Statistics\_Assignment\_Multiple.Regression\_C00246393

Pranav\_Deogaonkar

09/05/2020

Table of Contents

[1. Introduction 1](#_Toc40143167)

[1.1 Research Question: 1](#_Toc40143168)

[1.2 Research Rationale: 1](#_Toc40143169)

[2. Presentation of Dataset: 1](#_Toc40143170)

[3. Data Analysis: 2](#_Toc40143171)

[3.1 Data preparation: 2](#_Toc40143172)

[3.2 Model Creation: 7](#_Toc40143173)

[3.3 Model Diagnostics: 12](#_Toc40143174)

[3.4 Multiple Linear regression assumptions check: 18](#_Toc40143175)

[3.5 Model Validation on test data 22](#_Toc40143176)

[4. Conclusion: 22](#_Toc40143177)

# 1. Introduction

## 1.1 Research Question:

To predict the solar radiation level for four months (september-december 2016) at Hawaii islands. Do factors such as Temperature, Humidity, Wind Speed, DayLight duration and Wind Pressure can predict the Solar radiation level?

## 1.2 Research Rationale:

This research aims at predicting solar radiation, thus interest into meteorological data and facts and phenomena resulting from it made me to research on this topic.

# 2. Presentation of Dataset:

This dataset contains meteorological data from the Hawaii Space Exploration Analog and Simulation(HI-SEAS) weather station for the time period from september-december 2016.This dataset contains below columns in csv format file: The units of each dataset are:

Solar radiation: watts per meter^2 Temperature: degrees Fahrenheit Humidity: percent Barometric pressure: Hg Wind direction: degrees Wind speed: miles per hour Sunrise/sunset: Hawaii time

This dataset is from Kaggle and is part of NASA hackathon.

# 3. Data Analysis:

# Importing packages dplyr and ggplot2  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

## 3.1 Data preparation:

# Importing lubridate package  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

# Reading "csv" file and displaying top few records.  
Solar\_Dataset <- read.csv("D://Submissions//Statistics//SolarEnergy//SolarPrediction.csv")  
head(Solar\_Dataset)

## UNIXTime Data Time Radiation Temperature Pressure  
## 1 1475229326 9/29/2016 12:00:00 AM 23:55:26 1.21 48 30.46  
## 2 1475229023 9/29/2016 12:00:00 AM 23:50:23 1.21 48 30.46  
## 3 1475228726 9/29/2016 12:00:00 AM 23:45:26 1.23 48 30.46  
## 4 1475228421 9/29/2016 12:00:00 AM 23:40:21 1.21 48 30.46  
## 5 1475228124 9/29/2016 12:00:00 AM 23:35:24 1.17 48 30.46  
## 6 1475227824 9/29/2016 12:00:00 AM 23:30:24 1.21 48 30.46  
## Humidity WindDirection.Degrees. Speed TimeSunRise TimeSunSet  
## 1 59 177.39 5.62 06:13:00 18:13:00  
## 2 58 176.78 3.37 06:13:00 18:13:00  
## 3 57 158.75 3.37 06:13:00 18:13:00  
## 4 60 137.71 3.37 06:13:00 18:13:00  
## 5 62 104.95 5.62 06:13:00 18:13:00  
## 6 64 120.20 5.62 06:13:00 18:13:00

# Transforming sunrise and sunset column into time period datatype by using hms() from lubridate. hms() transforms into hour,minutes and seconds.  
  
Solar\_Dataset$TimeSunRise <- hms(Solar\_Dataset$TimeSunRise)  
Solar\_Dataset$TimeSunSet <- hms(Solar\_Dataset$TimeSunSet)  
  
# Taking difference between sunrise and sunset time and converting it into minutes.Created new column DayLight\_Minutes which gives how long sun radiation was available throughout the day in minutes.   
  
Solar\_Dataset$DayLight\_Minutes <- as.double(Solar\_Dataset$TimeSunSet-Solar\_Dataset$TimeSunRise,units="mins")

## Note: method with signature 'Period#ANY' chosen for function '-',  
## target signature 'Period#Period'.  
## "ANY#Period" would also be valid

# Dropping sunrise and sunset time columns as we created new column "DayLight\_Minutes" as above  
Solar\_Dataset <- Solar\_Dataset[,-c(10,11)]  
  
# Overview of basic statistics such as measures of centre (mean)and measures of spread (quantiles,sd) and distribution of data (histogram) for all variables in dataset.  
  
skimr::skim(Solar\_Dataset)

