation, as well as develop predictive visualizations that graph future climate scenarios based on changes in temperature, humidity, and solar radiation in regions all over the world.

Limitations & Future Studies

Lastly, it is important to acknowledge the limitations of our analysis. A major concern we had when creating a linear regression model with the original dataset was autocorrelation. To alleviate this problem, we split the data points by day and averaged them to get the daily average of each variable. However, we still will likely have some autocorrelation, as each day is recorded sequentially after another. Furthermore, this model may work for surface-level analyses/predictions, but we likely diminished underlying relationships that were present when they were separate data points. With this being said, future research should include more variables, a wider measured period of twelve months, and more regions surrounded by diverse climates and geographical latitudes. By incorporating a greater variety of observations, our solar radiation level prediction and analysis would dramatically increase in effectiveness.

References

E.O., Falayi. "The Impact of Cloud Cover, Relative Humidity, Temperature and Rainfall on Solar Radiation in Nigeria." *Energy and Power*, Scientific & Academic Publishing, 2013, article.sapub.org/10.5923.j.ep.20130306.03.html. Accessed 14 Mar. 2024

"Hawaii Trade Winds | Kona Winds Hawaii | High Wind in Hawaii - PDC." *Pacific Disaster Center*, pdc.bumpnetworks.com/iweb/high_wind.jsp?subg=1. Accessed 14 Mar. 2024.

"Is the Sun Causing Global Warming?" *NASA*, NASA, climate.nasa.gov/faq/14/is-the-sun-causing-global-warming/#:~:text=No.,goings%20of%20the%20ice%20ages. Accessed 14 Mar. 2024.

"Kona Winds Prompt Advisories for Big Island" *Big Island Now*, https://bigislandnow.com/2024/03/13/us-army-issues-big-island-convey-alerts-for-wednesday-through-friday/ Accessed 15 Mar. 2024

"Snow blankets Hawaii summits amid winter storm warning" *Phys.org*, The Associated Press, https://phys.org/news/2016-12-blankets-hawaii-summits-winter-storm.html Accessed 15 Mar. 2024