



A2: Regression

03.03.2020

Team: The Credibles

Menghong Han, Peihan Tian, Yanghe Liu, Laurence Finch

Part 1. Descriptive Statistics

Variable Description

First of all, we choose the relevant and important variables based on the distribution and boxplot from our last report, they are: "BIOGAS", "BIOMASS", "GEOTHERMAL", "SMALL.HYDRO", "SOLAR.PV", "SOLAR.THERMAL", "WIND.TOTAL", representing power production from various power sources (measured in megawatts). Also, because seasonality has a large influence on energy generation, we convert the variable "TIMESTAMP" into two categorical variables "Hour" and "Month".

Variable Type	Variable Name	Description	Type
Important Variables	BIOGAS	Biogas production in MW	Integer
	BIOMASS	Biomass production in MW	Integer
	GEOTHERMAL	Geothermal production in MW	Integer
	SOLAR.PV ¹	Solar Photovoltaic production in MW	Integer
	SOLAR.THERMAL ²	Solar thermal production in MW	Integer
	WIND.TOTAL	Wind power production in MW	Integer
	Hour	00:00 as 1, totally 24 values	Categorical
	Month	Totally 12 values	Categorical
Target Variable	SMALL.HYDRO ³	Small hydro production in MW	Integer

Figure 1. Variable Description

¹ Solar Photovoltaic (PV) is a technology that converts sunlight (solar radiation) into direct current electricity by using semiconductors. When the sun hits the semiconductor within the PV cell, electrons are freed and form an electric current.

² Solar thermal technologies capture the heat energy from the sun and use it for heating and/or the production of electricity[1]. This is different from photovoltaic solar

³ Small hydro energy production in California is defined as energy production from water related sources, at a facility with a capacity of 30MW or less. More information about small hydro can be found [here](<https://www.hydro.org/policy/technology/small-hydro/>).

Descriptive Analysis

And then we draw box-and-whisker plots for relevant and important variables.

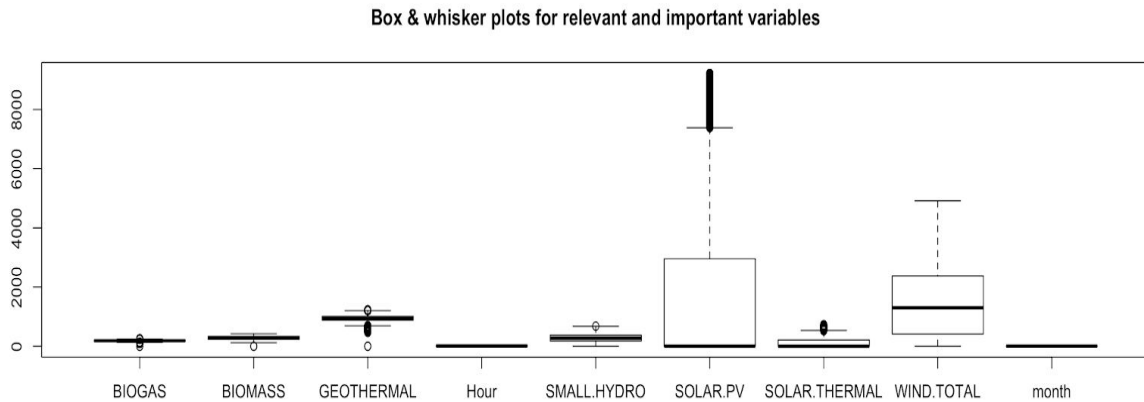


Figure 2. Box-and-whisker plots for predictor variables

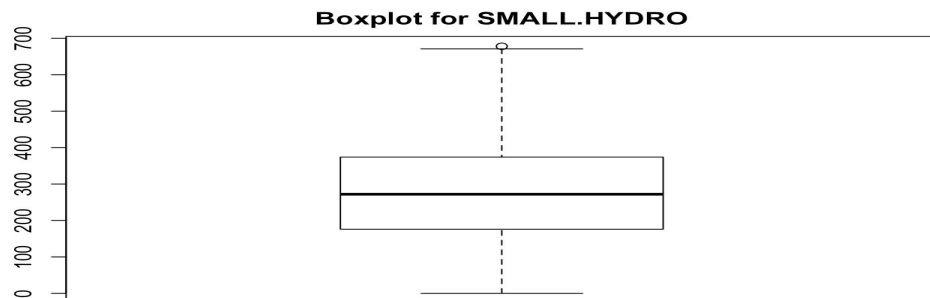


Figure 3. Box-and-whisker plots for target variable

From the plot, we can see that all variables have few outliers except for "SOLAR.PV", which shows a very large right deviation. We found that data highly concentrated on the frequency of 0 and shows positive skewness. However, this situation is reasonable since "SOLAR.PV" is Solar Photovoltaic which works only when the sun rises. Then, from the box chat of "SOLAR.PV" we can see the outliers show a large right deviation. Although Solar Photovoltaic works relatively fewer hours a day, it generates much more energy than others in a short period of time. Therefore, the large number of outliers would not destroy the data quality of "SOLAR.PV" and we should deal with the outliers later under the context of regression model.

Then we calculate the minimum, maximum, and average (mean, median, mode) and standard deviation and variance of important variables.