Data summary

|  |  |
| --- | --- |
| Name | Solar\_Dataset |
| Number of rows | 32686 |
| Number of columns | 10 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 2 |
| numeric | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| Data | 0 | 1 | FALSE | 118 | 10/: 288, 10/: 288, 10/: 288, 10/: 288 |
| Time | 0 | 1 | FALSE | 8299 | 16:: 24, 15:: 21, 18:: 21, 19:: 21 |

**Variable type: numeric**

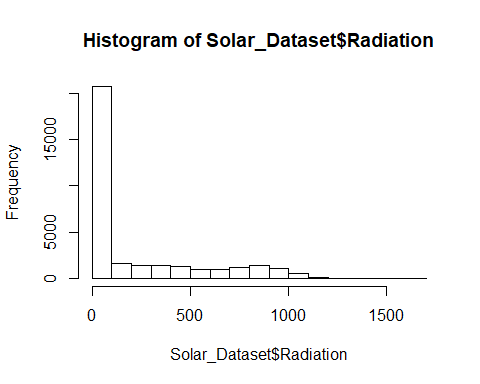
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| UNIXTime | 0 | 1 | 1.478047e+09 | 3005036.64 | 1.472724e+09 | 1.475546e+09 | 1.478026e+09 | 1.48048e+09 | 1.483265e+09 | ▇▇▇▇▇ |
| Radiation | 0 | 1 | 2.071200e+02 | 315.92 | 1.110000e+00 | 1.230000e+00 | 2.660000e+00 | 3.54240e+02 | 1.601260e+03 | ▇▁▁▁▁ |
| Temperature | 0 | 1 | 5.110000e+01 | 6.20 | 3.400000e+01 | 4.600000e+01 | 5.000000e+01 | 5.50000e+01 | 7.100000e+01 | ▁▇▇▃▁ |
| Pressure | 0 | 1 | 3.042000e+01 | 0.05 | 3.019000e+01 | 3.040000e+01 | 3.043000e+01 | 3.04600e+01 | 3.056000e+01 | ▁▁▃▇▁ |
| Humidity | 0 | 1 | 7.502000e+01 | 25.99 | 8.000000e+00 | 5.600000e+01 | 8.500000e+01 | 9.70000e+01 | 1.030000e+02 | ▁▂▂▃▇ |
| WindDirection.Degrees. | 0 | 1 | 1.434900e+02 | 83.17 | 9.000000e-02 | 8.223000e+01 | 1.477000e+02 | 1.79310e+02 | 3.599500e+02 | ▅▆▇▁▂ |
| Speed | 0 | 1 | 6.240000e+00 | 3.49 | 0.000000e+00 | 3.370000e+00 | 5.620000e+00 | 7.87000e+00 | 4.050000e+01 | ▇▂▁▁▁ |
| DayLight\_Minutes | 0 | 1 | 6.905500e+02 | 29.93 | 6.560000e+02 | 6.620000e+02 | 6.840000e+02 | 7.15000e+02 | 7.510000e+02 | ▇▃▃▃▂ |

### Performing shapiro wilk test for normality of data on the dataset by creating samples of size 4500 as shapiro wilk can be applied to datasets with max 5000 observations. In this the null hypothesis states that data are normally distributed.   
  
start= 1  
increment=4500  
result=""  
for(i in start:increment){  
print(start)  
print(increment)  
  
Solar\_Subset <- Solar\_Dataset[start:increment,]  
Solar\_Subset  
total\_records = nrow(Solar\_Dataset)  
  
if (increment <= total\_records) {  
result=shapiro.test(Solar\_Subset$Radiation)  
result\_epoch = result   
start= increment  
increment = increment +4500  
print(result\_epoch)  
}  
else {  
 break   
}  
  
}

## [1] 1  
## [1] 4500  
##   
## Shapiro-Wilk normality test  
##   
## data: Solar\_Subset$Radiation  
## W = 0.72539, p-value < 2.2e-16  
##   
## [1] 4500  
## [1] 9000  
##   
## Shapiro-Wilk normality test  
##   
## data: Solar\_Subset$Radiation  
## W = 0.67882, p-value < 2.2e-16  
##   
## [1] 9000  
## [1] 13500  
##   
## Shapiro-Wilk normality test  
##   
## data: Solar\_Subset$Radiation  
## W = 0.72153, p-value < 2.2e-16  
##   
## [1] 13500  
## [1] 18000  
##   
## Shapiro-Wilk normality test  
##   
## data: Solar\_Subset$Radiation  
## W = 0.71711, p-value < 2.2e-16  
##   
## [1] 18000  
## [1] 22500  
##   
## Shapiro-Wilk normality test  
##   
## data: Solar\_Subset$Radiation  
## W = 0.71282, p-value < 2.2e-16  
##   
## [1] 22500  
## [1] 27000  
##   
## Shapiro-Wilk normality test  
##   
## data: Solar\_Subset$Radiation  
## W = 0.71333, p-value < 2.2e-16  
##   
## [1] 27000  
## [1] 31500  
##   
## Shapiro-Wilk normality test  
##   
## data: Solar\_Subset$Radiation  
## W = 0.62714, p-value < 2.2e-16  
##   
## [1] 31500  
## [1] 36000

### From shapiro wilk test it is observed that p value is less than alpha(0.05), hence null hypotheis is rejected and we can infer that data is not normally distributed.

