# PREDICTION OF SOLAR RADIATION

Elianne Mora

Advance Regression and Prediction

# Table of Contents

# Introduction

Given the current climate situation, many countries are leading by example and innovation in their implementation of renewable energy sources. From geothermal, biomass, and hydro to wind, tidal and solar, renewable energy execution must be carefully planned and analyzed in order to outweigh the costs. For example, it is no secret that Seattle's autumn and winter days are especially cloudy and rainy; therefore, solar energy during these seasons will likely not be able to fully power a household if at all, so a different or supplemental renewable source may need to be considered.

In this project we explore meteorological data from the Hawaii Space Exploration Analog (HI-SEAS) station from four months (September – December 2016). Meteorological and time data were recorded every five minutes for a total of 32,686 observations and eleven variables defined as:

| ID | Description |
|---|---|
| *UNIXTime* | Time stamp in seconds (timestamp initialized 01-01-1970) |
| *Data* | Date in dd/mm/yyyy hh:mm format |
| *Time* | Time in hh:mm:ss format (24hr) |
| *Radiation* | Watts per square meter |
| *Temperature* | Degrees Fahrenheit |
| *Humidity* | Percentage |
| *Pressure* | Barometric pressure in Hg |
| *WindDirection.Degrees* | Wind direction in degrees |
| *Speed* | Wind speed in miles per hour |
| *TimeSunRise* | Hawaii recorded sunrise time |
| *TimeSunSet* | Hawaii recorded sunset time |

*Table 1. Original variable descriptions*

In this project we are interested in the predictive capabilities of different regression methods, where our target variable is *Radiation*. This analysis will help us understand and better evaluate the feasibility of implementing solar power in the island. For this reason, in this portion of the project we focus on predicting the average daily radiation during daylight for these 4 months.

# Data Processing and Visualization

Upon importing the dataset, it is noticed that there are no missing values and that all time-related variables (*UNIXTime, Data, Time, TimeSunRise, TimeSunSet*) have been imported as factors. We first focus on the time stamp variable, which, once converted into date, is used to create the variable *DayofYear*. We eliminate all measurements taken from the first month of 2017 (132 observations only), so that our range stays within 2016, from the 245$^{th}$-366$^{th}$ day. Also, the sunrise and sunset variables are properly converted into hours so that a *Daylight* variable is created, comprising the total hours of sunlight per day. The final dataset is as follows:

| ID | Type |
|---|---|
| *Radiation* | Numeric |
| *Daylight* | Numeric |
| *DayofYear* | Numeric |
| *Temperature* | Integer |
| *Pressure* | Numeric |
| *Humidity* | Integer |
| *WindDirection.Degrees* | Numeric |
| *Speed* | Numeric |

*Table 2. Final variables.*

In the plot below we can observe the fluctuations in the average daily temperatures and radiation. While there are significant variations, the overall trends between both temperature and radiation seem to follow a remarkably similar pattern.

