Predicting Solar Photovoltaic Production

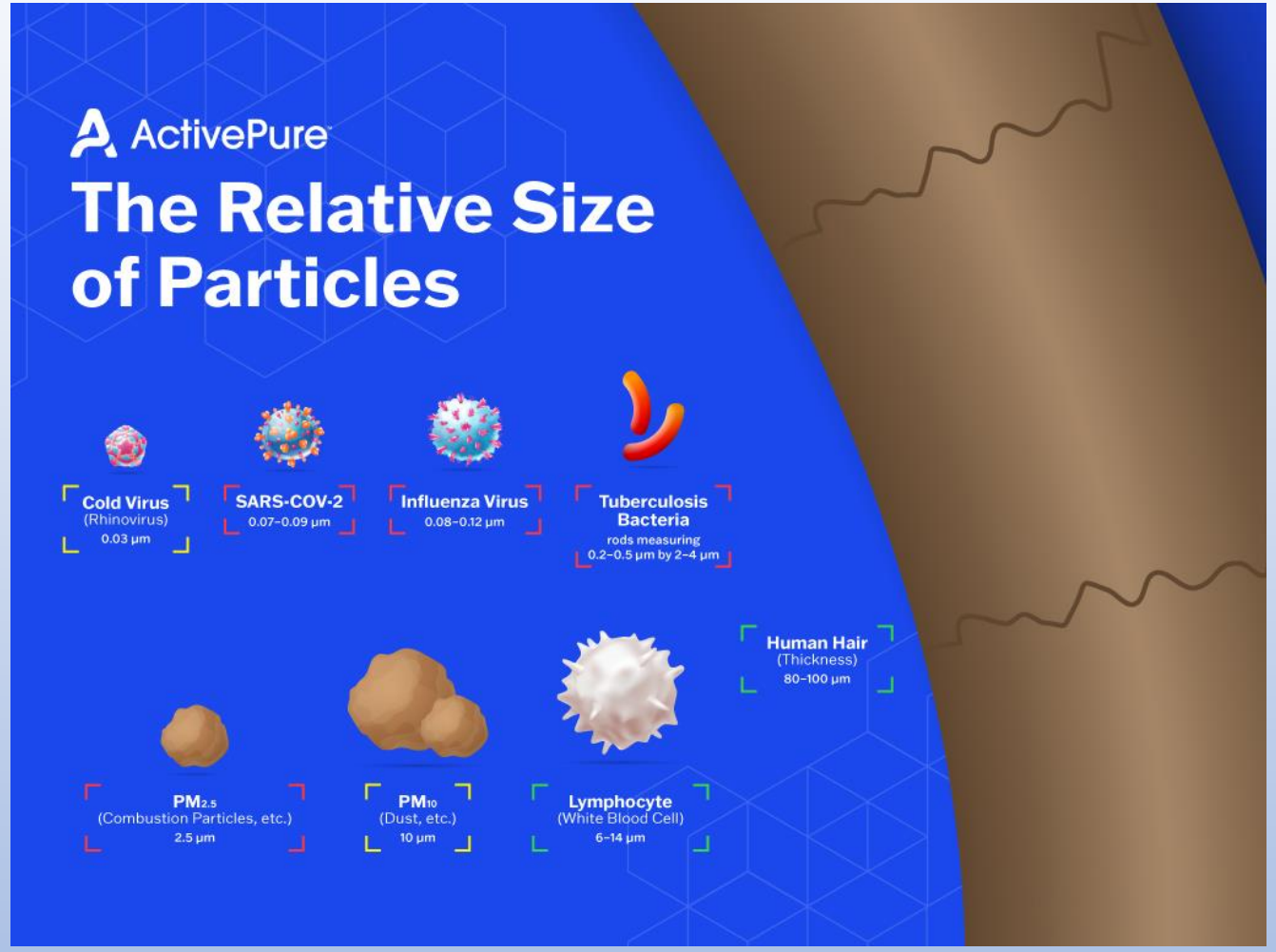
Outline

1. Research Question
2. Variables Explanation
3. Exploratory Data Analysis
4. Model 1 (Model selection)
5. Model 2 (Remove Outlying)
6. Model 3 (Log transformation)
7. Durbin Watson Test
8. Conclusion

Research Question

- How does wildfire smoke (mainly the pm2.5 in this research) affect solar photovoltaic production?

- The data is from The Department of Energy. It's collected based on a California solar power plant. Additionally, we added group data from a website specialized in collecting sunshine data.



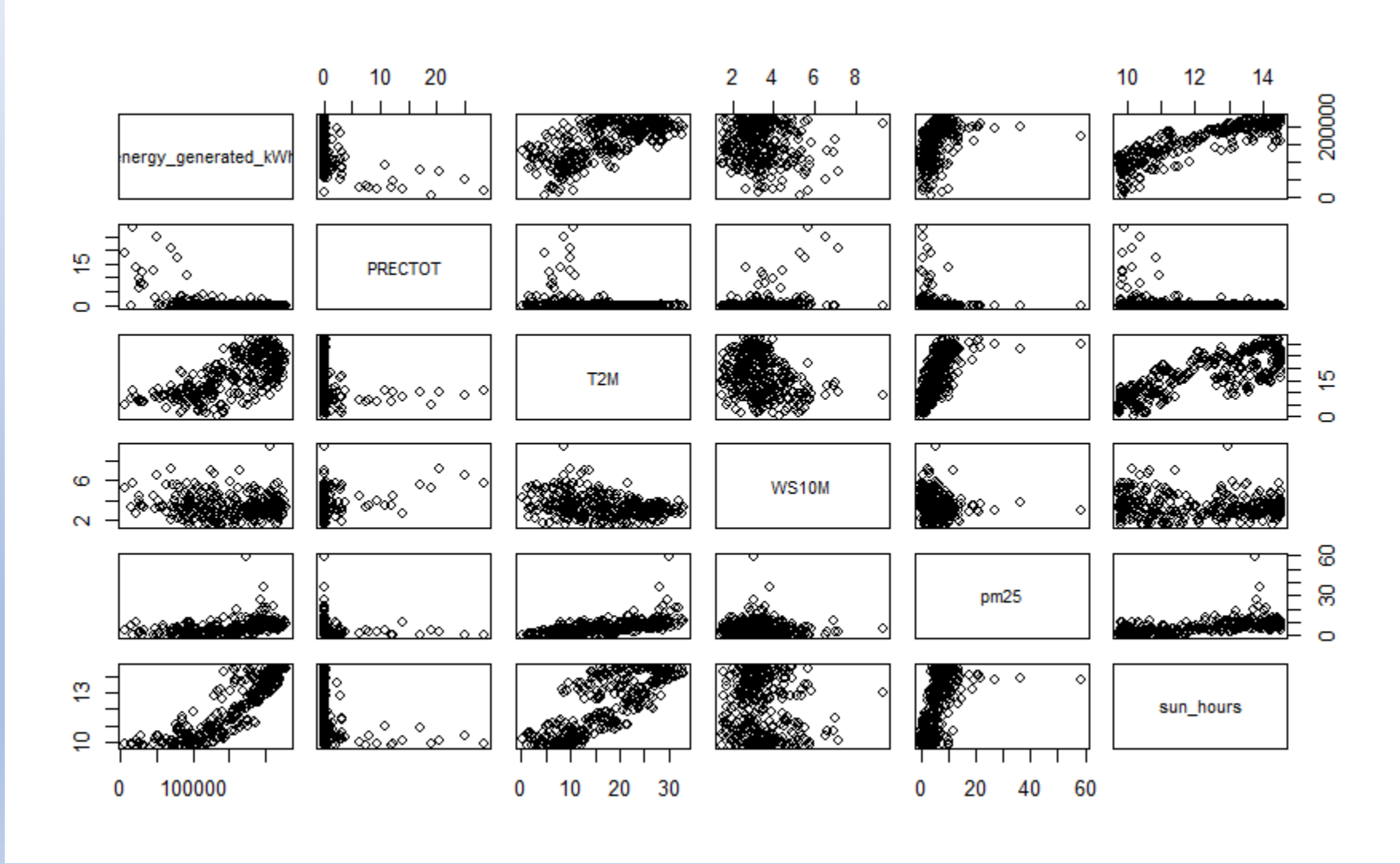
One micrometer is equal to one millionth of a meter (0.000001 meters or 10^-6 meters).