# Also, Checking normality through histogram plot.  
hist(Solar\_Dataset$Radiation)

 ### We can observe from above plot that data is skewed for Radiation variable.

# To check skewness in data variables.  
library(e1071)

## Warning: package 'e1071' was built under R version 3.6.2

skewness(Solar\_Dataset$Radiation)

## [1] 1.369556

skewness(Solar\_Dataset$Pressure)

## [1] -1.226849

skewness(Solar\_Dataset$Humidity)

## [1] -0.7762092

skewness(Solar\_Dataset$Speed)

## [1] 1.46968

### Positive value infers to right skewness and negative to left skewness of data.

## 3.2 Model Creation:

# Splitting the dataset into training and test data in the ratio 70:30 respectively.  
set.seed(100)  
index\_P <- sample(1:nrow(Solar\_Dataset),0.70\*nrow(Solar\_Dataset),replace = FALSE)  
training\_data <- Solar\_Dataset[index\_P,]  
testing\_data <- Solar\_Dataset[-index\_P,]

# Creating a simple linear regression model with Radiation as dependent variable and temperature as independent variable. As Radiation is skewed we perform log(natural logarithm) transformation on it.  
  
solar\_pred <- lm(log1p(Radiation)~Temperature,data = training\_data)  
solar\_pred

##   
## Call:  
## lm(formula = log1p(Radiation) ~ Temperature, data = training\_data)  
##   
## Coefficients:  
## (Intercept) Temperature   
## -11.8808 0.2926

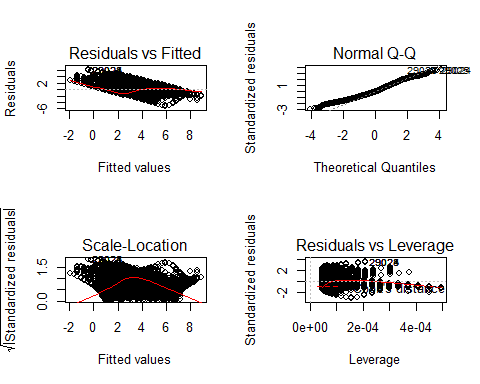
summary(solar\_pred)

##   
## Call:  
## lm(formula = log1p(Radiation) ~ Temperature, data = training\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1620 -1.3492 -0.2056 1.2275 6.1712   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.880769 0.094330 -125.9 <2e-16 \*\*\*  
## Temperature 0.292610 0.001834 159.5 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.72 on 22878 degrees of freedom  
## Multiple R-squared: 0.5267, Adjusted R-squared: 0.5266   
## F-statistic: 2.546e+04 on 1 and 22878 DF, p-value: < 2.2e-16

### Inference from summary statistics above for model1 :

1. R-squared value in the above model is 0.5266 which tells that 52.66% variance in dependent variable is explained by the independent variable.This is not very good enough to accept the model.  
2. Beta estimate tells that for every unit change in temperature the solar radiation increases by 0.29. So temperature is statistically significant predictor as beta coefficient is positive and p-value(2.2 \* 10^-16) is less than alpha(0.05)   
3. Residual std. error (std deviation of the residuals) of 1.72 on 22878 DF tells us that lower the value better is the model. Value of 1.72 is lower enough for the model to be fit.

# plotting graphs for above model   
# residuals vs fitted plot tells about homoscedasticity  
# QQ plots determine normality in the model  
  
par(mfrow=c(2,2))  
plot(solar\_pred)



### Multiple linear regression with radiation as dependent variable and temperature, wind speed, humidity, pressure and daylight\_min as independent variables.Performed log transformation on radiation and speed as they are skewed.

solar\_pred\_2 <- lm(log1p(Radiation)~Temperature+log1p(Speed)+Humidity+Pressure+DayLight\_Minutes,data = training\_data)  
solar\_pred\_2