VARIABLES	MIN	MAX	MEAN	MEDIAN	MODE	STD	VARIANCE
BIOGAS	0	248	185.7	187	199	19.9	397.9
BIOMASS	0	423	283.6	283	232	59.6	3552
GETHERMAL	0	1230	945.3	928	921	89.5	8011.5
SMALL.HYDRO	0	678	284.3	272	134	132.1	17454.3
SOLAR.PV	0	5558	1491	3	0	2031.2	4125921
SOLAR.THERMAL	0	725	117.3	0	0	118.7	35621.6
WIND.TOTAL	0	4914	1478.7	1301	129	1135	1288768
Month	1	12	6.6	7	12	3.5	12.4
Hour	1	24	12.5	12.5	1	6.9	47.9

Figure 4. Descriptive statistics of important variables

To figure out potentially linear or curvilinear relationships among variables, we create scatter plots as follows.

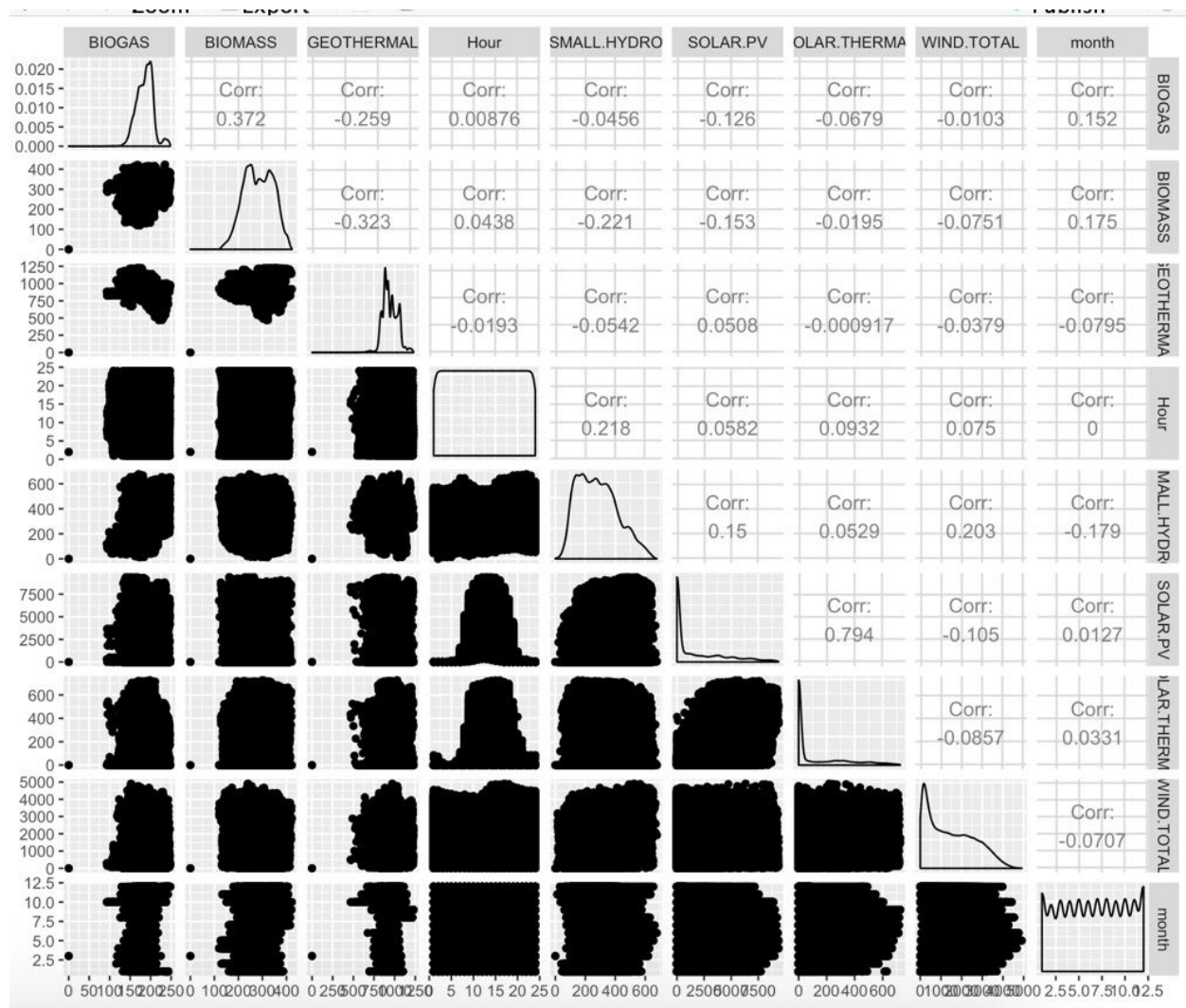


Figure 5. Correlation plots before sampling

Since there are too many data points, the scatterplots are hard to read, therefore, we randomly selected 100 variables to find clear patterns.

