




# Multiple Regression with R

Team TBD



## PART I Data Wrangling

Dataset Source: <https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather> (Changed from last time)

Major Wrangling steps are documented here in the preparation session, namely, dropping meaningless columns that are constituted by 0s and NaNs, filling in the blanks using linear interpolation with a forward direction, and concatenating weather and energy dataset together. The aim of these three major steps are respectively to raise the efficiency of future data mining, to get the complete observations and to introduce weather features to the correlation analysis with energy demand, mainly to avoid endogeneity.<sup>1</sup>

As the energy dataset is on nation (Spain)-scale while the weather set is on city-scale, we decide to average the indexes to make our upcoming analysis and prediction of domestic hourly electricity demand more reasonable. We do this process is because the demand of these five cities constitute a large proportion of the overall demand. Besides, the five cities, marked with red stars, are in different latitudes from top to bottom and longitudes from left to right of the territory of Spain. This is the most direct approach.



Graph 1 *Five Cities' Locations in Spain*

---

<sup>1</sup> Please refer to process.py

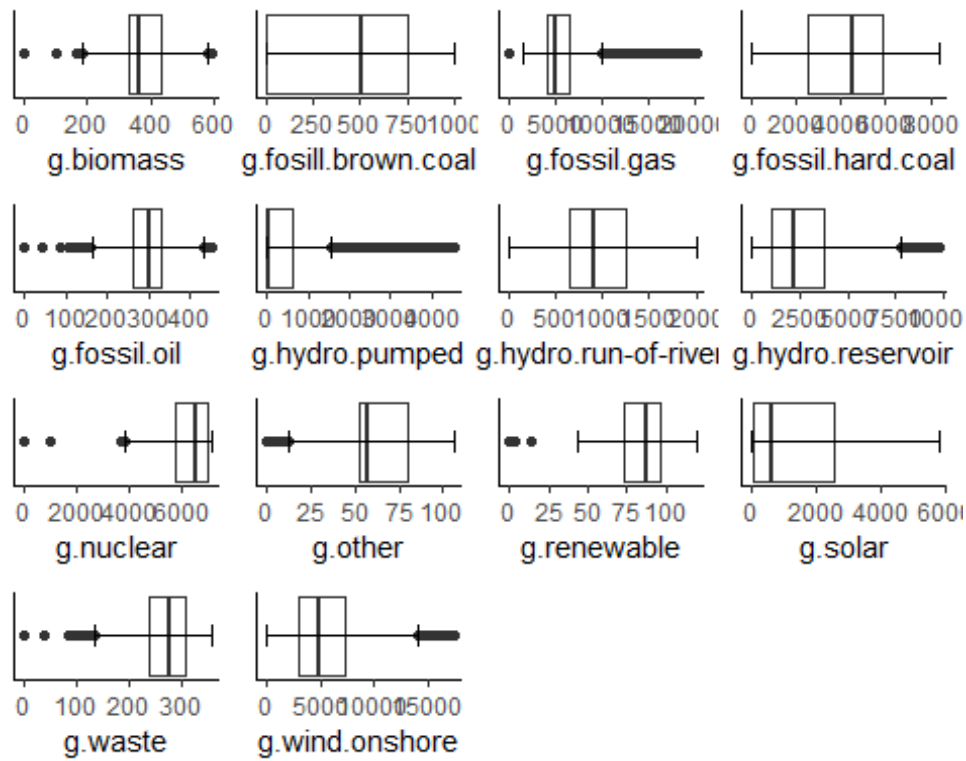
## PART II Descriptive Statistics

Multiple regression will be including many variables, so we need to confirm their validity and reliability. According to some reports and journals, we suspect electricity generations, weather features and time periods (we highly suspect) will have an impact on the hourly demand. Below is a chart and some box-and-whisker plots of key variables.

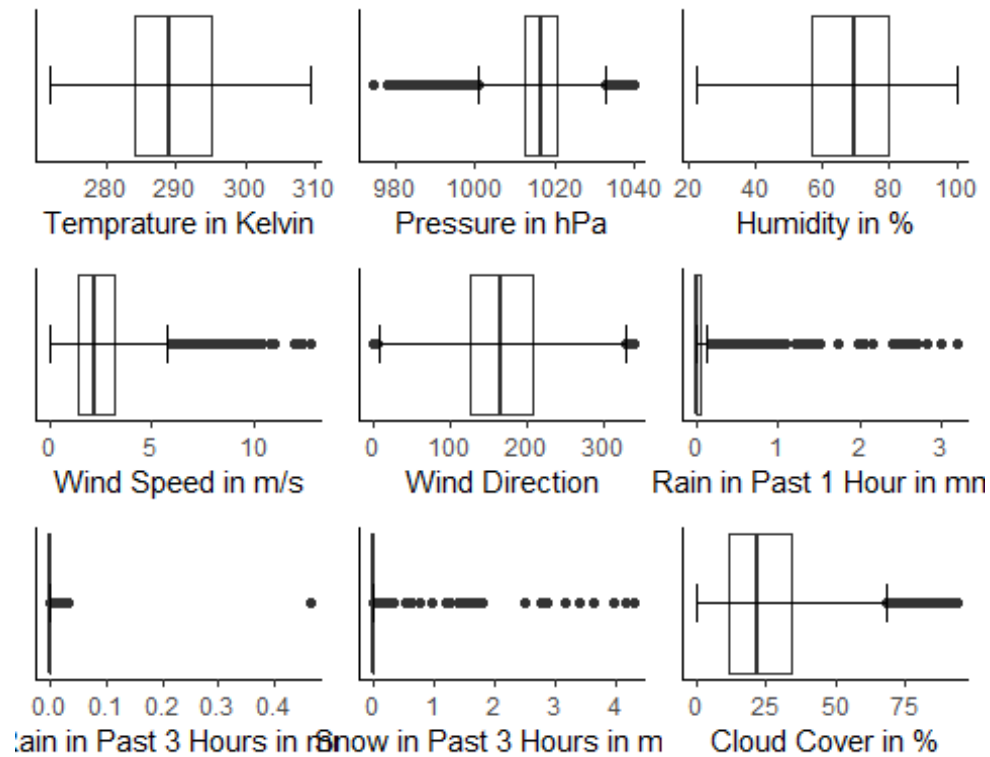
Variable	Mean	Median	Mode	Std. dev	var
generation.biomass	383.53	367	361	85.346	7284
generation.fossil.brown.coal.lignite	448.09	509	0	354.62	125754
generation.fossil.gas	5622.7	4969.5	3993	2201.5	5E+06
generation.fossil.hard.coal	4256.5	4475	5266	1962	4E+06
generation.fossil.oil	298.34	300	303	52.52	2758.3
generation.hydro.pumped.storage.consumption	475.58	68	0	792.31	627759
generation.hydro.run.of.river.and.poundage	972.2	906	600	400.71	160570
generation.hydro.water.reservoir	2605.5	2165	801	1835.2	3E+06
generation.nuclear	6263.5	6564	7102	840.27	706058
generation.other	60.226	57	57	20.239	409.61
generation.other.renewable	85.634	88	93	14.077	198.16
generation.solar	1432.8	616	26	1680	3E+06
generation.waste	269.42	279	317	50.218	2521.9
generation.wind.onshore	5465	4849.5	3932	3213.6	1E+07
total.load.actual	28698	28902	23665	4575.8	2E+07
price.actual	57.884	58.02	56.85	14.204	201.76
Temperature	289.71	289.04	294.5	7.2543	52.624
Pressure	1016.1	1016.8	1017	8.1863	67.016
Humidity	68.032	69.6	84.2	14.815	219.49
Wind Speed	2.469	2.2	1.4	1.3464	1.8128