##   
## Call:  
## lm(formula = log1p(Radiation) ~ Temperature + log1p(Speed) +   
## Humidity + Pressure + DayLight\_Minutes, data = training\_data)  
##   
## Coefficients:  
## (Intercept) Temperature log1p(Speed) Humidity   
## 172.58893 0.35847 0.31031 0.01350   
## Pressure DayLight\_Minutes   
## -5.79817 -0.01887

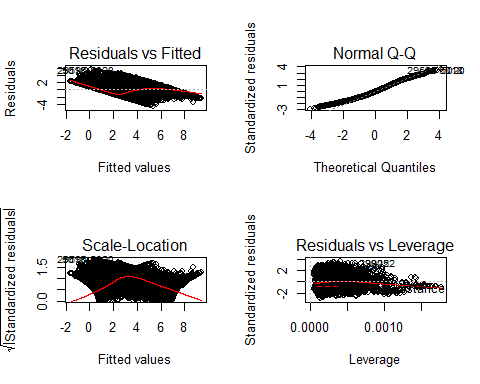
summary(solar\_pred\_2)

##   
## Call:  
## lm(formula = log1p(Radiation) ~ Temperature + log1p(Speed) +   
## Humidity + Pressure + DayLight\_Minutes, data = training\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.4950 -1.1782 -0.1448 1.0705 5.4860   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.726e+02 6.182e+00 27.92 <2e-16 \*\*\*  
## Temperature 3.585e-01 1.924e-03 186.26 <2e-16 \*\*\*  
## log1p(Speed) 3.103e-01 2.042e-02 15.20 <2e-16 \*\*\*  
## Humidity 1.350e-02 4.426e-04 30.50 <2e-16 \*\*\*  
## Pressure -5.798e+00 2.054e-01 -28.23 <2e-16 \*\*\*  
## DayLight\_Minutes -1.887e-02 3.886e-04 -48.56 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.562 on 22874 degrees of freedom  
## Multiple R-squared: 0.6102, Adjusted R-squared: 0.6101   
## F-statistic: 7160 on 5 and 22874 DF, p-value: < 2.2e-16

### Inference from summary statistics above for model2 :

1. R-squared value in the above model is 0.6101 which tells that 61.01% variance in radiation-dependent variable is explained by the independent variables.This is good enough than earlier model1. In short, how significant independent variables are to predict outcome variable.  
  
2. Beta estimate tells that for every unit change in temperature the solar radiation increases by 0.35, unit change in Log(speed) the radiation increases by 0.31, for unit change in humidity the solar radiation level increases by 0.0135 whereas for unit change in DayLight duration solar radiation decreases by 0.018 and for unit change in pressure solar radiation decreases by 5.79 . So temperature,Speed,humidity are statistically significant predictors as beta coefficient is positive and p-value(2.2 \* 10^-16) is less than alpha(0.05). Also, p-value for predictors Daylight\_Minutes and Pressure are less than 0.05 although have negative relationship with dependent variable can be considered as statistically significant independent variables.  
  
3. Residual std. error (std deviation of the residuals) of 1.52 on 22878 DF is lower than model1 hence it is better.

# plotting graphs for above model   
# residuals vs fitted plot tells about homoscedasticity  
# QQ plots determine normality in the model  
  
  
par(mfrow=c(2,2))  
plot(solar\_pred\_2)



# Checking the confidence interval.  
confint(solar\_pred\_2)

## 2.5 % 97.5 %  
## (Intercept) 160.47083301 184.70702052  
## Temperature 0.35470002 0.36224446  
## log1p(Speed) 0.27029334 0.35032771  
## Humidity 0.01263197 0.01436721  
## Pressure -6.20071314 -5.39562321  
## DayLight\_Minutes -0.01963262 -0.01810909

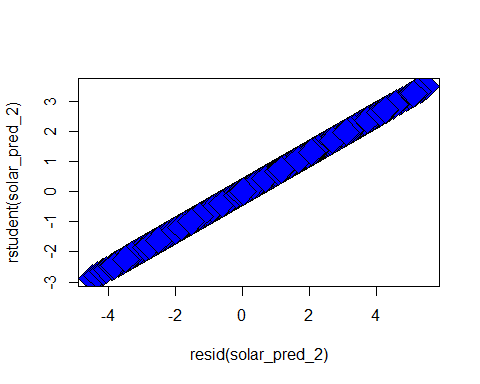
### Inference from confidence interval:

For independent variables Temperature, Humidity, DayLight\_Minutes the confidence interval is lower than 0.1. Also, for Speed and Pressure CI is lower than 1. As, CI is lower the Std error value becomes lower and better estimates of model are achieved.