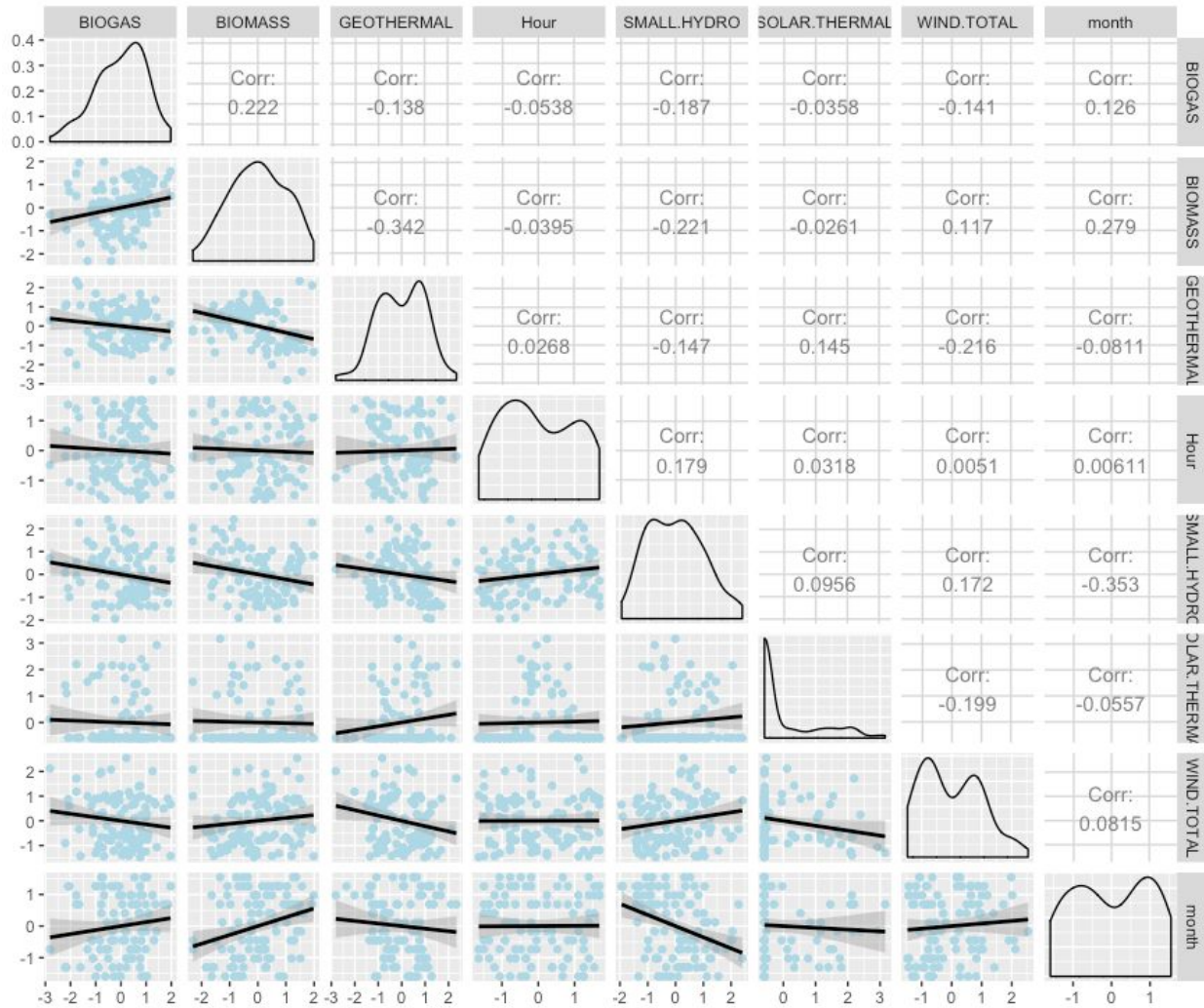



Figure 6. Correlation plots after sampling

Justify target variable

From the scatterplots and correlation coefficients, we can see the absolute value of correlation coefficients between “SMALL.HYDRO” and other variables are larger relatively, which we can see clearly from the slopes although all linear relationships are not that obvious. Therefore, finally we confirm our target variable “SMALL.HYDRO” which has better distribution and stronger linear relationships with others comparatively.

We chose small hydro as our target variable, after statistical and economic considerations. From the statistical side, SMALL.HYDRO follows the normal distribution, making it suitable for modelling. From the economic and business side, small hydro is one of the least



carbon-emitting energy sources, demand is likely to increase in the future, making understanding the factors that drive and complement its production highly relevant today. With the successful prediction of small hydro energy production, we can efficiently protect the environment, improve human welfare into the future as well as acting on responsible business values.

Part 2. Predictive Modeling: Multiple Regression

First, after loading our required libraries, we read in the data and divide it into training (80% of dataset) and validation sets.

For the reason given above, we choose SMALL.HYDRO as our target variable.

Small hydro energy production in California is defined as energy production from water related sources, at a facility with a capacity of 30MW or less. More information about small hydro can be found [here](<https://www.hydro.org/policy/technology/small-hydro/>).

We are interested in how the amount of small hydro energy production can be predicted using the other type of renewable energy and some time and month indicator variable.

We then fit a linear regression model, regressing the target variable, SMALL.HYDRO on all other variables. We used a forward stepwise fit to train our model. We also tried the backward and exhaustive regression, and the backward method has the highest R squared. As a result, we chose forward regression for further prediction. Then, we make predictions for the validation set.

Checking Regression Assumptions

Before interpreting the model and its predictions, we want to first check if the classical regression assumptions are satisfied, to enable us to conduct valid inference and prediction.

To check that the residuals are normally distributed, we plot a histogram of the residuals, and a Q-Q plot of the residuals. The plots below show that the residuals are approximately normally distributed.