Wind Direction	166.72	166.2	140	57.211	3273.1
Rain in past 1 hour	0.0693	0	0	0.1921	0.0369
Rain in Past 3 hours	0.0004	0	0	0.0034	1E-05
Snow in past 3 hours	0.0048	0	0	0.1011	0.0102
Clouds coverage rate	24.3441	22	0	16.9629	287.739

Table 1 *Statistic Description of Key Variables*

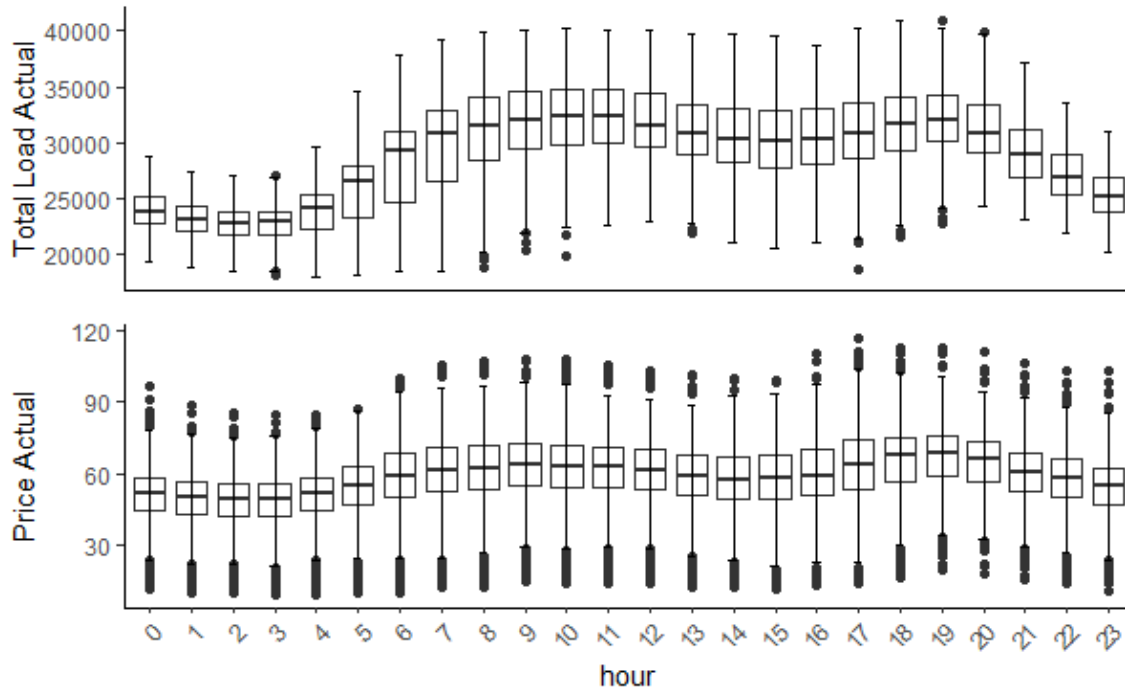


Plot 1 *Energy Generation Boxplot*



Plot 2 *Weather Feature Boxplot*

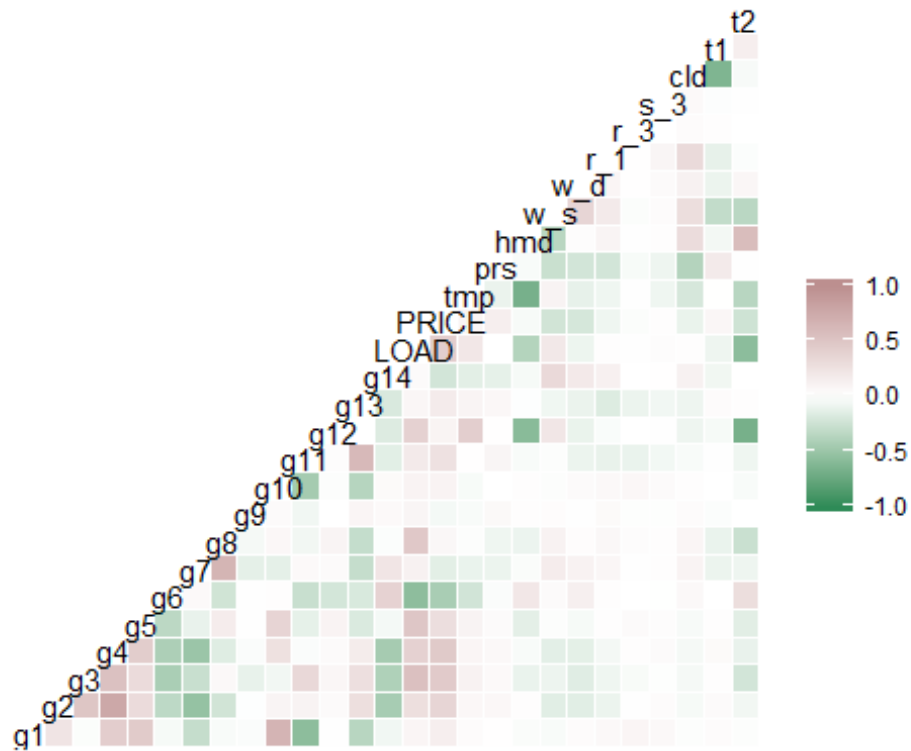
The table above tells us that variables are in different scales. Combined with box-and-whisker plots, we can see the distribution of all variables. Although we can see a lot of outliers in some potential variables, we are not arbitrarily removing them right now. One reason is that these variables are seasonal and naturally have outliers. The other is that we currently have no idea whether they will impact the multiple regression and the prediction model, so before the outcome it is better for us to save them for further analysis. In other words, we will leave it to the stepwise part to let the codes help us decide.