## 3.3 Model Diagnostics:

# Checking outliers using residuals(std residuals, internally studentized and externally studentized). Residuals are difference between actual and predicted values.In Studentized residuals observations are deleted and then regression model is assesed.  
  
training\_data$Residuals <- resid(solar\_pred\_2)  
training\_data$Int\_StudResiduals <- rstandard(solar\_pred\_2)  
training\_data$Ext\_StudResiduals <- rstudent(solar\_pred\_2)  
#training\_data

# plot of residuals vs studentized residuals   
plot(resid(solar\_pred\_2), rstudent(solar\_pred\_2), pch=23, bg='blue', cex=3)



# to check the number of observations as outliers using studentized residuals  
  
training\_data$Higher\_Residuals <- training\_data$Ext\_StudResiduals>1.96|training\_data$Ext\_StudResiduals< -1.96  
#training\_data

# Ext\_Studentized residuals with higher value than +/-1.96  
Higher\_residuals <- training\_data %>% filter(Higher\_Residuals==TRUE)

# Checking for outlier cases with std. residuals greater than 2.5 and less than -2.5  
# as per the rule only 1% observations(ie. 326) should fall within +/- 2.5 range.  
# As seen below less than 1% (0.63%) observations lie within specified range.  
  
Residuals\_greater2.5 <- training\_data%>% filter(Ext\_StudResiduals>2.5 | Ext\_StudResiduals < -2.5)

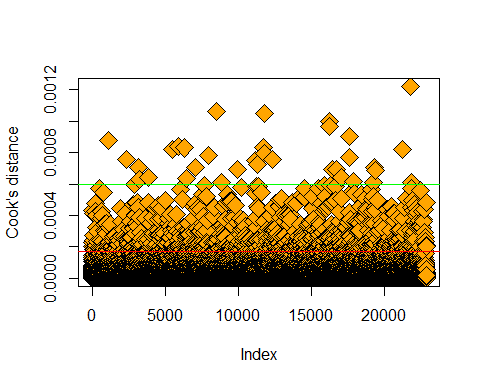
# Also, 99.7% cases while fall within 3 std residuals. Hence we check the condition below and look out for number of observations. We find less than 0.3% observations (18) as outliers   
  
Residuals\_99 <- training\_data%>% filter(Ext\_StudResiduals>3 | Ext\_StudResiduals < -3)

### Finding influential cases:

1. Cooks Distance:The cut off value as outlier is 4/N(N=sample size) so it is 4/22880= 0.00017

training\_data$CooksDistance\_value <- cooks.distance(solar\_pred\_2)

# Also, plotting cooks distance value to confirm/modify cut off value for deciding influential cases.  
  
plot(cooks.distance(solar\_pred\_2), pch=23, bg='orange', cex=2, ylab="Cook's distance")  
abline(h=0.00017,col="red")  
abline(h=0.0006,col="green")



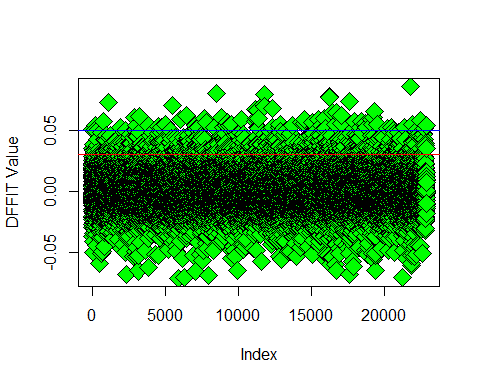
# From above plot it is observed that taking cut-off value as 0.0006 is more apt.  
#training\_data %>% filter(CooksDistance\_value> 0.0006)

1. DFFIT -

training\_data$DFFIT\_value <- dffits(solar\_pred\_2)  
#training\_data

# Critical value for DFFIT is calculated as "2\* sqrt(k+1/n)" ie 2\*sqrt(6/22880)~ 0.02  
  
#training\_data %>% filter(DFFIT\_value> 0.02)

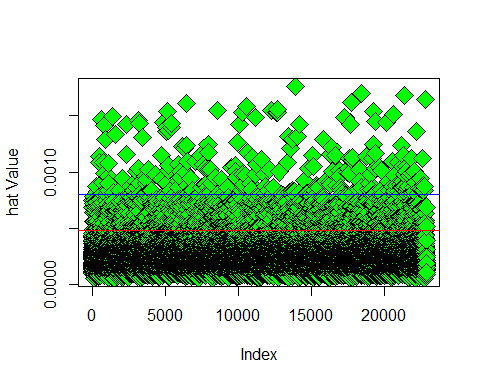
# From below plot it is depicted that cut off value for DFFIT of 0.05 separates outliers than 0.02  
  
plot(dffits(solar\_pred\_2), pch=23, bg='green', cex=2, ylab="DFFIT Value")  
abline(h=0.03,col="red")  
abline(h=0.05,col="blue")