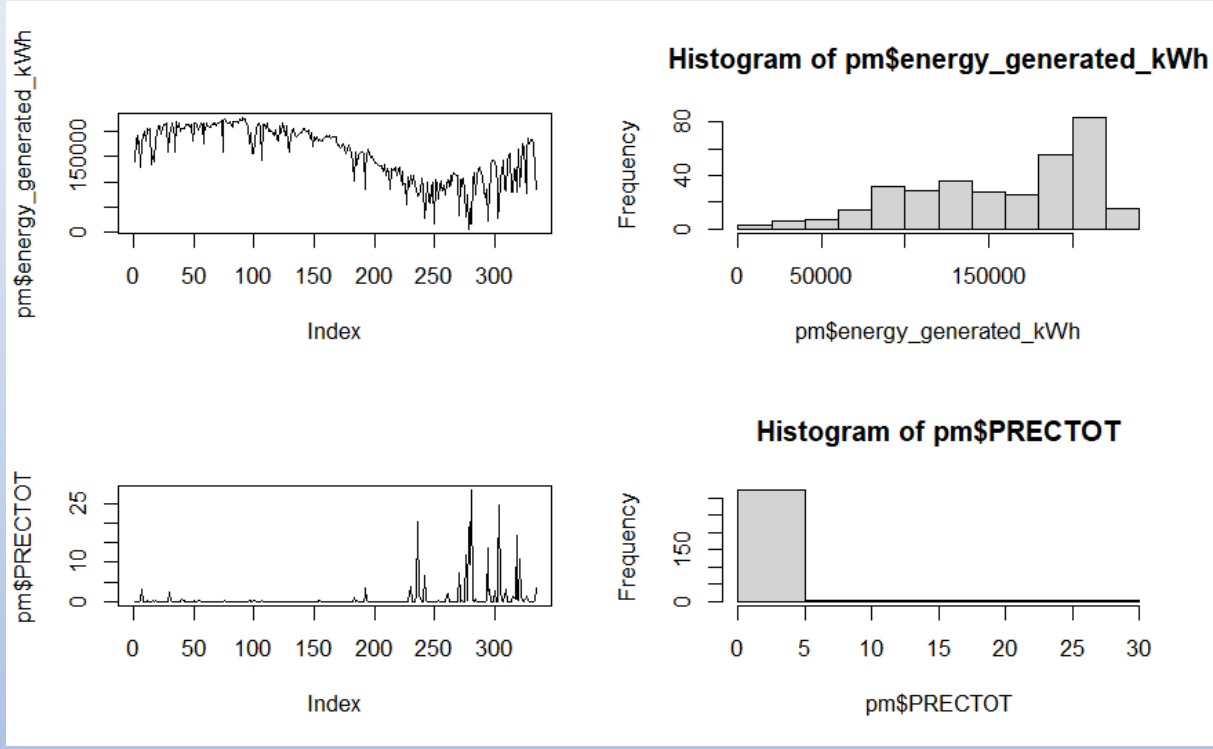
Variables

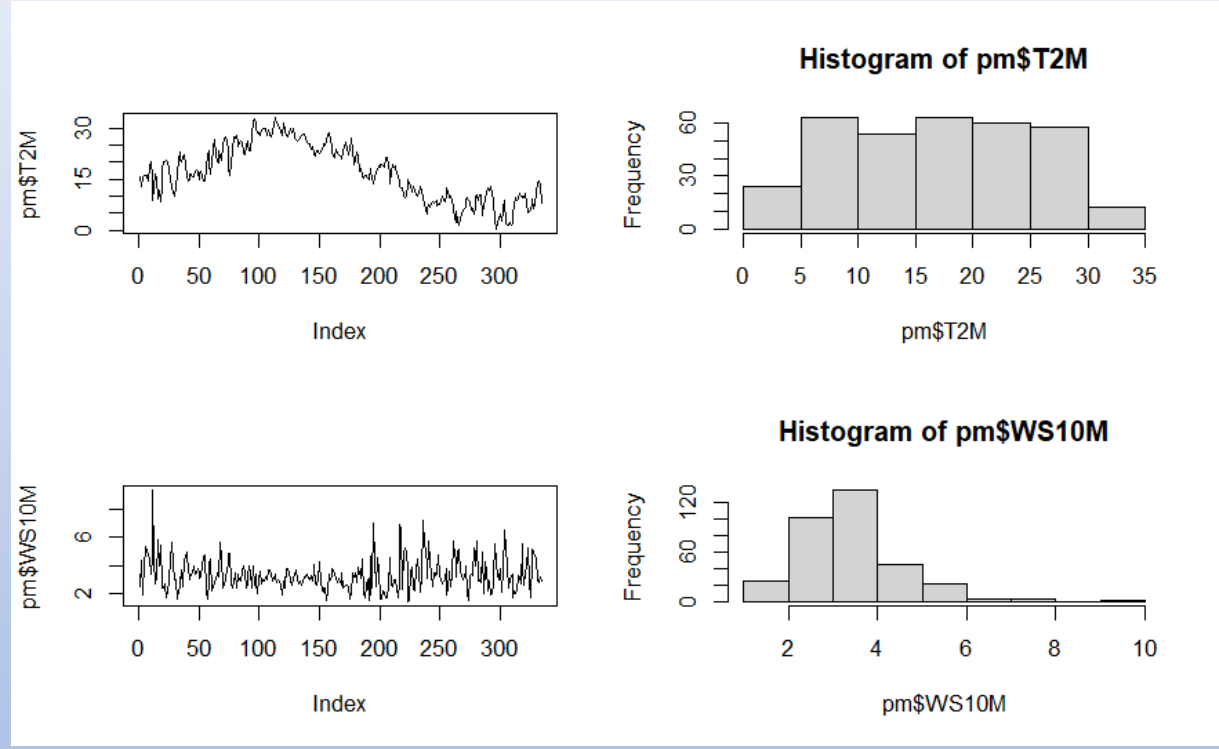
Predictor Variables:
PRECTOT (total precipitation at the surface of the earth in water mass)
T2M (the average daily air temperature at 2 meters above the surface of the earth)
WS10M (the average daily wind speed at 10 meters above the surface of the earth)
pm25 (daily average PM2.5 particulate matter where weights are based on theoretical hourly of the day with more sunshine)
Sunshine hours (daily sunshine length in hour)