Plot 3 *Demand and Price by Hour*

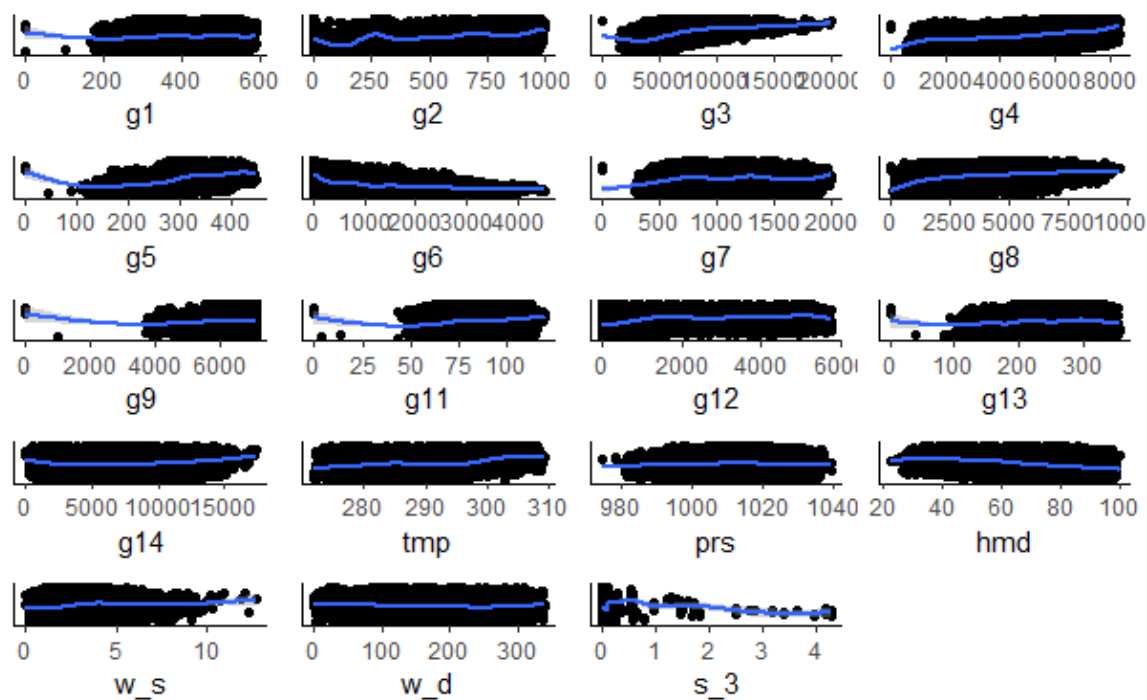
We mentioned before that we highly suspect LOAD and Price will be affected by time period. It is easy to catch that certain hours in a day, people will consume more electricity. And demand has a positive impact on price. Thus, we introduced plot 3 to verify the guess.

We can see an obvious volatility of LOAD over time, while a not so evident sign in PRICE. So, we split a day into two periods: 8 a.m. to 19 p.m. and 20 p.m. to 7 a.m. of the next day. Here we introduce two dummy variables t1 and t2 to distinguish between them. At this point, LOAD beats Price to be a better target variable.

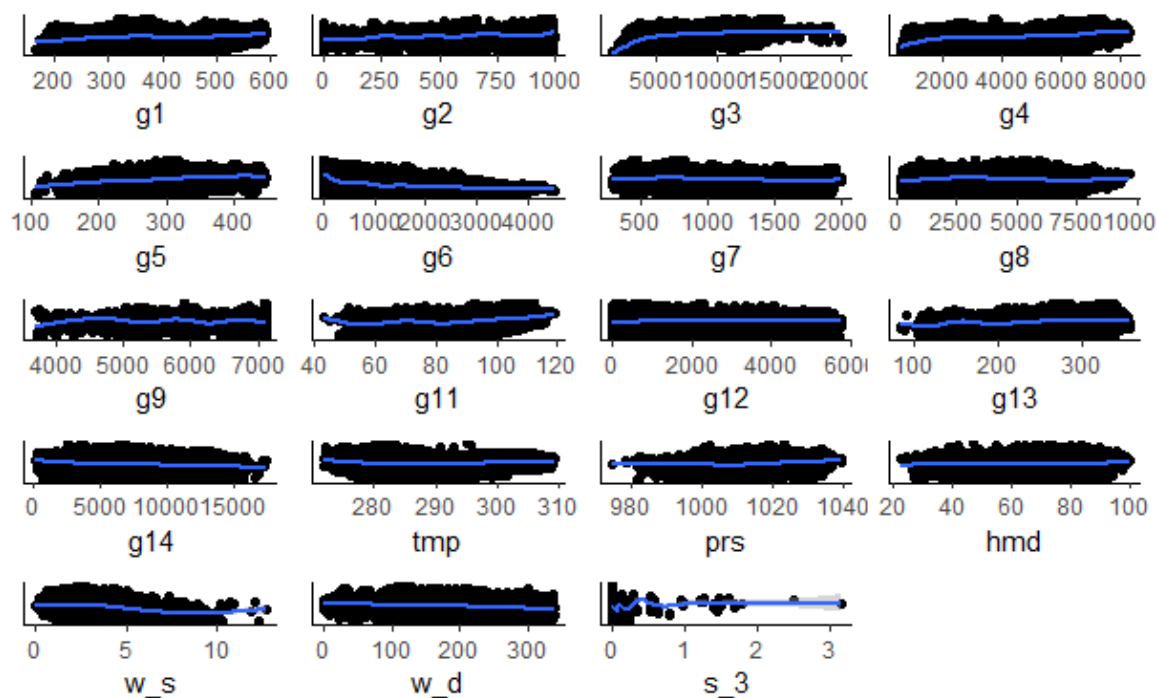


Plot 4 *Correlation Matrix*

The matrix gives us an impression of the correlation between every two randomly chosen variables. Based on common sense and the shade of squares, we think LOAD (energy demand) would be a better choice than PRICE to be the target variable.