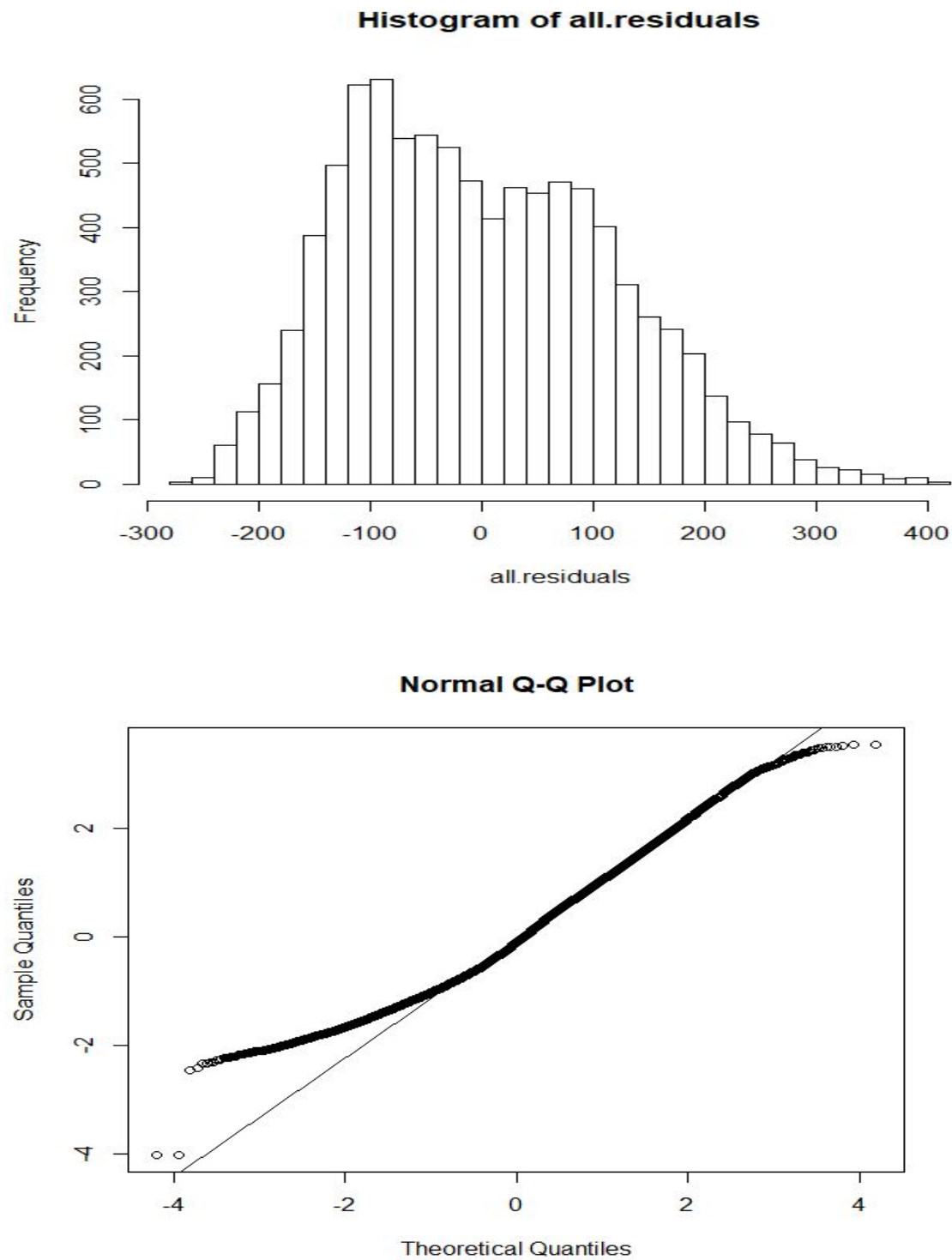
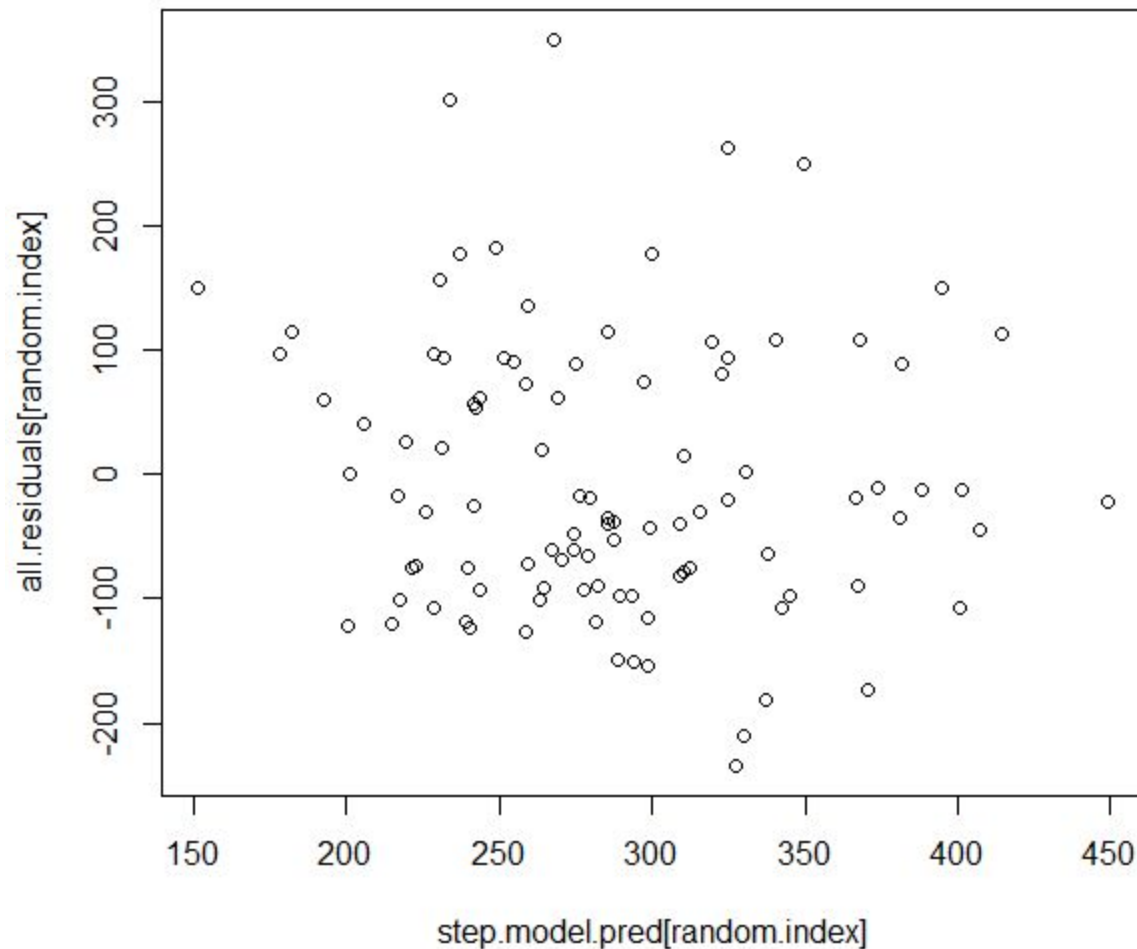