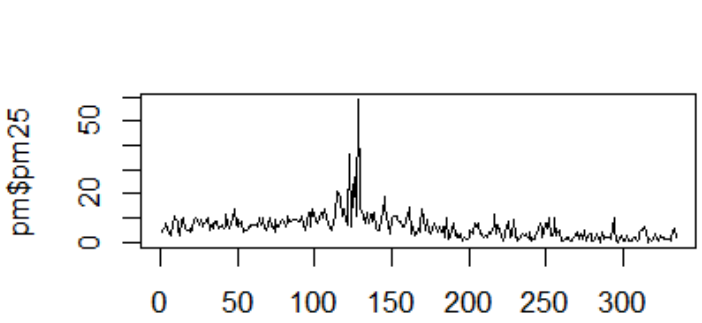
Response Variables:
Energy_generated_kWh (the daily production of the site)

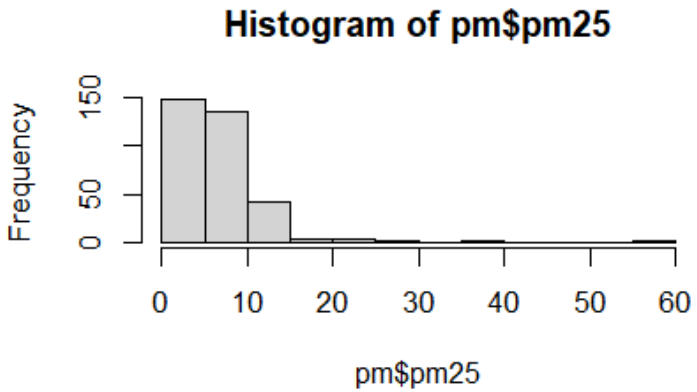
Exploratory Data Analysis

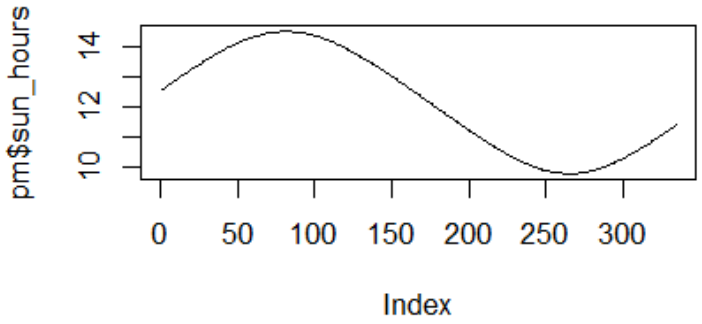


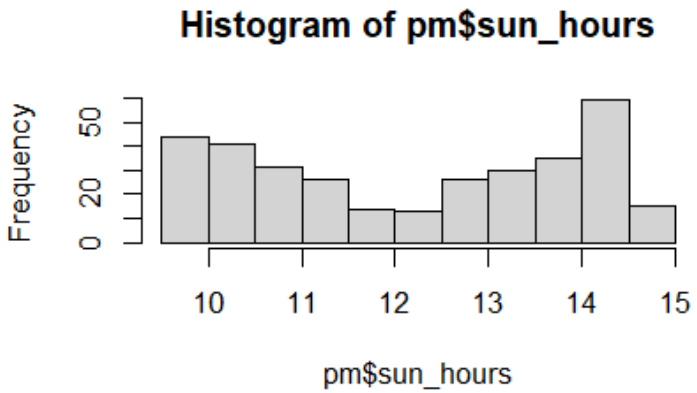

pm$energy_generated_kWh
0
150000
0
50
100
150
200
250
300
Index
Histogram of pm$energy_generated_kWh
Frequency
0
40
80
0
50000
150000
pm$energy_generated_kWh
pm$PRECTOT
0
10
25
0
50
100
150
200
250
300
Index
Histogram of pm$PRECTOT
Frequency
0
150
0
5
10
15
20
25
30
pm$PRECTOT


pm$T2M
0
15
30
0
50
100
150
200
250
300
Index
Histogram of pm$T2M
Frequency
0
30
60
0
5
10
15
20
25
30
35
pm$T2M
pm$WS10M
2
6
0
50
100
150
200
250
300
Index
Histogram of pm$WS10M
Frequency
0
60
120
2
4
6
8
10
pm$WS10M


pm$pm25
0
20
50
0
50
100
150
200
250
300
Index


Histogram of pm$pm25
Frequency
0
50
150
0
10
20
30
40
50
60
pm$pm25


pm$sun_hours
10
12
14
0
50
100
150
200
250
300
Index


Histogram of pm$sun_hours
Frequency
0
20
50
10
11
12
13
14
15
pm$sun_hours

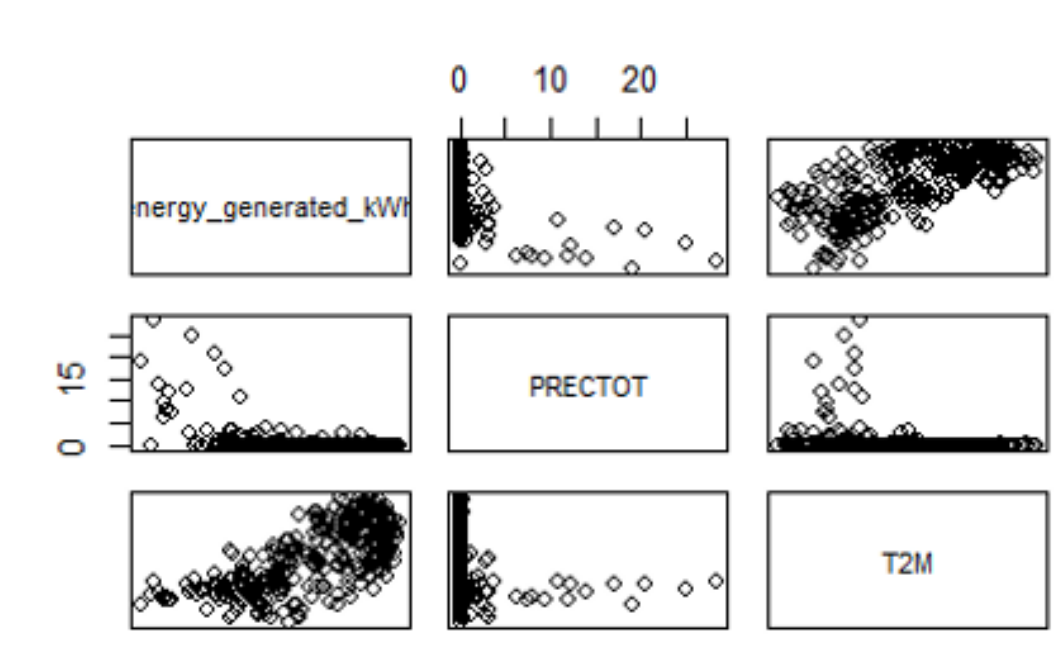
Model Selection

Preliminary model

energy_generated_kWh~PRECTOT+T2M+WS10M+pm25+sun_hours

Model 1 (Stepwise selection)

energy_generated_kWh=
-166758.6 - 3943.4*(PRECTOT) - 745.5*(pm25) + 27116*(sun_hours)