Plot 5 Scatterplot of Load vs. Potential Variables



Plot 6 Scatterplot of PRICE vs. Potential Variables



The scatterplots also prove our judgement. As the trend lines are relatively flatter in PRICE scatterplot than in LOAD scatterplot. Besides, according to two multiple regressions where these two variables are dependent variables and others independent ones, LOAD model has a much higher  $R^2$  than PRICE does. From what has been discussed above, we take LOAD as the target variable.

## Part III Prediction Model - Multiple Regression

We divide the dataset into two segmentation on a scale of eight-to-two, one for training purpose and the other for validation, in order to show the reliability of our best model.

The basic logic is that we generate a multiple regression and then achieve a prediction model by doing a stepwise simulation to it. Then we compare the accuracy of the multiple regression with fitted values and the prediction model with valid data seg. Before reaching the end, we monitor the residuals by drawing some plots to make sure the residuals (predictions minus true values) are reasonable and in normal distribution, during which we removed some outliers from the observations. We are not worried about the modifications as there are over 30,000 records in our dataset. Finally, if the RMSEs are approximate, then we can conclude that the model after stepwise can be useful for prediction.

We refer to forward-and-backward selection and best subset (exhaustive) search to decide our final best model.

The original model (multiple regression) explains over 90% of the dataset as the adjusted  $R^2$  is 0.9167, but the RMSE of prediction in valid set is 6395.828, more than 4 times of the training set's 1319.667.

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1320 on 27983 degrees of freedom
## Multiple R-squared:  0.9168, Adjusted R-squared:  0.9167
## F-statistic: 1.284e+04 on 24 and 27983 DF,  p-value: < 2.2e-16

load.lm.step.pred <- predict(load.lm.step.both,data = valid)
accuracy(load.lm.step.both$fitted.values, train$LOAD)

##              ME      RMSE      MAE      MPE      MAPE
## Test set 3.112739e-14 1319.667 1050.074 -0.2141934 3.716788

accuracy(load.lm.step.pred, valid$LOAD)

##              ME      RMSE      MAE      MPE      MAPE
## Test set -37.15714 6395.828 5186.479 -2.845234 18.67568
```

The automatic stepwise guides us to remove some variables that are not statistically significant and may have some bad impact on the model, like #PRICE.

```
Coefficients:
(Intercept) Estimate Std. Error t value Pr(>|t|)
g1          1.38489    0.16341    8.47 < 0.0000000000000002 ***
g2          0.90176    0.03887   23.20 < 0.0000000000000002 ***
g3          0.76193    0.00543  140.20 < 0.0000000000000002 ***
g4          0.68505    0.00814   84.16 < 0.0000000000000002 ***
g5          7.46838    0.21164   35.29 < 0.0000000000000002 ***
g6         -1.33205    0.01356  -98.26 < 0.0000000000000002 ***
g7          0.83964    0.03555   23.62 < 0.0000000000000002 ***
g8          1.00503    0.00717  140.19 < 0.0000000000000002 ***
g9          0.74183    0.01012   73.27 < 0.0000000000000002 ***
g10         1.03867    0.55414    1.87    0.061 .
g11        22.80719    0.93967   24.27 < 0.0000000000000002 ***
g12         0.79798    0.00756  105.58 < 0.0000000000000002 ***
g13         5.33084    0.23836   22.36 < 0.0000000000000002 ***
g14         0.71516    0.00352  203.32 < 0.0000000000000002 ***
tmp        16.31987    1.68652    9.68 < 0.0000000000000002 ***
prs        18.91031    1.10055   17.18 < 0.0000000000000002 ***
hmd         6.63698    0.98780    6.72    0.000000000019 ***
w_s        32.94721    8.10353    4.07    0.000048001242 ***
w_d        -0.71491    0.15947   -4.48    0.000007390355 ***
s_3       -183.40609   86.09082   -2.13    0.033 *
t1         -2.67409    1.33848   -2.00    0.046 *
t2        -541.39453   26.83017  -20.18 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1360 on 28028 degrees of freedom
Multiple R-squared:  0.912, Adjusted R-squared:  0.912
F-statistic: 1.33e+04 on 22 and 28028 DF,  p-value: <0.0000000000000002

              ME RMSE MAE MPE MAPE
Test set 0.0000000000000045 1355 1058 -0.23 3.7
```

There is also concern that when we include more variables, the  $R^2$  will not be so powerfully persuasive than before as there is a natural growth in the value.

Our solution is, without sacrificing the exactness of our model, we would manually control the potential variables no more than 20. According to the exhaustive-method table (*attached on the next page*), the variables as #PRICE, #rain in past 1 hour, #rain in past 3 hours, #Clouds Coverage Rate are deducted from the list. From a real world sense, except for PRICE, other variables may better be dummy variables from past experiences, and standardized if necessary.

There is another reason to leave out #PRICE. The endogeneity will happen. When price changes, demand changes accordingly. Vice Versa. Even though we are exploring a prediction model, we cannot abandon its function of explaining the phenomenon.