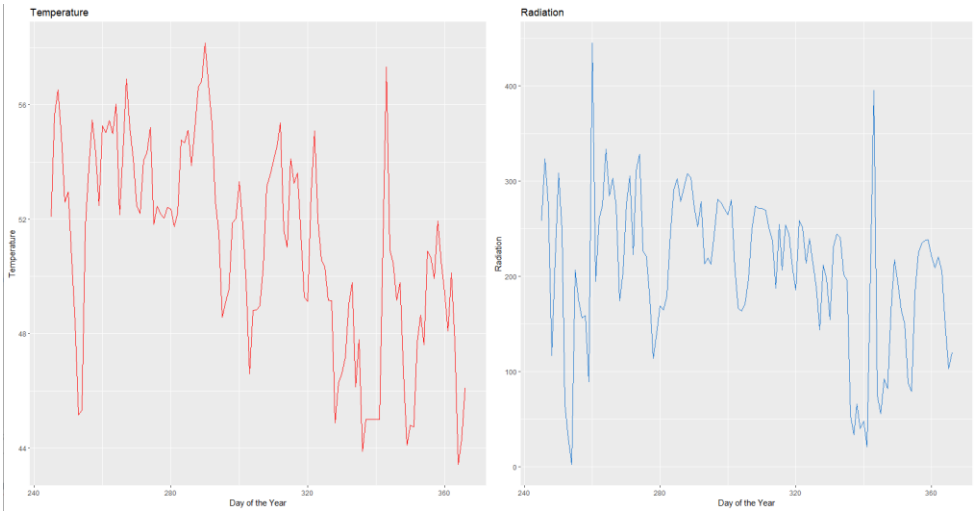



*Figure 1. Average daily temperature and radiaton.*

The final dataset for this study comprises the mean values of all measurements per day and is split so that 75% of observations are kept as a *training set* (92) and 25% are kept for *testing* (28) . Below we can see all the means for the different measurements within our training test.

| **Mean Values (September – December)** | | | | | | |
|---|---|---|---|---|---|---|
| Radiation | Daylight | Temperature | Pressure | Humidity | Wind Direction | Speed |
| 531.23 Watts/$m^2$ | 11.54 hours | 56.24 Farenheit | 30.43Hg | 73.50% | 120.56 degrees | 6.29 mph |

*Table 3. Average values of all measurements from September to December of 2016.*

Before we proceed with modeling, we plot the densities and correlations for all variables to better understand their behavior and relationship, if any. From figure 2 we can observe that most variables are skewed and have 2 peaks.

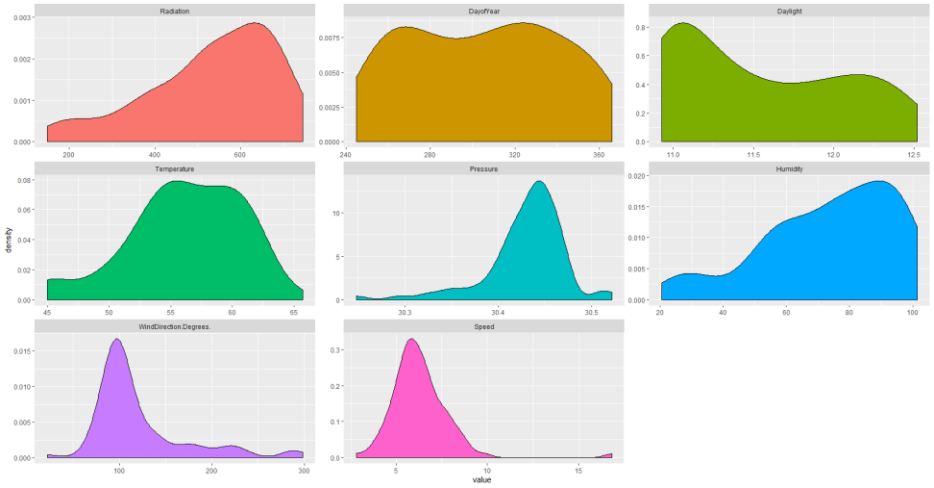



*Figure 2. Densities of all variables*

Below, the variable with the highest correlation to radiation is temperature at exactly 0.70 and there seems to be somewhat of a positive linear relationship. We can also see that humidity has a correlation of 0.64, but this may me do to its correlation to temperature. As expected, the hours of daylight are highly correlated to the day of the year so we may consider including one or the other in our models.