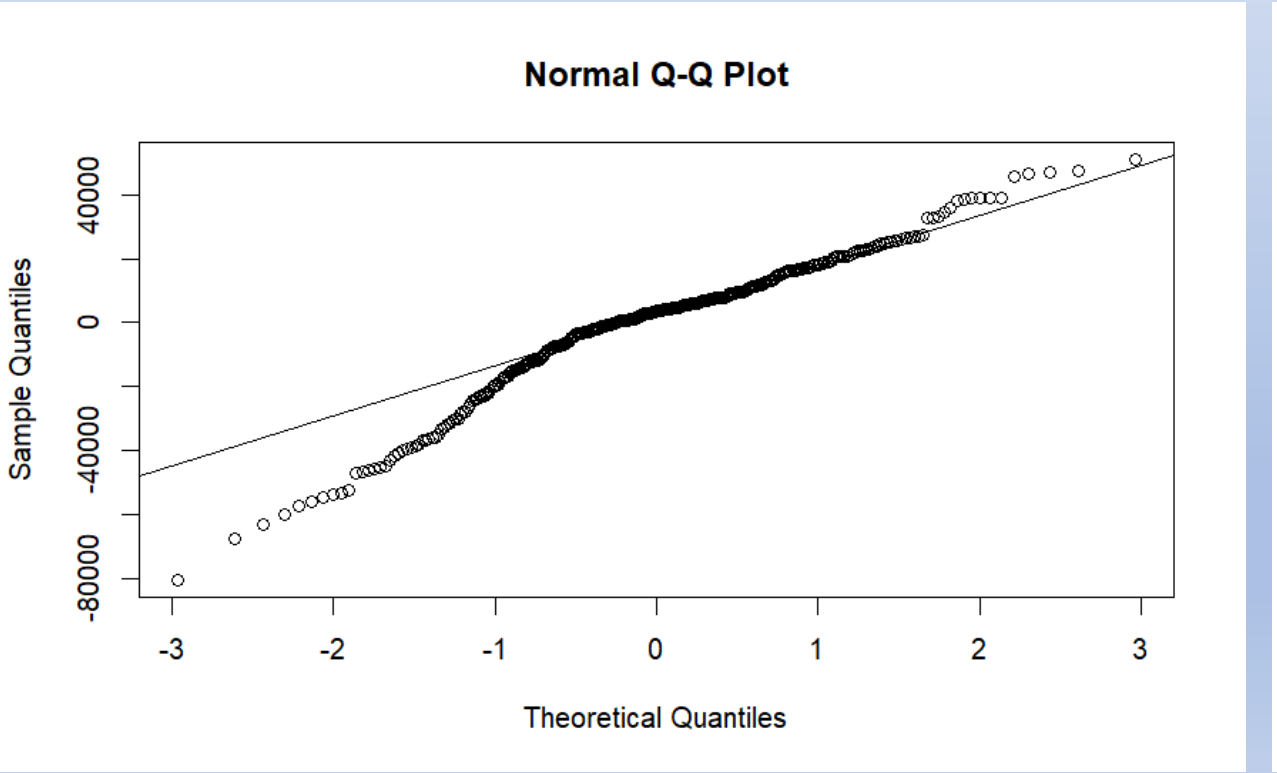


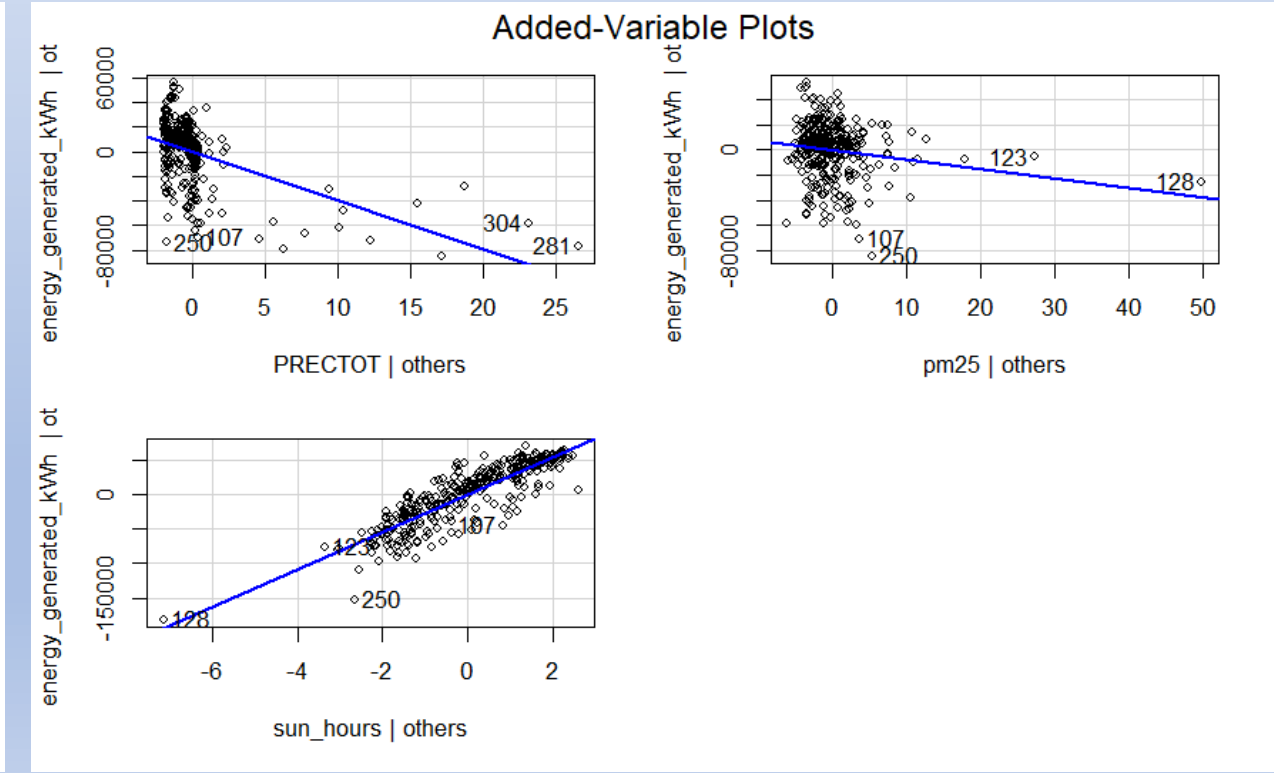
	Df	Sum of Sq	RSS	AIC
<none>			1.5146e+11	6665.4
+ WS10M	1	8.8062e+08	1.5058e+11	6665.5
+ T2M	1	1.9563e+08	1.5126e+11	6667.0
- pm25	1	3.6778e+09	1.5513e+11	6671.4
- PRECTOT	1	4.7778e+10	1.9923e+11	6755.0
- sun_hours	1	4.8139e+11	6.3285e+11	7141.0