```

## 1 subsets of each size up to 23
## Selection Algorithm: exhaustive
##
##      g1  g2  g3  g4  g5  g6  g7  g8  g9  g10 g11 g12 g13 g14 PRICE tmp prs  hmd w_s w_d r_1 r_3 s_3 cld t1 t2
## 1  ( 1 ) " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 2  ( 1 ) " " " " "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 3  ( 1 ) " " " " "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 4  ( 1 ) " " " " "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 5  ( 1 ) " " " " "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 6  ( 1 ) " " " " "*" "*" " " " " "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 7  ( 1 ) " " " " "*" "*" " " " " "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 8  ( 1 ) " " " " "*" "*" " " " " "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 9  ( 1 ) " " " " "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 10 ( 1 ) " " " " "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 11 ( 1 ) " " " " "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 12 ( 1 ) " " " " "*" "*" "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 13 ( 1 ) " " "*" "*" "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 14 ( 1 ) " " "*" "*" "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 15 ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 16 ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 17 ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 18 ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 19 ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 20 ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 21 ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 22 ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "
## 23 ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " "

```

Table 2 Exhaustive-method Generated Tab

The final regression model is like below. The coefficients might be a little bit different as the train and validation sets are generated randomly:

Dependent Variable	Independent Variables
LOAD (Demand)	g1 - generation.biomass g2 - generation.fossil.brown.coal.lignite g3 - generation.fossil.gas g4 - generation.fossil.hard.coal g5 - generation.fossil.oil g6 - generation.hydro.pumped.storage.consumption g7 - generation.hydro.run.of.river.and.poundage g8 - generation.hydro.water.reservoir g9 - generation.nuclear g11 - generation.other.renewable g12 - generation.solar g13 - generation.waste g14 - generation.wind.onshore
	tmp - Temperature prs - Pressure hmd - Humidity w_s - Wind Speed w_d - Wind Direction s_3 - snow in past 3 hours t2

Table 3 *Original Best Model*

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-21745.30340	1375.96102	-15.80	< 0.0000000000000002	***
g1	1.51133	0.14848	10.18	< 0.0000000000000002	***
g2	0.89855	0.03885	23.13	< 0.0000000000000002	***
g3	0.76125	0.00543	140.32	< 0.0000000000000002	***
g4	0.68645	0.00811	84.60	< 0.0000000000000002	***
g5	7.50603	0.21071	35.62	< 0.0000000000000002	***
g6	-1.33029	0.01353	-98.31	< 0.0000000000000002	***
g7	0.83732	0.03551	23.58	< 0.0000000000000002	***
g8	1.00646	0.00714	140.89	< 0.0000000000000002	***
g9	0.74336	0.01010	73.57	< 0.0000000000000002	***
g11	22.83026	0.93962	24.30	< 0.0000000000000002	***
g12	0.79785	0.00756	105.55	< 0.0000000000000002	***
g13	5.27239	0.23682	22.26	< 0.0000000000000002	***
g14	0.71552	0.00351	203.92	< 0.0000000000000002	***
tmp	16.68613	1.67499	9.96	< 0.0000000000000002	***
prs	18.80694	1.09905	17.11	< 0.0000000000000002	***
hmd	7.20196	0.95745	7.52	0.0000000000000056	***
w_s	36.97396	7.77632	4.75	0.000001997444182	***
w_d	-0.71858	0.15943	-4.51	0.000006598594940	***
s_3	-185.66912	86.09290	-2.16	0.031	*
t2	-551.26089	26.56026	-20.76	< 0.0000000000000002	***

All the generation ways except g6 (hydro pumped storage consumption) have a positive impact on the demand. For example, a unit increase in g11(generation renewable) will have a 22.83 increase in the demand. This might be because the daily composition of electricity usage in Spain relies highly on this kind of power generation method. Similarly, when temperature, air pressure, humidity, and wind speed go up, load follows. Meanwhile, when there is snow in the past three hours, load shrink. For example, a unit change in temperature will cause a 16.68 movement in demand, which means the demand for electricity can be reflected by temperature change. It is easy to understand that people need air-conditioner when it is hot or cold. We can draw inferences from the instances explained. Lastly, when it is in the night, people tend to demand much less, a 555.26 down in demand for electricity. It makes every sense because people tend to use less power during inactive periods.

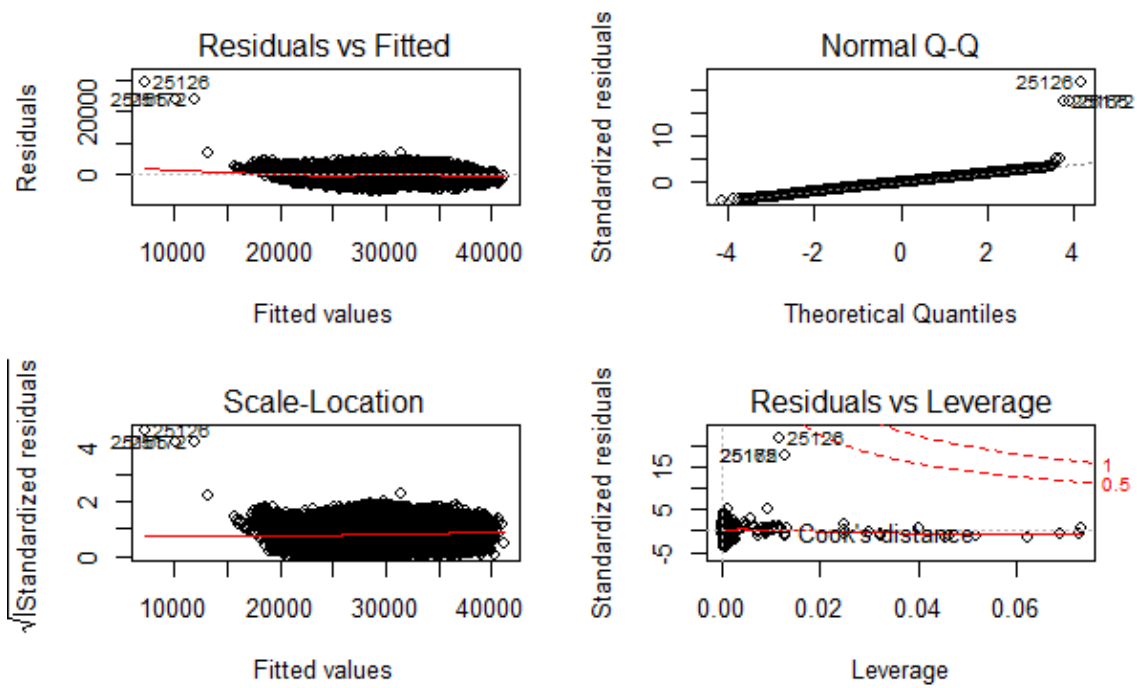
The new adjusted  $R^2$  is 0.912, slightly lower than before but still at a very high level from experiences. Meanwhile, we now get a model that plays seemingly much better in prediction as the new RMSE is reduced to 1288. But it is not the truth.