 3. leverage: It is standardized distance to mean of predictors.

training\_data$leverage <- hatvalues(solar\_pred\_2)  
head(training\_data)

## UNIXTime Data Time Radiation Temperature  
## 20170 1479301801 11/16/2016 12:00:00 AM 03:10:01 1.25 48  
## 16887 1480287918 11/27/2016 12:00:00 AM 13:05:18 873.39 57  
## 3430 1474173303 9/17/2016 12:00:00 AM 18:35:03 1.20 55  
## 3696 1474091704 9/16/2016 12:00:00 AM 19:55:04 1.27 54  
## 30707 1481405140 12/10/2016 12:00:00 AM 11:25:40 243.21 54  
## 31705 1480880733 12/4/2016 12:00:00 AM 09:45:33 73.59 45  
## Pressure Humidity WindDirection.Degrees. Speed DayLight\_Minutes  
## 20170 30.44 23 143.64 9.00 671  
## 16887 30.44 68 0.41 9.00 664  
## 3430 30.46 83 134.74 5.62 734  
## 3696 30.47 96 117.20 5.62 735  
## 30707 30.40 88 44.96 0.00 657  
## 31705 30.40 93 77.27 4.50 660  
## Residuals Int\_StudResiduals Ext\_StudResiduals Higher\_Residuals  
## 20170 -0.851088 -0.5451143 -0.5451060 FALSE  
## 16887 1.145681 0.7337682 0.7337608 FALSE  
## 3430 -2.760015 -1.7676509 -1.7677330 FALSE  
## 3696 -2.468862 -1.5812141 -1.5812659 FALSE  
## 30707 1.026101 0.6573639 0.6573558 FALSE  
## 31705 2.526442 1.6180428 1.6181001 FALSE  
## CooksDistance\_value DFFIT\_value leverage  
## 20170 1.376406e-05 -0.009087456 0.0002778447  
## 16887 1.757509e-05 0.010268808 0.0001958150  
## 3430 7.734112e-05 -0.021542744 0.0001484924  
## 3696 7.862276e-05 -0.021720209 0.0001886411  
## 30707 5.404903e-05 0.018007946 0.0007498970  
## 31705 5.580177e-05 0.018298477 0.0001278685

# Critical value for leverage > (2k+1)/N ie (2\*5+1)/22880~0.00048  
plot(hatvalues(solar\_pred\_2), pch=23, bg='green', cex=2, ylab="hat Value")  
abline(h=0.00048,col="red")  
abline(h=0.0008,col="blue")



#training\_data %>% filter(leverage> 0.0008)

1. Covariance ratio:

training\_data$covariance.ratios<-covratio(solar\_pred\_2)  
head(training\_data)

## UNIXTime Data Time Radiation Temperature  
## 20170 1479301801 11/16/2016 12:00:00 AM 03:10:01 1.25 48  
## 16887 1480287918 11/27/2016 12:00:00 AM 13:05:18 873.39 57  
## 3430 1474173303 9/17/2016 12:00:00 AM 18:35:03 1.20 55  
## 3696 1474091704 9/16/2016 12:00:00 AM 19:55:04 1.27 54  
## 30707 1481405140 12/10/2016 12:00:00 AM 11:25:40 243.21 54  
## 31705 1480880733 12/4/2016 12:00:00 AM 09:45:33 73.59 45  
## Pressure Humidity WindDirection.Degrees. Speed DayLight\_Minutes  
## 20170 30.44 23 143.64 9.00 671  
## 16887 30.44 68 0.41 9.00 664  
## 3430 30.46 83 134.74 5.62 734  
## 3696 30.47 96 117.20 5.62 735  
## 30707 30.40 88 44.96 0.00 657  
## 31705 30.40 93 77.27 4.50 660  
## Residuals Int\_StudResiduals Ext\_StudResiduals Higher\_Residuals  
## 20170 -0.851088 -0.5451143 -0.5451060 FALSE  
## 16887 1.145681 0.7337682 0.7337608 FALSE  
## 3430 -2.760015 -1.7676509 -1.7677330 FALSE  
## 3696 -2.468862 -1.5812141 -1.5812659 FALSE  
## 30707 1.026101 0.6573639 0.6573558 FALSE  
## 31705 2.526442 1.6180428 1.6181001 FALSE  
## CooksDistance\_value DFFIT\_value leverage covariance.ratios  
## 20170 1.376406e-05 -0.009087456 0.0002778447 1.0004624  
## 16887 1.757509e-05 0.010268808 0.0001958150 1.0003170  
## 3430 7.734112e-05 -0.021542744 0.0001484924 0.9995912  
## 3696 7.862276e-05 -0.021720209 0.0001886411 0.9997951  
## 30707 5.404903e-05 0.018007946 0.0007498970 1.0008995  
## 31705 5.580177e-05 0.018298477 0.0001278685 0.9997035