T2M removed in model selection

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -166758.6     9595.5 -17.379  < 2e-16 ***
PRECTOT       -3943.4      386.5 -10.203  < 2e-16 ***
pm25           -745.5      263.4  -2.831  0.00493 **
sun_hours     27116.0      837.3  32.386  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21420 on 330 degrees of freedom
Multiple R-squared:  0.8397,    Adjusted R-squared:  0.8383
F-statistic: 576.3 on 3 and 330 DF,  p-value: < 2.2e-16
```

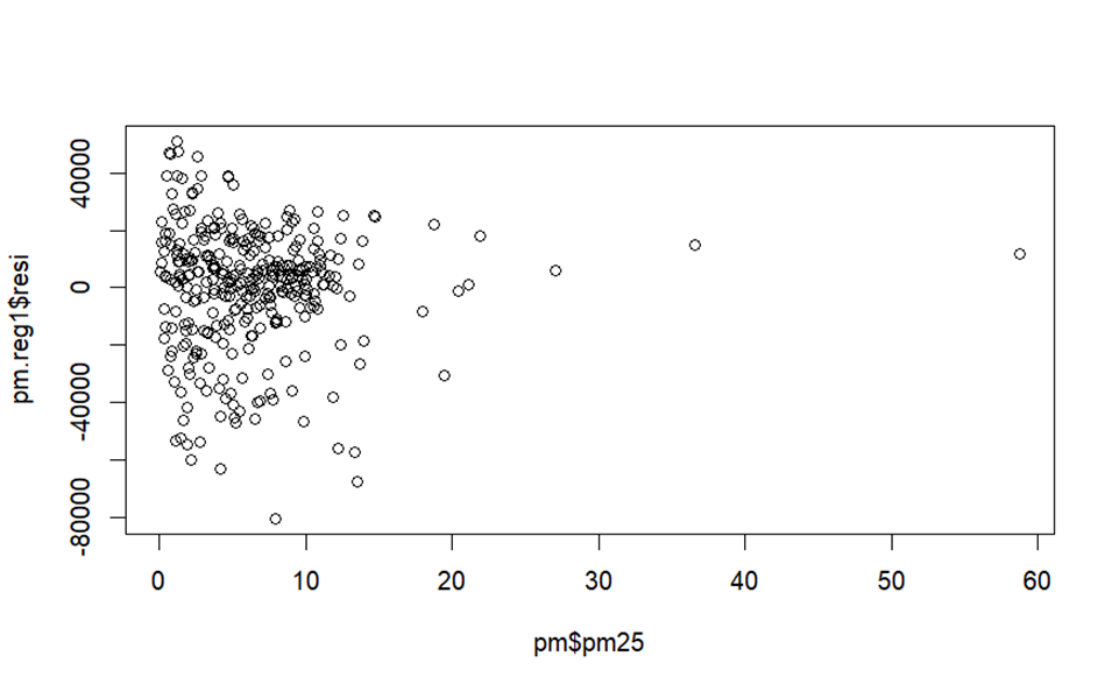




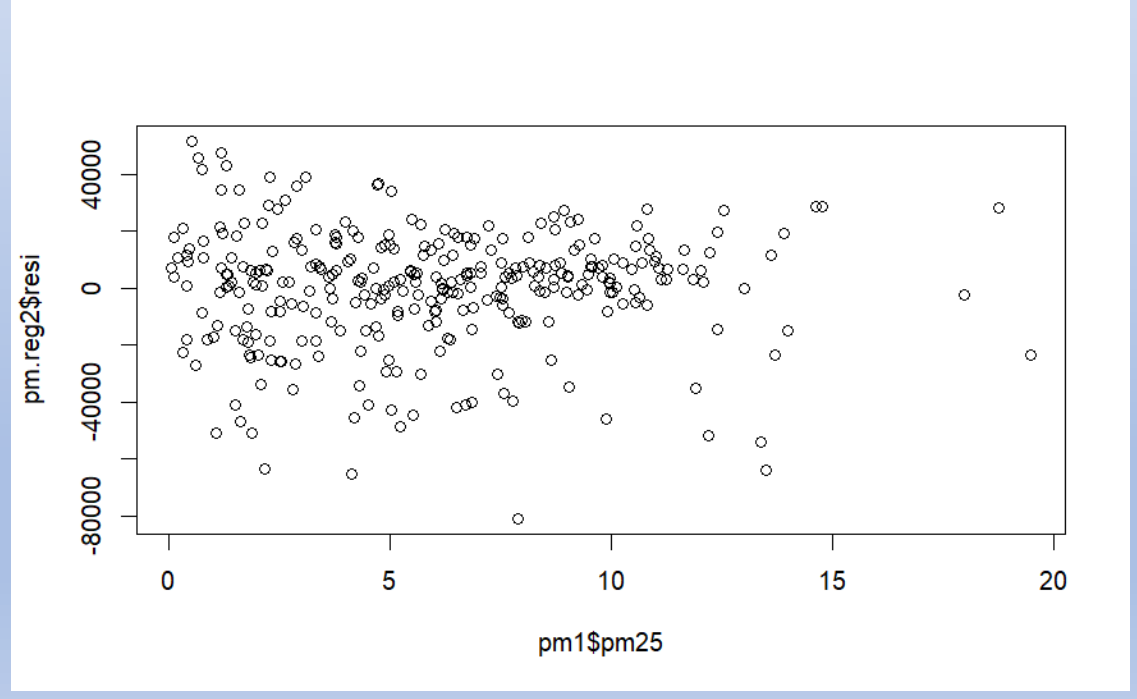
Model 2 (Remove Outlying)

energy_generated_kWh=
-160638.1 - 10397.4*(PRECTOT) - 1343.8*(pm25) + 27025.4*(sun_hours)