```
Residual standard error: 1360 on 28030 degrees of freedom
Multiple R-squared:  0.912,    Adjusted R-squared:  0.912
F-statistic: 1.46e+04 on 20 and 28030 DF,  p-value: <0.0000000000000002
```

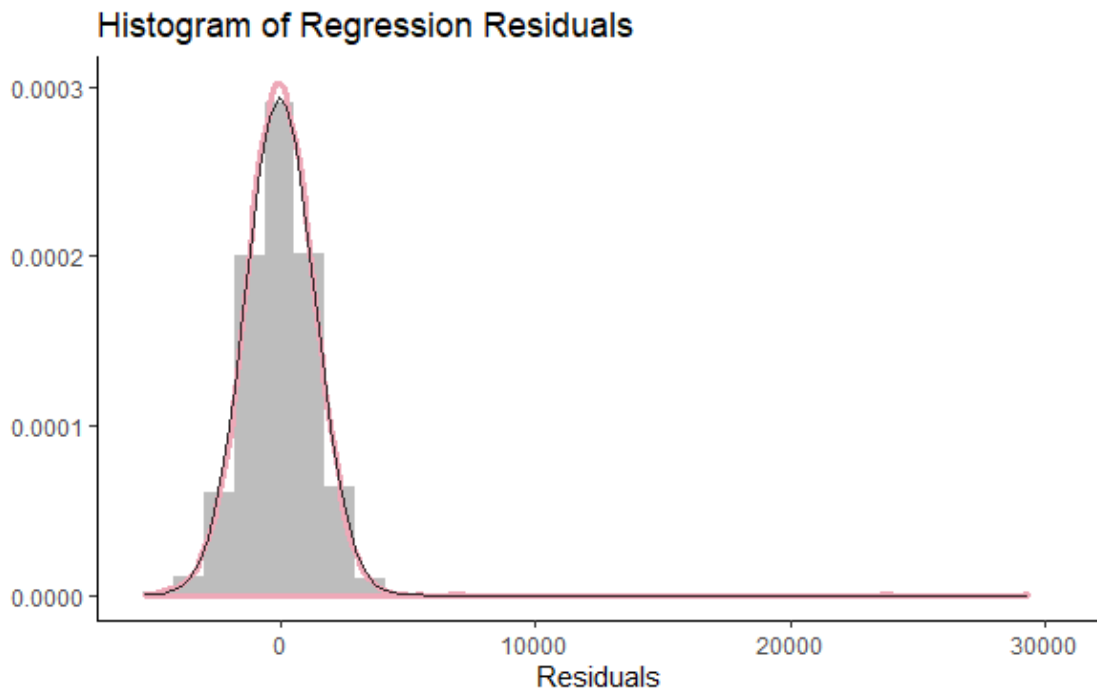
```