# Critical value is 1+[3(k+1)/n]= 1+[(3\*6/22880)]~ 1.00078  
  
#training\_data %>% filter(covariance.ratios> 1.0008)

### Deleting influential cases as per above diagnostic approaches(Cooks distance, DFFIT, leverage and covariance ratio):

Training\_Data\_Without\_Outliers <- training\_data[-c(which(cooks.distance(solar\_pred\_2) > 0.0006 | dffits(solar\_pred\_2) > 0.03 | hatvalues(solar\_pred\_2)> 0.0008 | covratio(solar\_pred\_2) > 1.0008)),]

#Training\_Data\_Without\_Outliers

### 3.4 Creating 3rd multiple regression model after deleting influential cases

solar\_pred\_3 <- lm(log1p(Radiation)~Temperature+log1p(Speed)+Humidity+Pressure+DayLight\_Minutes,data = Training\_Data\_Without\_Outliers)  
solar\_pred\_3

##   
## Call:  
## lm(formula = log1p(Radiation) ~ Temperature + log1p(Speed) +   
## Humidity + Pressure + DayLight\_Minutes, data = Training\_Data\_Without\_Outliers)  
##   
## Coefficients:  
## (Intercept) Temperature log1p(Speed) Humidity   
## 185.59403 0.37208 0.45706 0.01351   
## Pressure DayLight\_Minutes   
## -6.26657 -0.01862

summary(solar\_pred\_3)

##   
## Call:  
## lm(formula = log1p(Radiation) ~ Temperature + log1p(Speed) +   
## Humidity + Pressure + DayLight\_Minutes, data = Training\_Data\_Without\_Outliers)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.4834 -1.0881 -0.1094 1.0262 4.5865   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.856e+02 6.769e+00 27.42 <2e-16 \*\*\*  
## Temperature 3.721e-01 1.867e-03 199.28 <2e-16 \*\*\*  
## log1p(Speed) 4.571e-01 2.172e-02 21.05 <2e-16 \*\*\*  
## Humidity 1.351e-02 4.344e-04 31.09 <2e-16 \*\*\*  
## Pressure -6.267e+00 2.242e-01 -27.95 <2e-16 \*\*\*  
## DayLight\_Minutes -1.862e-02 3.758e-04 -49.56 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.465 on 21604 degrees of freedom  
## Multiple R-squared: 0.6559, Adjusted R-squared: 0.6558   
## F-statistic: 8236 on 5 and 21604 DF, p-value: < 2.2e-16

### Inference from summary statistics above for model3 :

1. R-squared value in the above model is 0.6558 which tells that 65.58% variance in radiation-dependent variable is explained by the independent variables.This is better than earlier model2.   
  
2. Beta estimate tells that for every unit change in temperature the solar radiation increases by 0.37, unit change in Log(speed) the radiation increases by 0.45, for unit change in humidity the solar radiation level increases by 0.0135 whereas for unit change in DayLight duration solar radiation decreases by 0.018 and for unit change in pressure solar radiation decreases by 6.26 . So temperature,Speed,humidity are statistically significant predictors as beta coefficient is positive and p-value(2.2 \* 10^-16) is less than alpha(0.05). Also, p-value for predictors Daylight\_Minutes and Pressure are less than 0.05 although have negative relationship with dependent variable can be considered as statistically significant independent variables.  
  
3. Residual std. error (std deviation of the residuals) of 1.46 on 22878 DF is lower than model2 which is good fit.

## 3.4 Multiple Linear regression assumptions check:

1. Independence of residual errors - Measure of autocorrelation in residuals

library(car)

## Warning: package 'car' was built under R version 3.6.3

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

durbinWatsonTest(solar\_pred\_3)

## lag Autocorrelation D-W Statistic p-value  
## 1 0.002898974 1.994146 0.646  
## Alternative hypothesis: rho != 0

### From above durbin watson test the statistic value of 1.99 is closer to 2 which satisfies criteria of no correlation among residual errors.