Figure 7. Histogram of residuals and Q-Q plot before improving



To check the normal distribution, we plot the QQplot and the Histogram of Residuals, from which we can see that the residuals have a left-skewed distribution and the residuals are a little far away from the QQ line, which means we can improve the residual plots later.

To check the linearity assumption, we plot a random sample of the residuals against the fitted values. The plot above shows that the model is linear in parameters. The residuals do look randomly distributed, so we have confidence in the linearity assumption.

So the regression assumptions seem to hold, meaning we can be confident in the validity of the regression.

Evaluating First Model

	RMSE (Validation set)	Adjusted R squared
Value	116.44	0.1889

We find an adjusted R-squared of 0.1889 and a RMSE of 116.44. Given this low adjusted R-squared and high RMSE, we feel we can do better by including some polynomial terms (quadratics and cubics) for the various types of renewable energy and some indicator variables for time and year.

Improving the model

We proceed by wiping our environment and building the model and parameters from the ground up for clarity.

In order to improve the R Squared and RMSE, we decided to add dummy, cubic, quadratic and interaction variables.

For categorical variables such as 'Hour' and 'Month', we can convert them into dummy variables. For continuous variables such as 'Solar power', 'Solar thermal', we can create polynomial terms for each type of energy production.

Model Training and Assumption Checking

Following the same steps as above, we partition the data, fit a model on the training set and use this model to train a stepwise model using a forward, AIC method.

We then make the same diagnostic plots as earlier, to check the regression assumptions.