              ME RMSE  MAE   MPE MAPE
Test set 0.000000000000001 1355 1058 -0.23  3.7
              ME RMSE  MAE   MPE MAPE
Test set 14 1288 1021 -0.17  3.6
```

We would like to assess the accuracy of our model and whether it will be having perfect multicollinearity.



Plot 7 *Residual Combo of Original Model (With Outliers)*



Plot 8 *Residual histogram vs. Normal Distribution Line (With outliers)*

g1	g2	g3	g4	g5	g6	g7	g8	g9	g11	g12	g13	g14	tmp	prs	hmd	w_s	w_d
2.5	2.9	2.2	3.9	1.9	1.8	3.1	2.6	1.1	2.7	2.5	2.2	1.9	2.3	1.2	3.1	1.7	1.3
s_3	t2																
1.0	2.7																

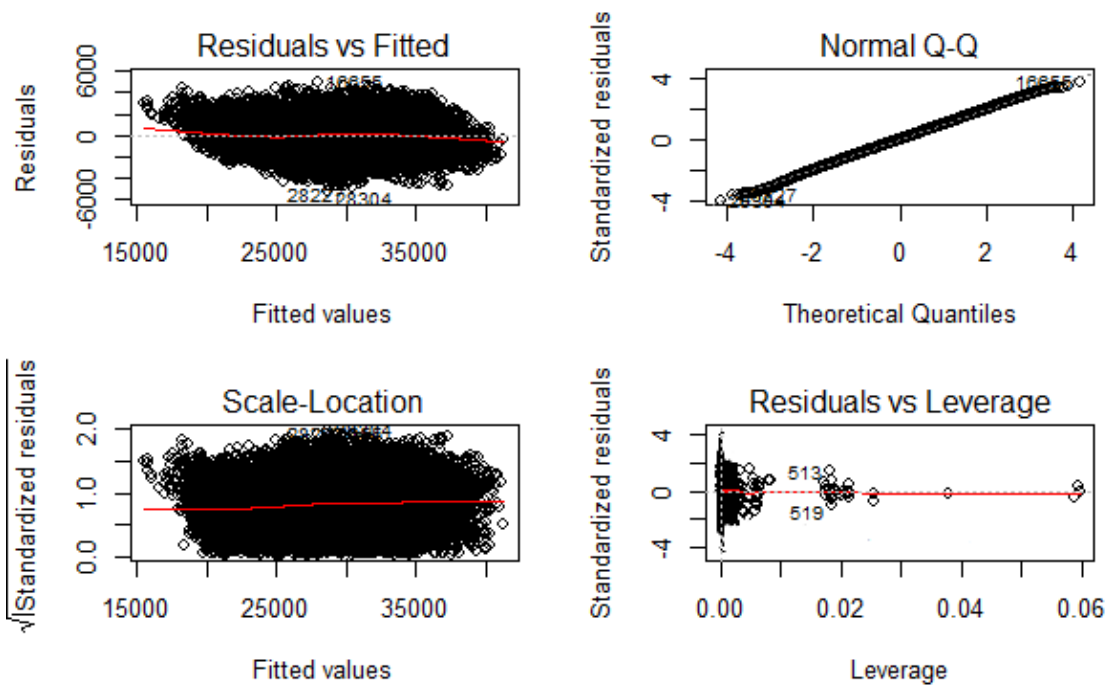


Table 4 *Perfect Multicollinearity of Original Best Model*

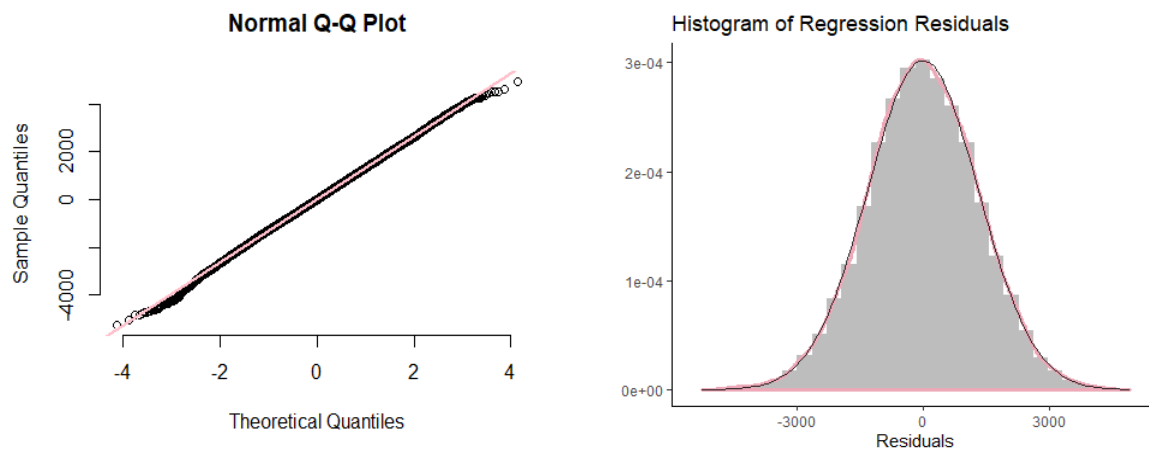
We could see many outliers existing in the observations. So we delete them gradually. The good news is that the model is not suffering from perfect multicollinearity as all values are far under 5 in vif call.

After removing the outlier, we get another set of the residual plots and reports. Also we attached the original best model when we first knit this .Rmd file.

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.361e+04  1.344e+03 -17.568  < 2e-16 ***
## g1           2.020e+00  1.463e-01  13.801  < 2e-16 ***
## g2           8.580e-01  3.784e-02  22.676  < 2e-16 ***
## g3           7.510e-01  5.270e-03 142.509  < 2e-16 ***
## g4           6.965e-01  7.924e-03  87.899  < 2e-16 ***
## g5           7.845e+00  2.049e-01  38.286  < 2e-16 ***
## g6          -1.316e+00  1.329e-02 -99.032  < 2e-16 ***
## g7           8.848e-01  3.470e-02  25.499  < 2e-16 ***
## g8           1.010e+00  6.987e-03 144.608  < 2e-16 ***
## g9           7.709e-01  9.915e-03  77.745  < 2e-16 ***
## g11          2.569e+01  9.204e-01  27.906  < 2e-16 ***
## g12          8.103e-01  7.418e-03 109.225  < 2e-16 ***
## g13          5.704e+00  2.312e-01  24.675  < 2e-16 ***
## g14          7.223e-01  3.426e-03 210.804  < 2e-16 ***
## tmp          1.909e+01  1.636e+00  11.666  < 2e-16 ***
## prs          1.895e+01  1.072e+00  17.670  < 2e-16 ***
## hmd          9.137e+00  9.325e-01   9.798  < 2e-16 ***
## w_s          3.393e+01  7.603e+00   4.463 8.11e-06 ***
## w_d         -8.298e-01  1.551e-01  -5.350 8.88e-08 ***
## s_3          1.009e+03  2.302e+02   4.384 1.17e-05 ***
## t2          -5.559e+02  2.591e+01 -21.456  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```



Plot 9 *Residual Combo of Original Model (Without Outliers)*



Left \_ Plot 10 *QQ Plot of Original Model (Without Outliers)*

Right \_ Plot 11 *Residual histogram vs. Normal Distribution Line (Without outliers)*

We removed around 50 outliers. Compared to the 35,000+ observations, it is only 0.14%. As there's no cook's distance and the x-scale are very low, we think the margin of continue removing outliers would not be so beneficial. The reason why there is still leverage point is probably because of the dataset itself has a highly internal volatility or maybe the model still needs other modifications and improvements to predict as more than a multiple regression model.

Now we are eligible to say the model works fine. The plot at the left corner of the combo shows a randomly scattered residuals and an almost flat line parallel with line  $y=0$ . The QQ Plot indicates that the dots fit so well with line  $y=x$ . Besides, the Residuals vs. Leverage plot reflects points within a normal range and a fitted line almost overlap with  $y=0$ . Lastly, we see a residual trend line in pink and a normal distribution line in black that fit perfectly.

## Part IV Regression with Polynomial Term

Under the same logical flow, we attempt to add some polynomial term into the regression. The base lies where we find a curving trend line in the scatterplot above. Under the instruction of the document, we do not replace the original variable with a polynomial one, so the variance inflation factor (vif) call may well lose some authorities.

The new adjusted  $R^2$  is slightly higher, but as is reasoned above, it is not so reliable. The RMSE is a little bit lower, not so obvious. Our thinking is that these variables, despite the curve trend, do not necessarily has the attribution of a higher-order term. That is, we are not able to capture a trend when the two ends are high/low and the middle is low/high, not to mention that we can find ways to give practical meanings to them.

And the dataset also does not allow us to match some interactive terms.

These index changes cannot support a better model with polynomial term theory.

```
load.ip <- lm(Load ~ . - PRICE - g10 - r_1 - r_3 - cld - t1 + I(g8^2) + I(g3^2)
+ I(g6^2), data = train)
load.ip.pred <- predict(load.ip, valid)
summary(load.ip)

##
## Call:
## lm(formula = Load ~ . - PRICE - g10 - r_1 - r_3 - cld - t1 +
##      I(g8^2) + I(g3^2) + I(g6^2), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5444.3  -843.1    -3.8    858.4   4692.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.149e+04  1.305e+03 -16.467  < 2e-16 ***
## g1           2.162e+00  1.421e-01  15.218  < 2e-16 ***
## g2           8.160e-01  3.668e-02  22.248  < 2e-16 ***
```

```

## g3          1.232e+00  1.910e-02  64.497 < 2e-16 ***
## g4          6.284e-01  7.855e-03  79.996 < 2e-16 ***
## g5          7.238e+00  2.003e-01  36.141 < 2e-16 ***
## g6         -2.075e+00  3.282e-02 -63.213 < 2e-16 ***
## g7          8.518e-01  3.432e-02  24.818 < 2e-16 ***
## g8          1.129e+00  1.684e-02  67.052 < 2e-16 ***
## g9          7.547e-01  9.618e-03  78.474 < 2e-16 ***
## g11         2.093e+01  8.999e-01  23.259 < 2e-16 ***
## g12         7.926e-01  7.215e-03 109.849 < 2e-16 ***
## g13         5.191e+00  2.250e-01  23.067 < 2e-16 ***
## g14         7.212e-01  3.356e-03 214.914 < 2e-16 ***
## tmp         1.197e+01  1.595e+00   7.503 6.42e-14 ***
## prs         1.865e+01  1.041e+00  17.921 < 2e-16 ***
## hmd         7.019e+00  9.053e-01   7.754 9.23e-15 ***
## w_s         3.221e+01  7.368e+00   4.371 1.24e-05 ***
## w_d        -8.218e-01  1.503e-01  -5.466 4.63e-08 ***
## s_3         8.651e+02  2.231e+02   3.877 0.000106 ***
## t2         -5.726e+02  2.514e+01 -22.778 < 2e-16 ***
## I(g8^2)     -2.230e-05  2.080e-06 -10.718 < 2e-16 ***
## I(g3^2)     -3.078e-05  1.122e-06 -27.429 < 2e-16 ***
## I(g6^2)     2.858e-04  1.034e-05  27.641 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1280 on 27984 degrees of freedom
## Multiple R-squared:  0.9218, Adjusted R-squared:  0.9217
## F-statistic: 1.434e+04 on 23 and 27984 DF,  p-value: < 2.2e-16

vif(load.ip)