Some extreme situations

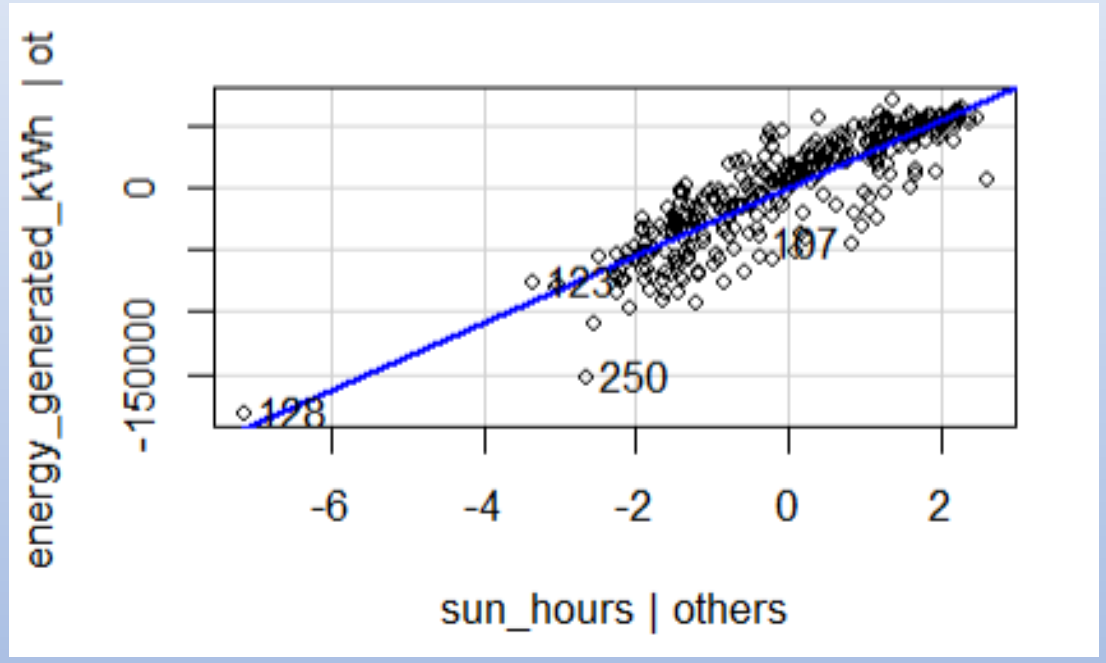


Model 1

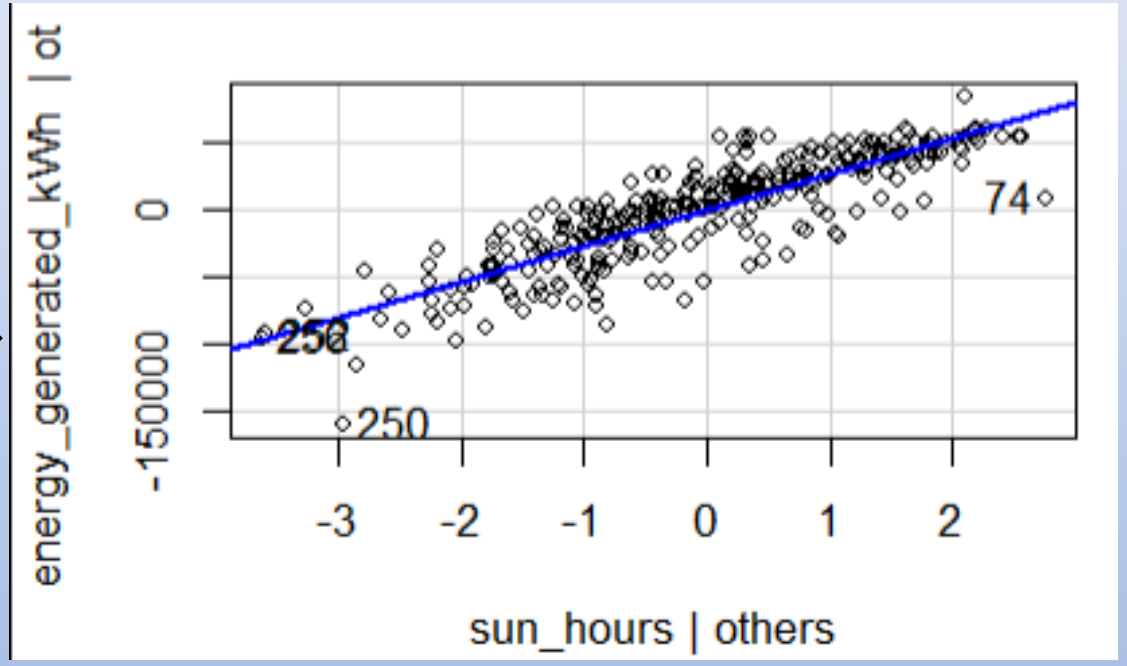


Model 2

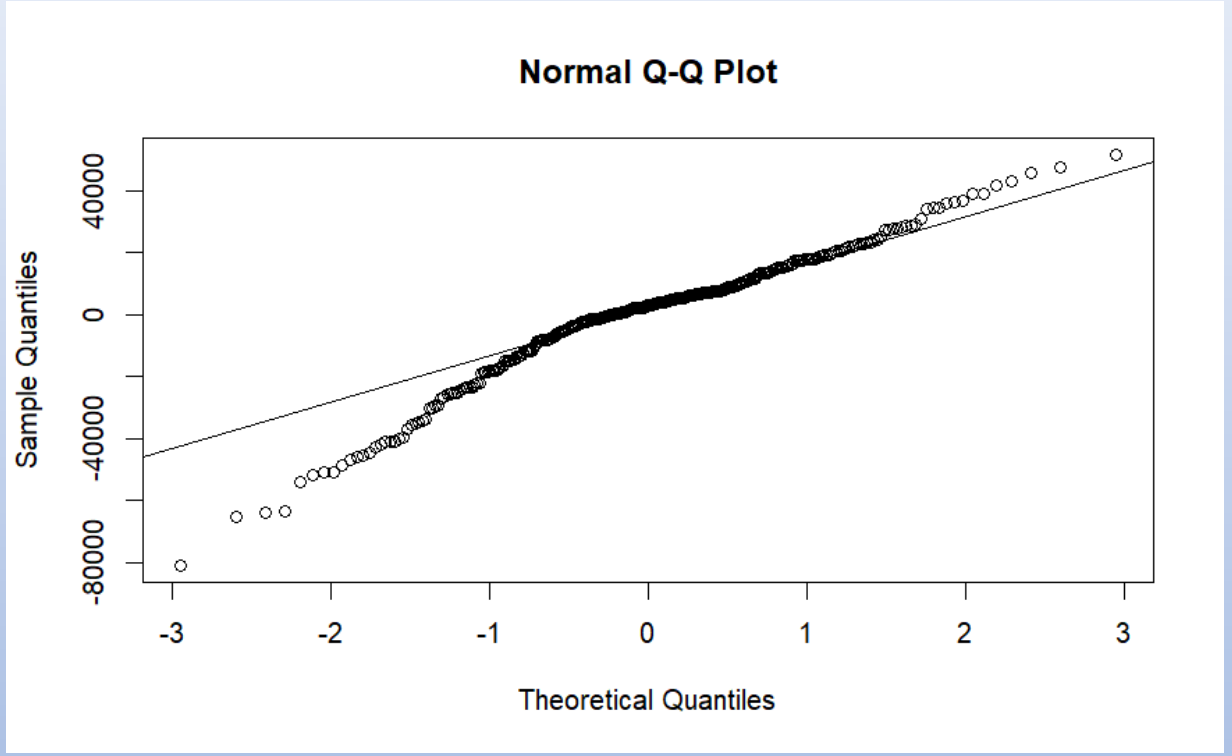
Some extreme situations: this will not in our research scope

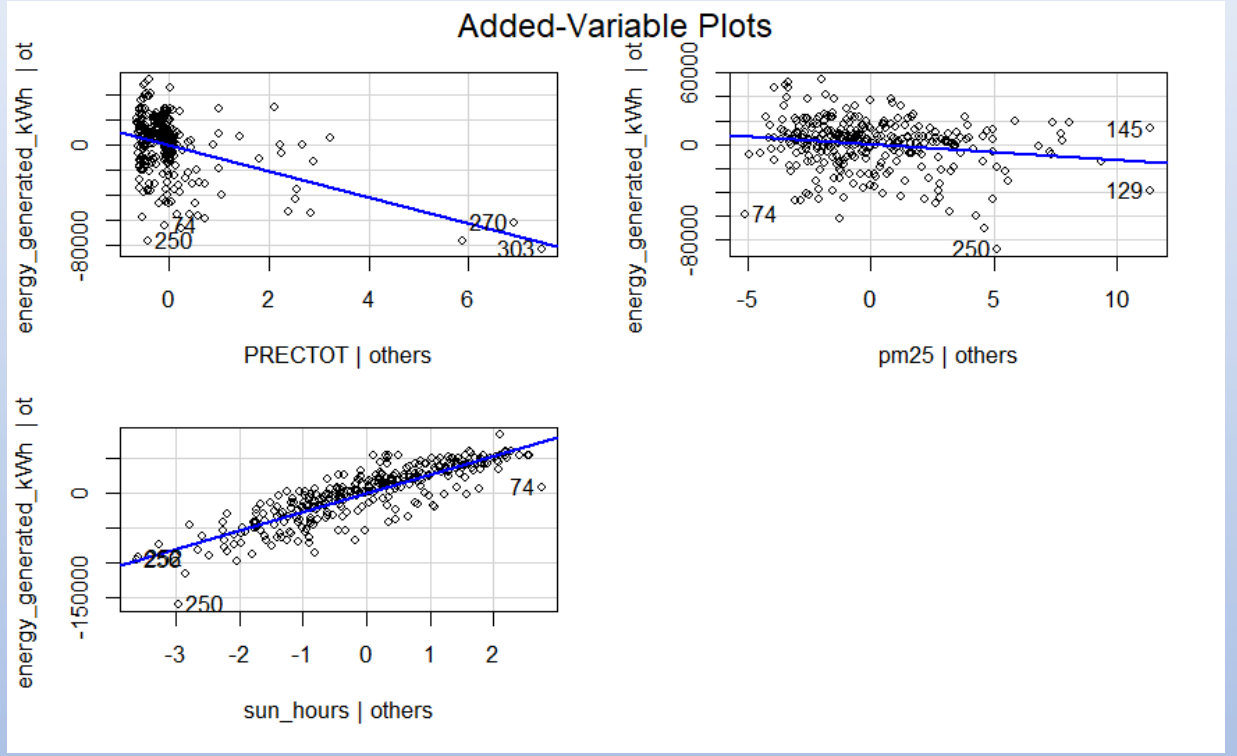


Model 1



Model 2

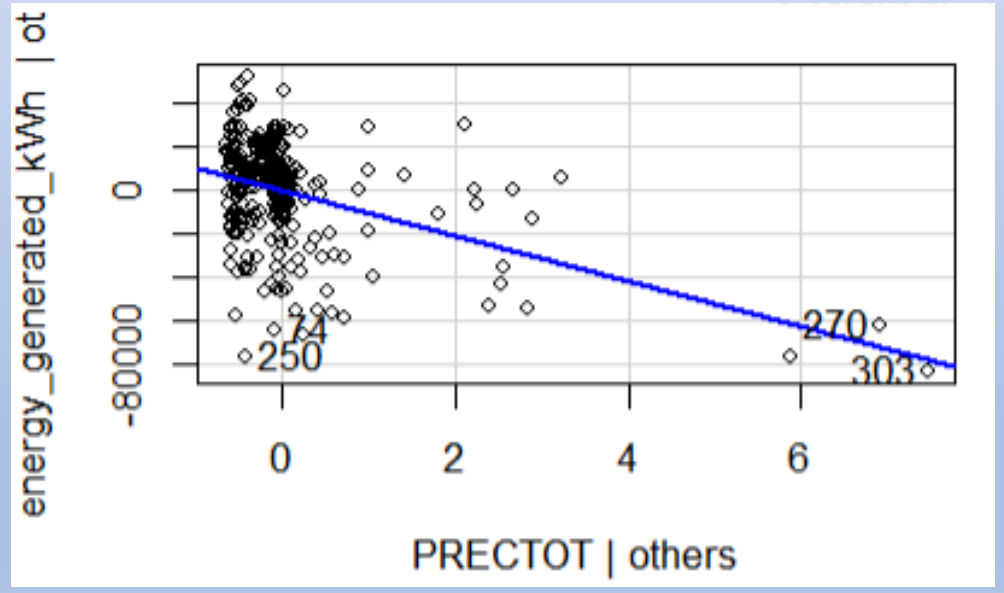

Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
40000
0
-40000
-80000
-3
-2
-1
0
1
2
3


Added-Variable Plots
energy_generated_kWh | ot
PRECTOT | others
pm25 | others
sun_hours | others
74
250
270
303
145
129
60000
0
-80000
-150000
-5
0
5
10
-3
-2
-1
1
2
4
6

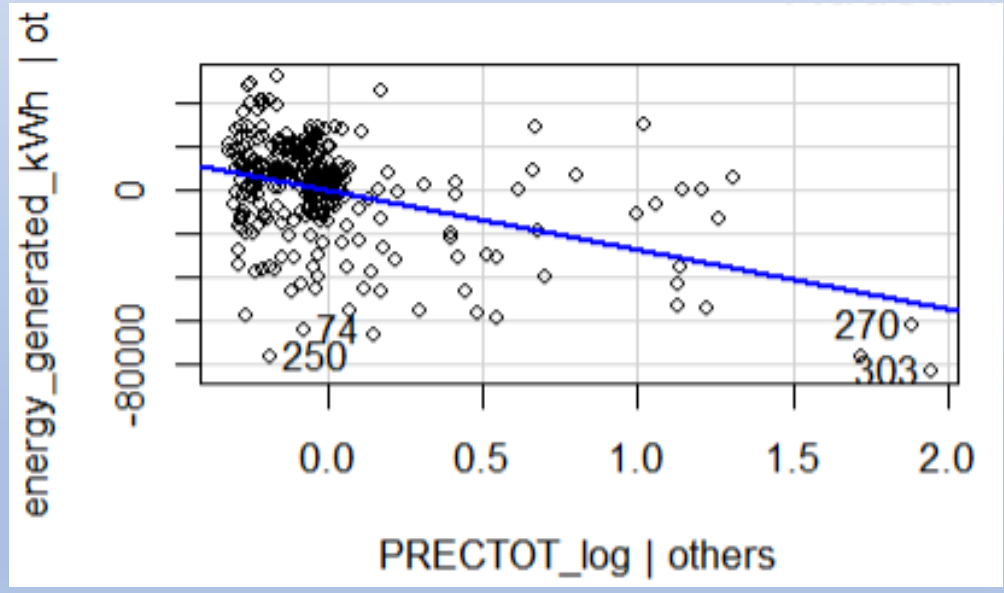
Model 3 (Log transformation)