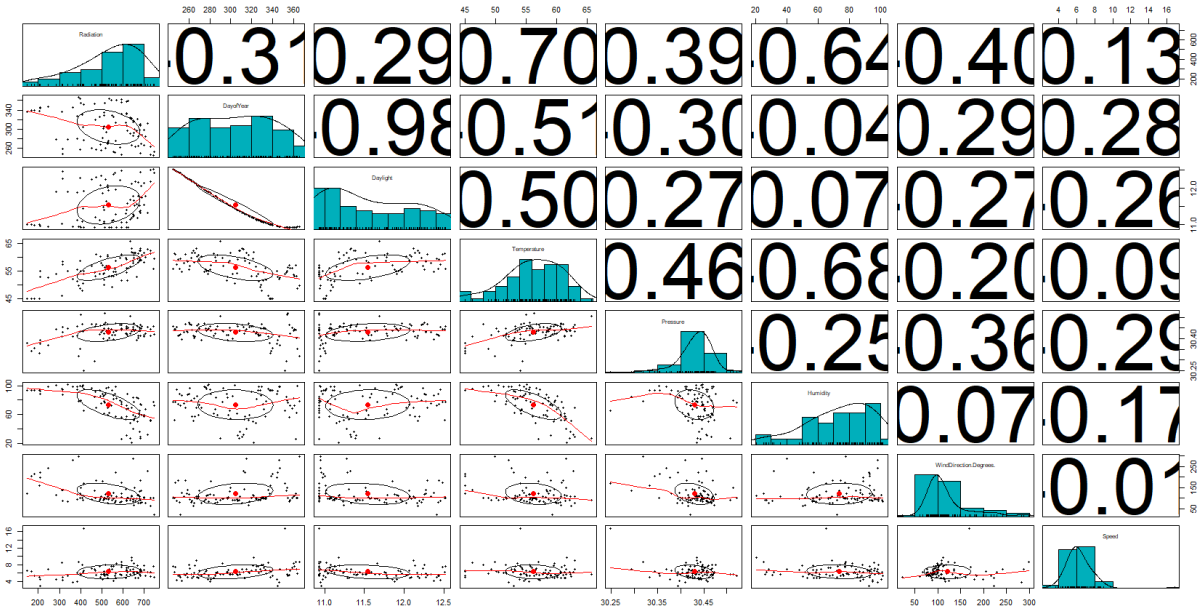
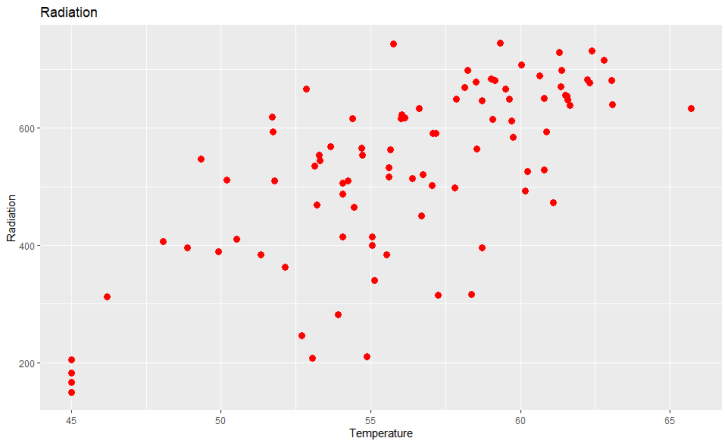



*Figure 3. Correlations, densities, and plots against other variables.*

# Regression Models

## Simple and Multiple Regression Models

Exploring the relationship between radiation and temperature, we can see in figure 4 below that there seems to be in fact a linear relation; however, there is significant variability. We fit a simple linear regression model with only one predictor. The model estimates the mean as a negative value of -741.74 where temperature has a positive relationship, such that for a one unit increase in temperature the radiation is to increase by 22.636 watts per square meter.




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

Residuals:
    Min      1Q  Median      3Q     Max
-290.45  -67.49   11.99   80.13  222.27