##          g1          g2          g3          g4          g5          g6          g7
g8
##  2.495293  2.885266 30.589877  4.051990  1.887949 11.496833  3.230031
16.204181
##          g9          g11          g12          g13          g14          tmp          prs
hmd

```

```
## 1.110705 2.729591 2.508959 2.171948 1.987219 2.280147 1.239515
3.073958
##          w_s          w_d          s_3          t2    I(g8^2)    I(g3^2)    I(g6^2)
## 1.677384 1.265505 1.028158 2.701547 12.630485 24.658884 9.049004

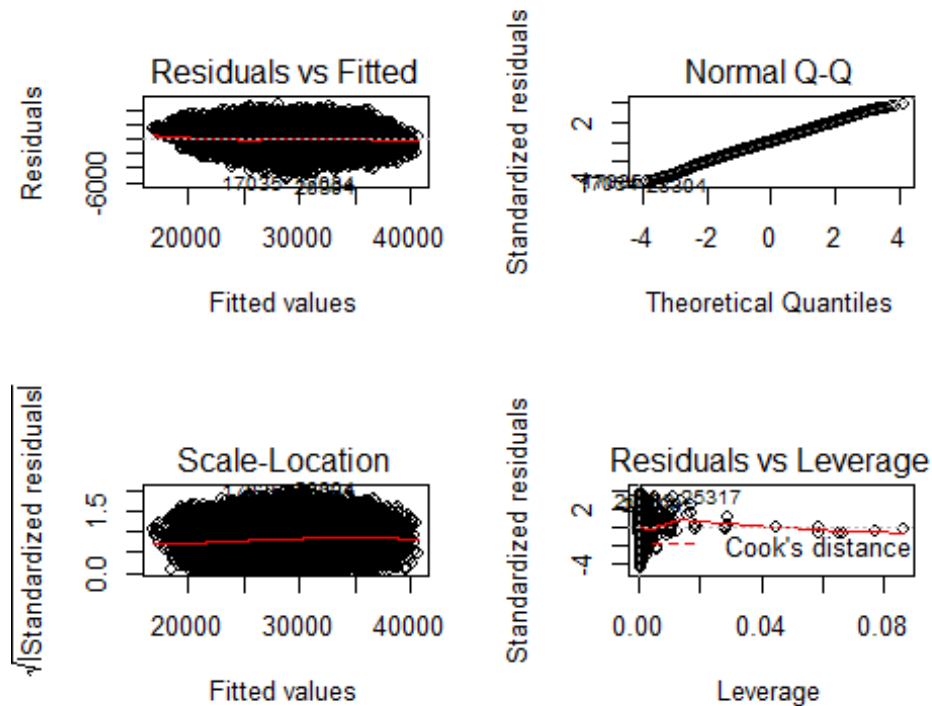
accuracy(load.ip$fitted.values, train$LOAD)

##              ME      RMSE      MAE      MPE      MAPE
## Test set 7.052061e-15 1279.129 1012.844 -0.1994319 3.584798

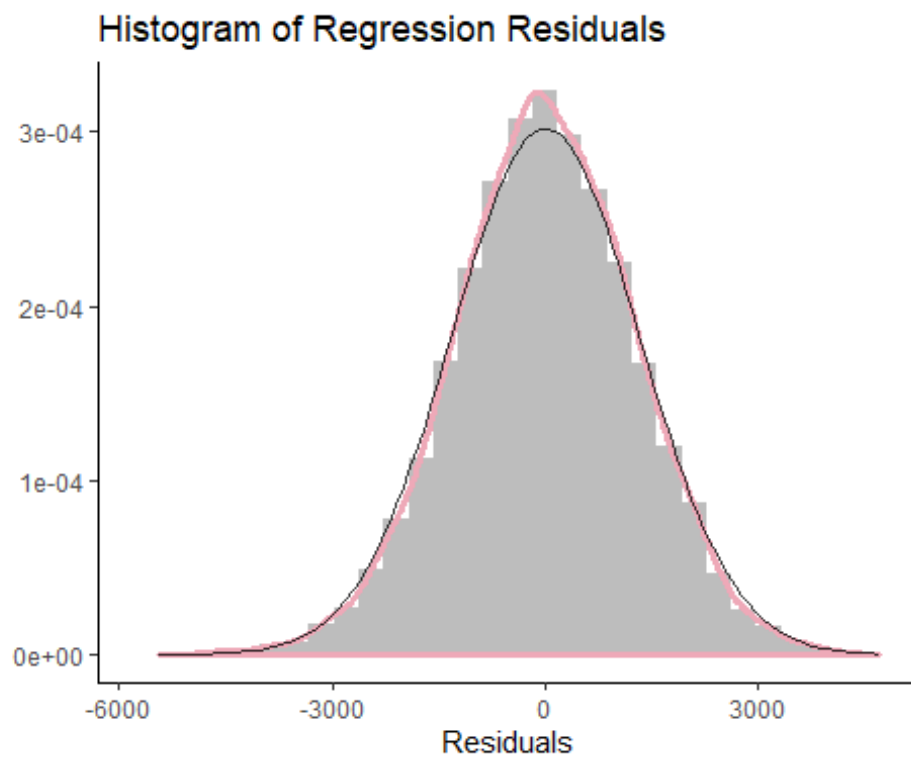
accuracy(load.ip.pred, valid$LOAD)

##              ME      RMSE      MAE      MPE      MAPE
## Test set 1.183412 1255.173 989.8391 -0.1814663 3.50437

par(mfrow=c(2,2))
plot(load.ip)
```



Plot 12 *Residual Combo of Original Model (With Polynomial Term)*



Plot 13 *Residual histogram vs. Normal Distribution Line (With Polynomial Term)*