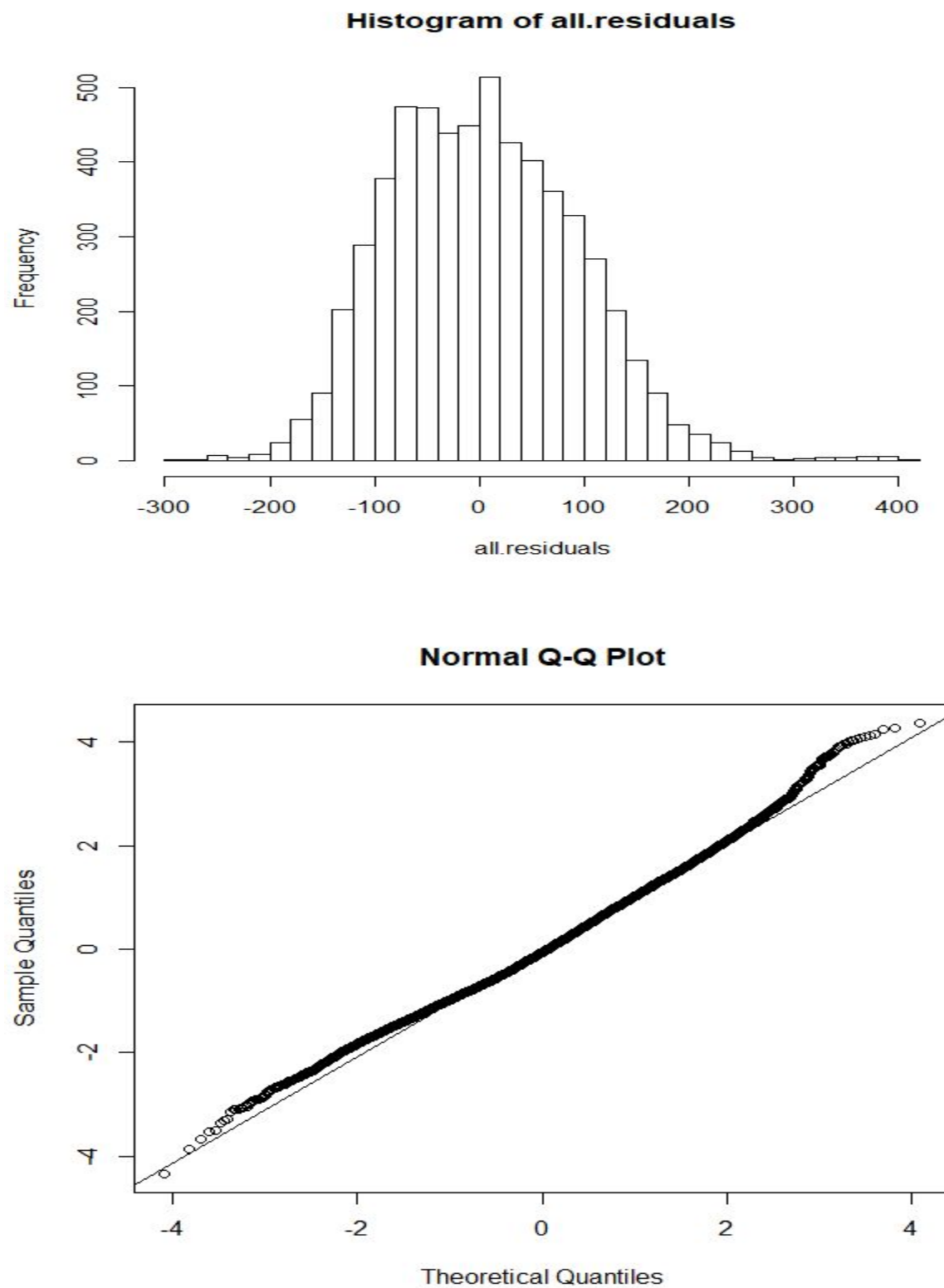
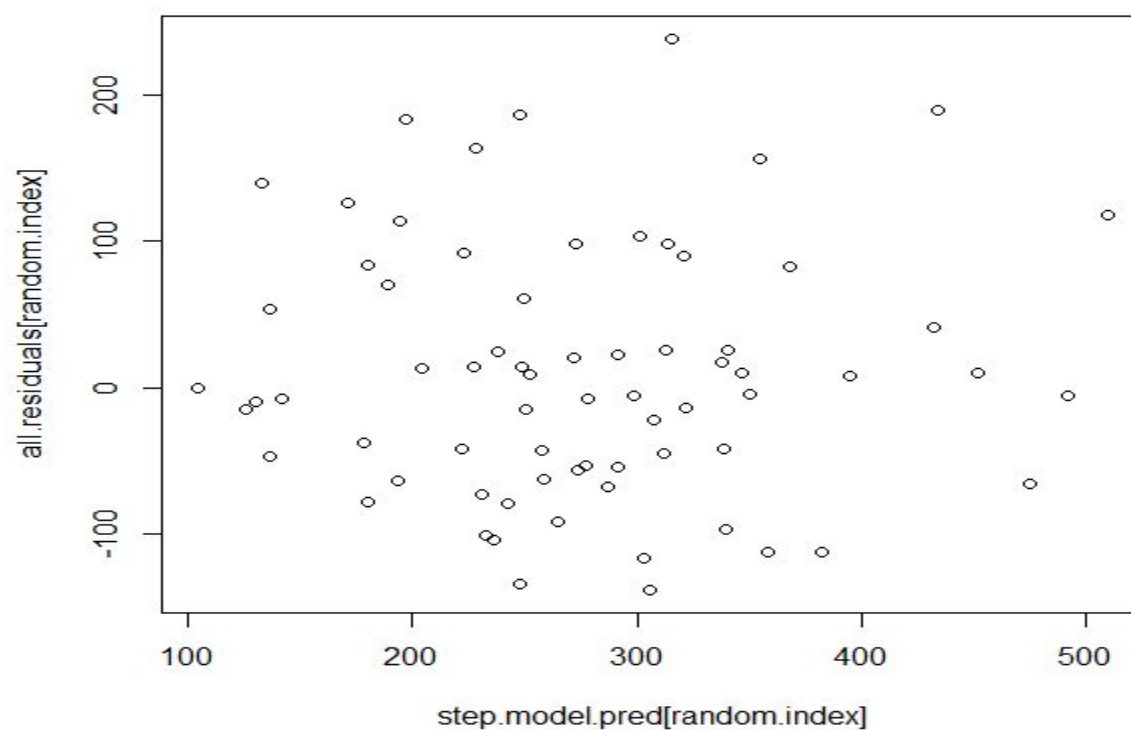


Figure 7. Histogram of residuals and Q-Q plot after improving



As before, we find that the residuals appear normally distributed and the linearity assumption holds. We can therefore be confident in using the model for predictions.

Comparing the residuals plots before and after adding more variables, we can find that the Histogram of Residuals looks Gaussian; the Normal Probability plot of Residual tracks the diagonal line more closely than before.

To make the Residual vs Fitted Value Plot more straightforward, we just randomly chose 100 samples from the whole dataset. We can see that the residuals look randomly distributed.

However, we find out there are few outliers after we graph the Residuals vs Fitted plot. So we decide to remove them and check if the performance of regression will be better.