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -741.745    137.695  -5.387 5.69e-07 ***
Temperature   22.636      2.441   9.275 9.20e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 106.7 on 90 degrees of freedom
Multiple R-squared:  0.4887,    Adjusted R-squared:  0.483
F-statistic: 86.03 on 1 and 90 DF,  p-value: 9.203e-15
```

*Figure 4. Radiation vs. temperature plot. Simple linear regression model.*

While the diagnostics plots below do not show any significant trends or anomalies, the model's $R^2$ is roughly 48% and it does not account well for some of the variability.

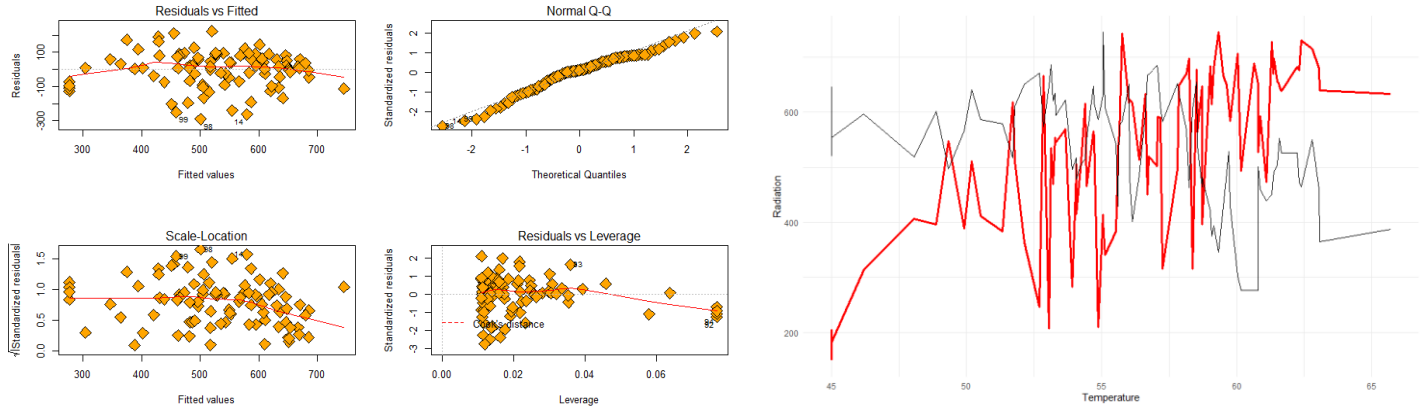



*Figure 5. Simple regression model diagnostics.*

Since we know radiation and temperature are correlated we fit a polynomial model. Both models below show a slight improvement in $R^2$, but interpretability becomes diffciult.

```
Call:
lm(formula = Radiation ~ poly(Temperature, 2), data = training)

Residuals:
    Min      1Q  Median      3Q     Max
-302.54  -54.21   16.43   73.45  210.53

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            531.23      11.08  47.936  < 2e-16 ***
poly(Temperature, 2)1  989.69     106.29   9.311 8.53e-15 ***
poly(Temperature, 2)2 -138.33     106.29  -1.301    0.196
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 106.3 on 89 degrees of freedom
Multiple R-squared:  0.4983,    Adjusted R-squared:  0.487
F-statistic: 44.19 on 2 and 89 DF,  p-value: 4.688e-14
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.81229 -0.10635  0.04915  0.16570  0.46305

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -5.7669     1.2865  -4.483 2.16e-05 ***
log(Temperature)  2.9779     0.3195   9.322 7.37e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2562 on 90 degrees of freedom
Multiple R-squared:  0.4912,    Adjusted R-squared:  0.4856
F-statistic: 86.89 on 1 and 90 DF,  p-value: 7.369e-15
```

*Figure 6. Polynomial simple model and model with log-transformations.*

To fint the best multiple regression model we use the *StepAIC* function in a backward process and the final model with the lowest AIC includes speed, humidty, temperature and wind direction. However, when we fit this model we see that speed is not significant at a 0.05 level. Refitting the model without speed yields a slightly lower $R^2$ at roughly 61%, so we keep the original AIC model (modAIC). In this model temperature and wind speed have a positive relationship with radiation as 1 unit increase in either leads to a 14.512 and/or 10.164 increase in radiation respectively. In contrasta, humidity and wind direction have a negative impact.