1. Multicollinearity assumption: To check whether independent variables are correlated with each other using variance inflation factor(vif) values and test statistic(1/vif)

print("Vif-Values")

## [1] "Vif-Values"

print(vif(solar\_pred\_3))

## Temperature log1p(Speed) Humidity Pressure   
## 1.324484 1.081753 1.287252 1.196709   
## DayLight\_Minutes   
## 1.252868

print("Tolerance Statistics")

## [1] "Tolerance Statistics"

print(1/vif(solar\_pred\_3))

## Temperature log1p(Speed) Humidity Pressure   
## 0.7550108 0.9244259 0.7768486 0.8356249   
## DayLight\_Minutes   
## 0.7981687

print("Average Vif")

## [1] "Average Vif"

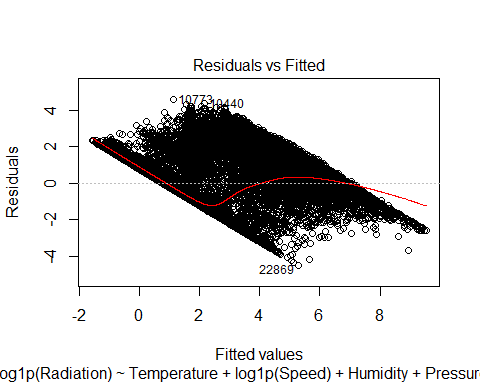
print(mean(vif(solar\_pred\_3)))

## [1] 1.228613

### From above vif values we can infer that all vif values are less than 10 and tolerance statistics is >0.2. Hence there is no multicollinearity among independent variables.

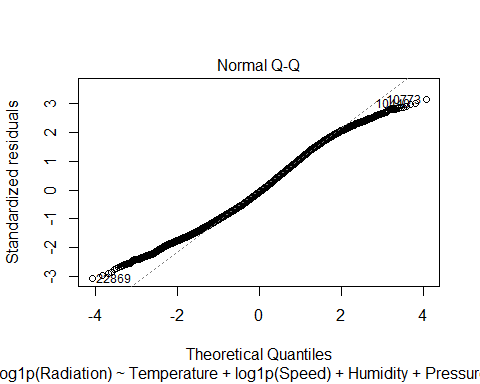
1. Assumption of Normal residuals:

# Residuals vs Fitted values plot  
plot(solar\_pred\_3,which=1)



### From above plot we can observe that there is no funnel shape,that is constant variance is seen. Heteroscedasticity(lot of variation) leads to bias in std. errors which in return leads to issues in hypothesis testing. If hypothesis testing is biased it leads to wrong inferences.But our model doesn’t suffer from heteroscedasticity, it adheres to the assumption of homoscedasticity.

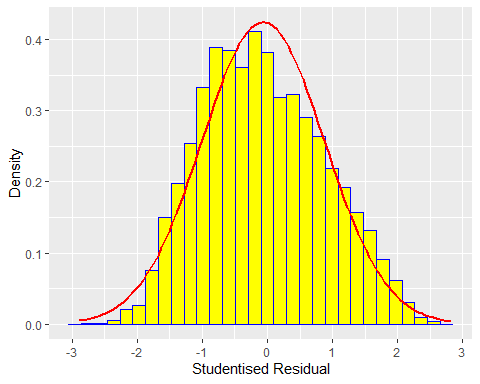
# Quantile Quantile (QQ) plot  
plot(solar\_pred\_3,which=2)



### Above QQ plot tells us that most data points lie on regression line following the normality pattern.

# Histogram of residuals:  
histogram<-ggplot(Training\_Data\_Without\_Outliers, aes(Ext\_StudResiduals))  
histogram+geom\_histogram(aes(y=..density..), colour="blue", fill="yellow") +  
 labs(x="Studentised Residual", y= "Density") +  
 stat\_function(fun=dnorm, args = list(mean= mean(Training\_Data\_Without\_Outliers$Ext\_StudResiduals, na.rm = TRUE),   
 sd= sd(Training\_Data\_Without\_Outliers$Ext\_StudResiduals, na.rm = TRUE)),   
 colour = "red", size =1)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



### Above histogram has approximately normally distributed studentized residuals.

## 3.5 Model Validation on test data

predicted\_values\_testdata <- predict(solar\_pred\_3,newdata = testing\_data)  
predicted\_values\_testdata <- expm1(predicted\_values\_testdata)  
test\_actual\_values <- testing\_data$Radiation  
testdata\_df <- data.frame(predicted\_values\_testdata,test\_actual\_values)  
#(testdata\_df)

# 4. Conclusion:

We can conclude from good fit model statistics(R-squared value of 0.65 and std residual error 1.46) and above assumptions pertaining to multiple linear regression, that factors such as Temperature, Humidity, Wind Speed, DayLight and Wind Pressure are significant in predicting the Solar Radiation level.