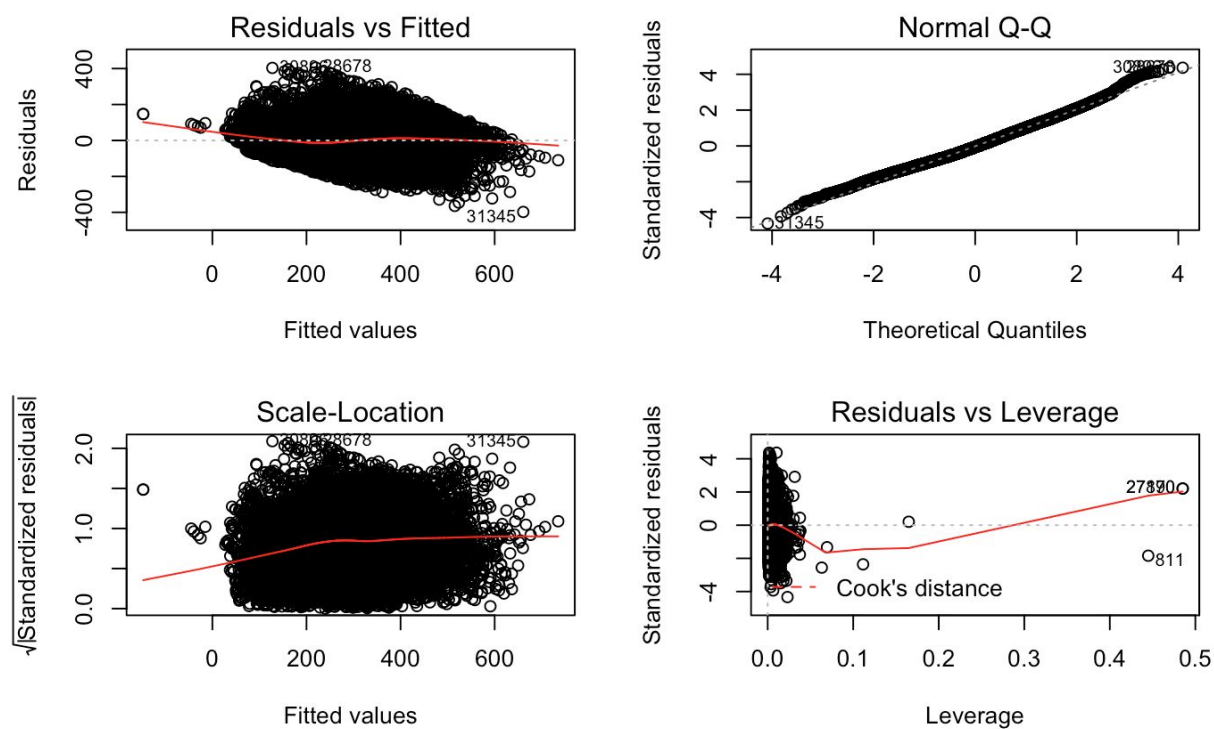


Figure 8. Residuals plots before removing outliers

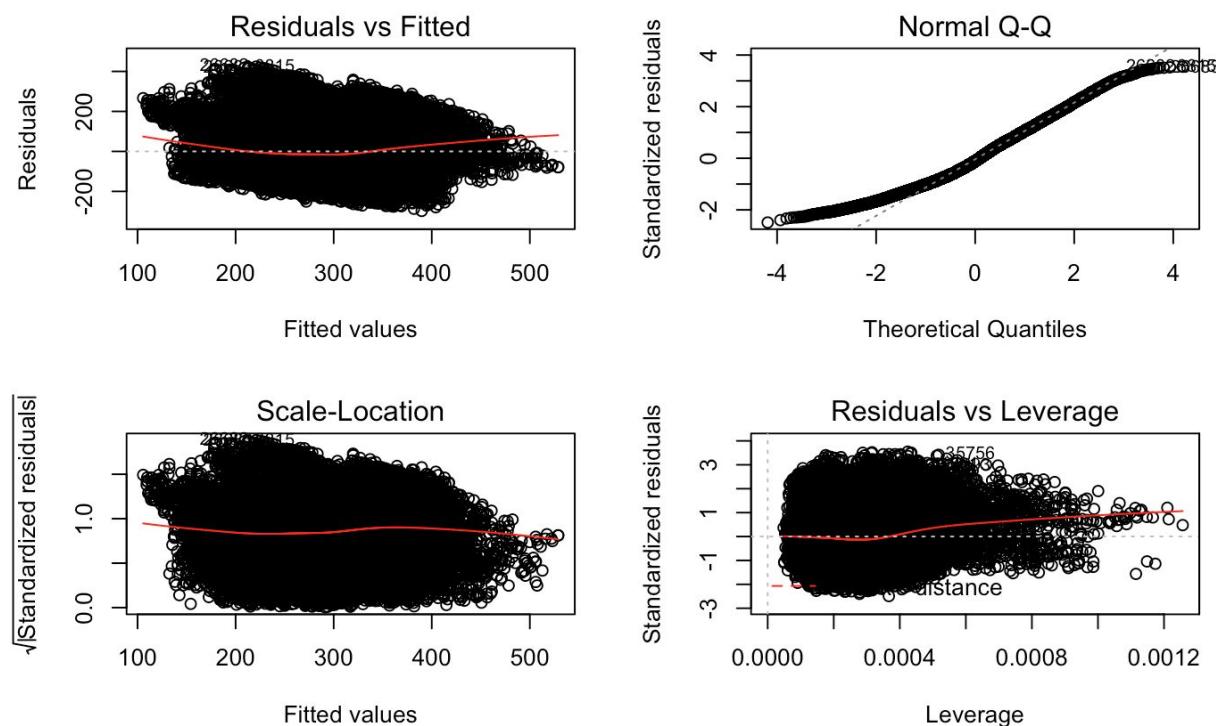


Figure 9. Residuals plots after removing outliers

After removing 12, which is 0.2% data points from the dataset, the residual scatter plot seems more scattered and more random distributed, also the some of the leverage points left in since extreme leverage points are removed.