```
Step:  AIC=835.31
Radiation ~ Temperature + Humidity + WindDirection.Degrees. +
    Speed

                      Df Sum of Sq    RSS    AIC
<none>                             723978 835.31
- Speed                1     23831 747809 836.29
- Humidity             1     87115 811094 843.76
- WindDirection.Degrees.  1    161480 885459 851.83
- Temperature          1    192073 916051 854.96

Call:
lm(formula = Radiation ~ Temperature + Humidity + WindDirection.Degrees. +
    Speed, data = training)

Coefficients:
          (Intercept)            Temperature               Humidity
              -88.9144                14.5122                -2.1401
WindDirection.Degrees.                  Speed
               -0.8515                10.1636
```

```
Call:
lm(formula = Radiation ~ Temperature + Humidity + WindDirection.Degrees. +
    Speed, data = training)

Residuals:
    Min      1Q  Median      3Q     Max
-191.74  -68.85   13.29   60.82  197.07

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -88.9144   227.5230  -0.391  0.69691
Temperature            14.5122     3.0207   4.804 6.43e-06 ***
Humidity               -2.1401     0.6614  -3.236  0.00172 **
WindDirection.Degrees.  -0.8515     0.1933  -4.405 3.00e-05 ***
Speed                  10.1636     6.0060   1.692  0.09418 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 91.22 on 87 degrees of freedom
Multiple R-squared:  0.6388,    Adjusted R-squared:  0.6222
F-statistic: 38.46 on 4 and 87 DF,  p-value: < 2.2e-16
```

*Figure 7. Backward StepAIC final model selection and model summary (modAIC).*

Testing our model, we make some predictions and test against our testing set wich yields and $R^2$ of roughly 50%.

## Advanced Regression Methods

To calibrate, estimate and evaluate different models, we set a five-fold cross-validation of 4 repeats, testing first our *modAIC* which includes temperature, humidity, speed and wind direction. The results for the testing set show a lower $R^2$ at roughly 54%, while the training was significantly higher at 61%.

```
Pre-processing: scaled (4), centered (4)
Resampling: Cross-Validated (5 fold, repeated 4 times)
Summary of sample sizes: 74, 74, 73, 73, 74, 73, ...
Resampling results:

  RMSE      Rsquared   MAE
  96.60316  0.6093769  76.60581

Tuning parameter 'intercept' was held constant at a value of TRUE
> postResample(pred = test_results$lm,  obs = test_results$Radiation)
      RMSE   Rsquared        MAE
106.9381326  0.5387079  84.4338694
```

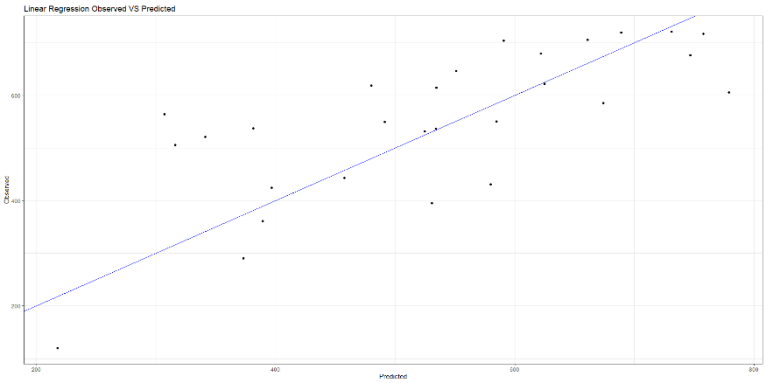
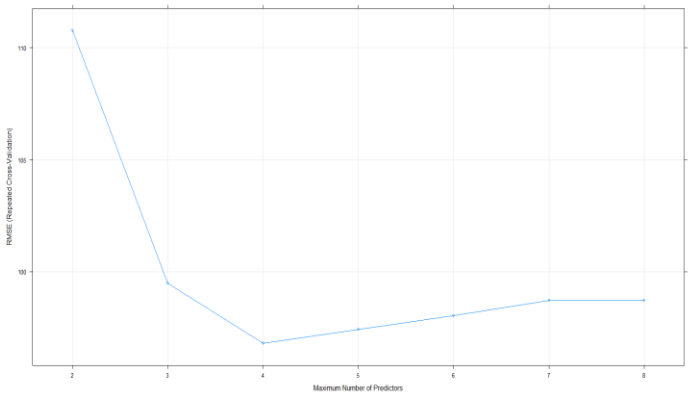



*Figure 8. Results from 5-fold CV of 4 repeats.*

We now try a stepwise regression with CV. The initial model includes temperature, humidity, speed, wind direction, daylight and day of the year and an interaction between. This process also selects the same four variables our *stepAIC* and we can see that the optimum number of variables to be included in the model, so that RMSE is minimized is four. The predictive results are the same as with the previous method.