energy_generated_kWh=
-158544.1 - 27172.9*(PRECTOT_log) - 1488.2*(pm25) +
26996.1*(sun_hours)

Method: log(pm1$PRECTOT+1)



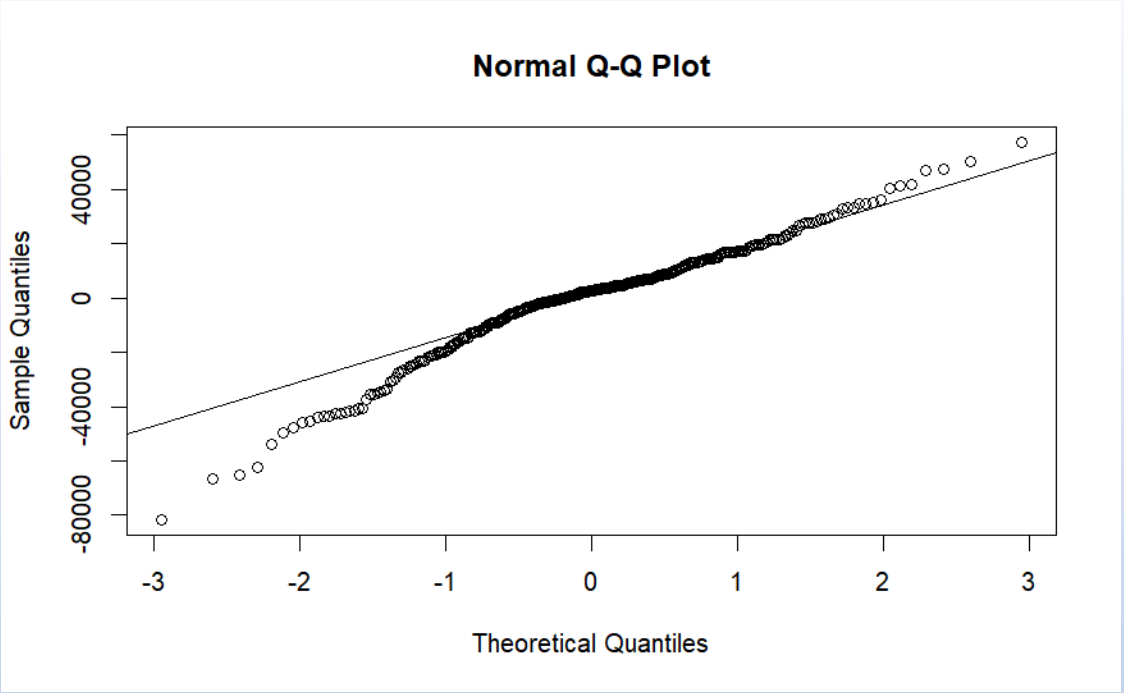
Model 2

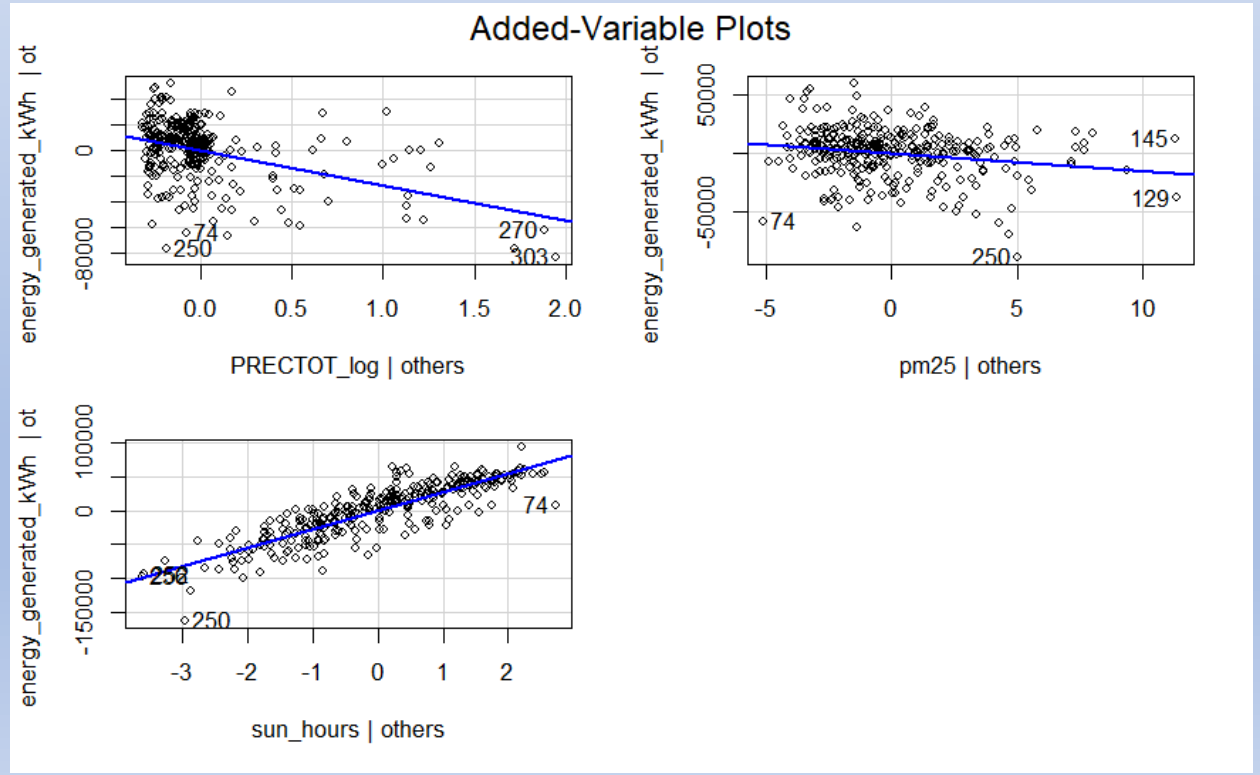


Model 3

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -158544.1    10006.5 -15.844  < 2e-16 ***
PRECTOT_log  -27172.9     3386.3  -8.024 2.05e-14 ***
pm25           -1488.2      417.7  -3.563 0.000424 ***
sun_hours      26996.1      917.3  29.430  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20540 on 314 degrees of freedom
Multiple R-squared:  0.8338,    Adjusted R-squared:  0.8322
F-statistic: 525.2 on 3 and 314 DF,  p-value: < 2.2e-16
```