Model Evaluation and Interpretation

	RMSE (Validation set)	Adjusted R squared
Value	89.948	0.5087

Now we get an ****adjusted R-squared of 0.5087****. This is much higher than previously, although still not that high. Thinking about the model however, this is not very surprising. There is no clear, direct causal path between small-hydro power production and other energy sources. For example, solar power will depend primarily on the amount of sunshine, whereas hydro power shouldn't be very correlated with the amount of sunshine. However some relationship between small-hydro power and our predictions is plausible. For example, if one renewable energy source is being particularly productive, other sources could be neglected to focus on the more productive source, or alternatively, the other


sources of energy could also be productive at the same time, if environmental conditions are favourable to the production of both sources of energy.

Looking at specific parameter estimates, we see that there is a strong positive relationship between biogas production and small-hydro production (coefficient of 47.48 on BIOGAS, meaning an increase in 1 unit of biogas production predicts an increase of 47.48 units of small hydro production, not considering the small polynomial terms), although this effect gets slightly weaker at higher levels of biogas production, seen by the negative coefficient on the BIOGAS_² coefficient.

SMALL.HYDRO is also predicted increasing with BIOMASS energy production (coefficient of 1.91, meaning an increase of one unit of biomass energy predicts a 1.91 unit increase in small hydro production, not including the small polynomial terms) and decreasing in GEOTHERMAL (coefficient of -5.53, meaning an increase of one unit of geothermal energy predicts a decrease of 5.53 units of small hydro, not considering the small polynomial terms). The coefficients on SOLAR.PV, SOLAR.THERMAL and WIND.TOTAL are small, so are not very correlated with small hydro power production. (SOLAR.PV and SOLAR.THERMAL coefficients are not statistically significant, possibly because of their high collinearity)

The coefficients on the polynomial terms are generally negative, meaning that small hydro power is predicted to be increasing slightly less at higher levels of alternative renewable energy production, given the generally positive relationships between small hydro and alternative power sources. The coefficient on BIOGAS_² is -0.2616 meaning that for an increase of one unit of biogas production, small hydro production is predicted to be 0.2616 x biogas energy production level, not including the standalone effect of biogas production. Similarly, the coefficient on BIOMAS_² is -0.04421, meaning that a one unit increase in biomass energy production predicts a 0.04421 x biomass energy production level in small hydro, not considering the standalone effect of biomass energy production on small hydro. These findings are in agreement with our intuition that higher energy production from other sources could take focus away from small hydro, even if environmental conditions or energy demand mean that small hydro production is higher when alternative energy production is higher.

Perhaps the most interesting and most interpretable is the coefficients on the time and month indicator variables. For time, we use midnight (hour 24) as the omitted reference hour. In the early hours, before 7am, small.hydro energy production is below the midnight reference (negative coefficients: coefficient on hour 2 of -29.14 meaning small hydro energy production is 29.14 units per hour less at 2am than at midnight for example). After



7am production is above that of midnight (positive coefficients: the coefficient of 116.4 on hour 12, being the peak production hour, meaning at 12pm hourly small hydro production is 116.4 units higher than at midnight), with the highest production between 10am and 3pm.

For months, December (month 12) is the omitted reference month. Small hydro energy production is predicted lower than December in the months of September, October and November, and higher in all other months..

The interaction terms are very small and vary in sign with the largest being BIOGAS:BIOMASS at 0.0061. These coefficients are hard to interpret and possible not as useful as the above findings, so we include the interaction terms mainly to increase the accuracy of the model

Nearly all of the coefficients are highly statistically significant, with very small p-values. Given this and the satisfaction of the regression assumptions, we are confident that our model can be used to predict small hydro energy production.

With our improved model, we get a much better ****RMSE of 89.948****. This increase in accuracy is largely the result of including the time and month indicator variables. Although the new polynomial terms are mostly statistically significant, and are plausible as argued above, the larger coefficients on the time and month indicator variables are the main improvement in our model.



Conclusion and Result Insights

The ability to predict small hydro energy production using alternative energy sources and time and month categories is useful for all who care about specific sources of energy production for whatever reason. For policymakers and energy planners, it is useful to be able to estimate a time/year dependence of small hydro energy production, to be able to plan for energy demand or supply shocks. For example, in the case of a particularly cloudy summer where solar energy supply drops, policy makers could look to our model to estimate knock-on or indirect effects to small hydro energy production, that might not be immediately obvious otherwise.

Small hydro energy production is very environmentally friendly, producing very little if any carbon emissions. We believe that our model is most useful for obtaining a better understanding of the renewable energy production landscape, for identifying complementary and non complementary energy sources to small hydro and finally for identifying and forecasting time and month specific production.