```
> coef(step_tune$finalModel, step_tune$bestTune$nvmax)
          (Intercept)           Temperature              Humidity
            531.22753             66.51321            -44.47934
WindDirection.Degrees.                 Speed
            -43.21381             17.13525
> postResample(pred = test_results$seq,  obs = test_results$Radiation)
      RMSE   Rsquared        MAE
106.9381326  0.5387079  84.4338694
```

*Figure 9. Stepwise regression results with CV.*

## Ridge Regression

Ridge regression  is used in high dimension to mitigate overfitting. It adds some bias to the estimation to reduce variance, so it yields a better MSE than OLS. Our implementation of ridge regression centers the data matrix of predictors. We implement ridge regression with a grid of size 50 hyperparameters. Using the second lambda to conservatively predict the future (0.068), the results displayed below show that the $R^2$ has decresed in comparison with previous models.

```
     RMSE  Rsquared       MAE
0.2715211 0.4701161 0.1982662
```

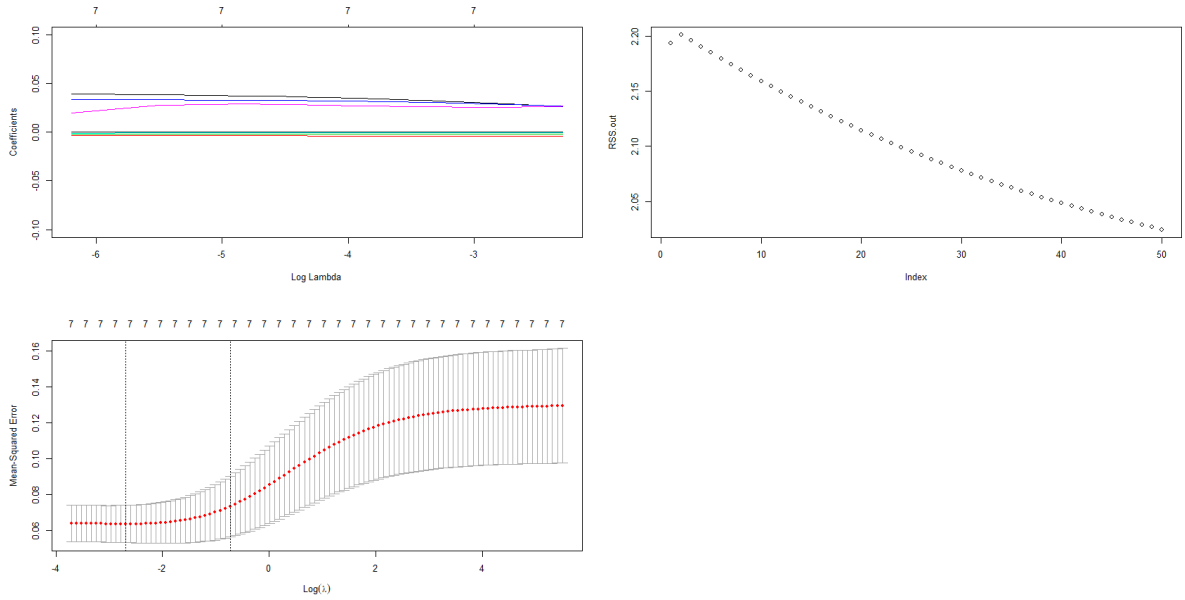
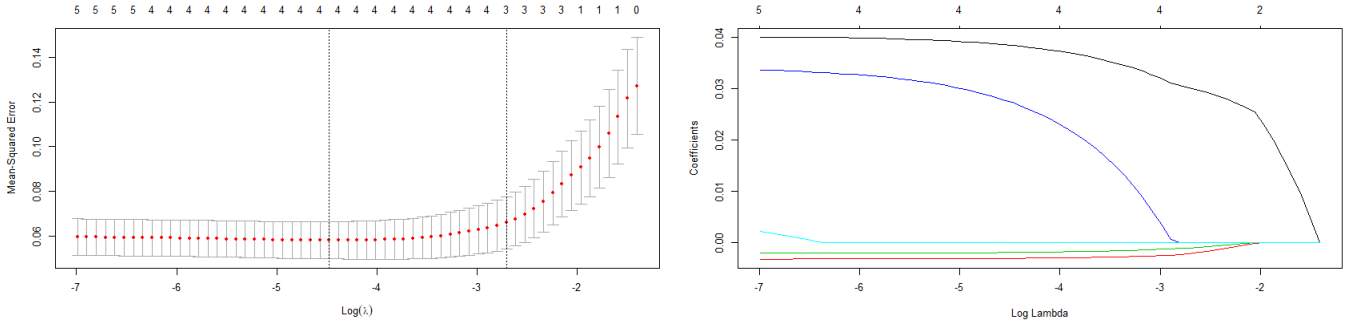



*Figure 10. Ridge regression coefficients as log rho increases, RSS over the grid index, optimal lambda to minimize MSE.*

### Lasso

Lasso regression is used in high dimensions to mitigate overfitting. In our function, we replace the alpha value to a 1 to achieve $L_1$ regularization. This method also adds some bias to the estimation to reduce the variance. We start with 7 variables and as lambda is increased, the variables go to zero in each step as displayed on the right figure below. This method yields and optimal lambda of 0.011 and an $R^2$ of 54%; however, the RMSE and MAE are the highest reported amongst all models.




| RMSE | Rsquared | MAE |
|---|---|---|
| 560.7020283 | 0.5480349 | 542.9558376 |