Constant variance assumption valid

Durbin Watson Test

energy_generated_kWh=
-158544.1 - 27172.9*(PRECTOT_log) - 1488.2*(pm25) + 26996.1*(sun_hours)

```
        Durbin-Watson test

data:  pm.reg3
DW = 1.3463, p-value = 2.801e-09
alternative hypothesis: true autocorrelation is not 0
```

- Reject the null hypothesis. The residuals (errors) are not independent of each other. This is likely caused by a relationship over time.
- Remedial measures outside the scope of this model (Generalized Least Squares).
- 1.3463 is a moderate positive autocorrelation

Conclusion

The results meet our assumptions:

- pm25 affects solar photovoltaic production
- Sun hours have the dominant effect on power generation
- Presence of autocorrelation (weather typically has seasonal trends).

Strengths	Weaknesses
- 3 different iterations of the original model - Conclusions are logical	- Additional remedial measures such as analysis of dffits / dfbetas could be done - 3 variables in final model - No interaction terms

References

Image source:

https://www.archdaily.com/908571/california-approves-rule-requiring-solar-panels-on-new-homes

Data sources:

https://catalog.data.gov/dataset/dataset-for-evaluating-the-impact-of-wildfire-smoke-on-solar-photovoltaic-production

https://www.sciencedirect.com/science/article/pii/S0306261923006670?via%3Dihub#sec1.2

https://www.timeanddate.com/sun/usa/california-city?month=10&year=2018