## Conclusions

While we have focused on the predictive capabilities of different regression methods for solar radiation, the true application of this study concerns the feasibility of implementing a PV grid in the island given its meteorological tendencies during fall and the beginning of winter. It is known that solar panels (PV) do not work are night or under severe weather conditions due to the lack of solar radiation exposure, or simply when it is significantly low. In these events the PV uses their backup battery; therefore, we only considered observations where the recorded radiance reports a value of 100

Watts/$\mathrm{m}^2$ – the minimum radiance required to power very few appliances. This translates into a time between 9a.m. and 6p.m. so that the mean solar radiation per day was focused on the amount of daylight 2 hours after sunrise.

Given the predictive focus of our project, we are not concerned about overfitting. However, when comparing all the methods used, the final method to be selected would be between stepwise CV and ridge regression. Stepwise CV offered a $\mathrm{R}^2$ 53.87% on our testing set, while ridge offered a lower 47% given its conservative approach. However, ridge offered a lower RMSE, so it would come down to how conservative we wish to be in predicting the mean solar radiation for any given day.

In the second part of this project, the focus will shift to predicting solar radiation given hourly average data instead of daily values, where we implement machine learning tools. Hopefully, implementing different methods and advanced tools will allow us to make better predictions and finally asses solar energy applications.

# Bibliography

Solar Radiation Prediction Task from NASA Hackathon - https://www.kaggle.com/dronio/SolarEnergy

Solar Radiation Basics, U.S. Office of Energy Efficiency & Renewable Energy - https://www.energy.gov/eere/solar/articles/solar-radiation-basics