MULTILINEAR REGRESSION MODEL TO PREDICT ENERGY GENERATION FROM TURBINE

Sahil Nasa

AIM:

Apply O.L.S in real life situations (Multiple regression of at least 5 variables). Check autocorrelation, multi correlation and Heteroscedasticity.

INTRODUCTION:

The dataset contains 36733 instances of 11 sensor measures aggregated over one hour (by means of average or sum) from a gas turbine located in Turkey's north western region for the purpose of studying flue gas emissions, namely CO and NOx (NO + NO2). The data comes from the same power plant as the dataset used for predicting hourly net energy yield. By contrast, this data is collected in another data range (01.01.2011 - 31.12.2015), includes gas turbine parameters (such as Turbine Inlet Temperature and Compressor Discharge pressure) in addition to the ambient variables. Note that the dates are not given in the instances but the data are sorted in chronological order. See the attribute information and relevant paper for details. Kindly follow the protocol mentioned in the paper (using the first three years' data for training/ cross-validation and the last two for testing) for reproducibility and comparability of works. The dataset can be well used for predicting turbine energy yield (TEY) using ambient variables as features.

Independent Variables:

1. Ambient temperature (AT) /X1
2. Ambient pressure (AP) mbar /X2
3. Ambient humidity (AH) (%) /X3
4. Air filter difference pressure (AFDP) /X4
5. Gas turbine exhaust pressure (GTEP) mbar /X5
6. Turbine inlet temperature (TIT) C /X6
7. Turbine after temperature (TAT) C /X7
8. Compressor discharge pressure (CDP) mbar /X8
9. Carbon monoxide (CO) mg/m3 /X9
10. Nitrogen oxides (NOx) mg/m3 /X10

Dependent Variable:

1. Turbine energy yield (TEY) MWH /Y: Total Energy yielded from the mixture produced by NO(x) and CO.

Dataset:

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
2	116.27	1.9532	1020.1	84.985	2.5304	20.116	1048.7	544.92	10.799	7.4491	113.25
3	109.18	1.2191	1020.1	87.523	2.3937	18.584	1045.5	548.5	10.347	6.4684	112.02
4	125.88	0.94915	1022.2	78.335	2.7789	22.264	1068.8	549.95	11.256	3.6335	88.147
5	132.21	1.0075	1021.7	76.942	2.817	23.358	1075.2	549.63	11.702	3.1972	87.078
6	133.58	1.2858	1021.6	76.732	2.8377	23.483	1076.2	549.68	11.737	2.3833	82.515
7	133.58	1.8319	1021.7	76.411	2.841	23.495	1076.4	549.92	11.829	2.0812	81.193
8	131.53	2.074	1022	75.974	2.7981	22.945	1073.7	549.98	11.687	2.2529	83.171
9	133.18	1.7824	1022.6	73.535	2.8327	23.337	1075.7	550.01	11.745	3.735	85.749
10	135.38	1.593	1023.2	72.873	2.8729	23.654	1078.5	550.06	11.772	3.6398	86.491
11	134.86	1.6819	1023.8	72.441	2.9058	23.463	1077.9	550.12	11.742	3.5866	86.328
12	134.98	1.9002	1024.5	71.376	2.9126	23.562	1078.2	550.12	11.77	3.5605	84.117
13	134.21	1.7797	1025.1	68.528	2.8725	23.276	1077	550.03	11.782	3.6902	85.317
14	134.44	1.1722	1025.3	65.626	2.85	23.215	1077.3	550.09	11.873	3.1766	86.431
15	133.82	0.72327	1025.4	64.393	2.8395	23.101	1076.2	550.15	11.797	2.562	84.708
16	135.36	0.48348	1025.8	64.144	2.8746	23.299	1078.2	549.96	11.897	2.1854	84.104
17	134.47	0.42953	1025.9	64.504	2.866	23.069	1077	549.73	11.805	2.4286	83.869
18	115.92	0.48238	1026.1	63.519	2.5481	20.185	1037.1	535.4	10.86	12.659	118.27
19	111.2	0.059447	1026	52.124	2.4651	18.478	1048.4	548.78	10.384	4.6896	104.56
20	115.46	-0.21598	1026.7	51.978	2.5301	19.438	1050.4	546.71	10.767	7.0983	116.96

CLEANED AND WORKING DATA

METHODOLOGY:

• The cleaned and working data set was taken from UCI Machine Learning Repository. It is a secondary source of data and statical concepts of multiple linear regression was used. A multiple linear model was fitted taking Y as a dependent variable and the Xi's (i=1 to 10) as explanatory variables.

• The model is further tested for multicollinearity, heteroscedasticity and auto correlation and treated accordingly.

• The entire project and the statistical tests in it are carried using the R software.

ECONOMETRIC ANALYSIS:

1. FITING THE MODEL:

Taking Y as dependent variable, a multiple linear regression model was fitted and the Xi's (i=1 to 7) as explanatory variables .

$$Y=B0+B1X1+B2X2+B3X3+B4X4+B5X5+B6X6+B7X7+B8X8+B9X9+B10X10+U$$

Where U is the disturbance term. Bi's are the ith parameter associated with explanatory variable Xi.

<u>The fitted Model is:</u>

```
> df=read.csv(file.choose())
> head(df)
       Y      X1     X2     X3     X4     X5     X6     X7     X8     X9     X10
1 116.27 1.95320 1020.1 84.985 2.5304 20.116 1048.7 544.92 10.799 7.4491 113.250
2 109.18 1.21910 1020.1 87.523 2.3937 18.584 1045.5 548.50 10.347 6.4684 112.020
3 125.88 0.94915 1022.2 78.335 2.7789 22.264 1068.8 549.95 11.256 3.6335  88.147
4 132.21 1.00750 1021.7 76.942 2.8170 23.358 1075.2 549.63 11.702 3.1972  87.078
5 133.58 1.28580 1021.6 76.732 2.8377 23.483 1076.2 549.68 11.737 2.3833  82.515
6 133.58 1.83190 1021.7 76.411 2.8410 23.495 1076.4 549.92 11.829 2.0812  81.193
> X= cbind(df$X1,df$X2,df$X3,df$X4,df$X5,df$X6,df$X7,df$X8,df$X9,df$X10)
> Y= df$Y
> Model1= lm(Y~X)
> Model1

Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)          X1           X2           X3           X4           X5           X6           X7
 -183.12571    -0.29975     -0.05815     -0.01569     -0.85491     -0.02519      0.68088     -0.67006
         X8          X9          X10
    1.58265     0.05408     -0.02678
```

ANOVA OF THE MODEL:

```
> anova(Model1)
Analysis of Variance Table

Response: Y
            Df  Sum Sq Mean Sq F value    Pr(>F)
X           10 1929632  192963  475479 < 2.2e-16 ***
Residuals 7373    2992       0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

HYPOTHESES TESTING

H0 : B0=B1=B2=B3=B4=B5=B6=B7=0

H1 : At least one of Bi is not equal to zero. (i=0 to 10)

The F statistics obtained from ANOVA is 475479 with it's p value being less than 0.05. Thus, taking the level of significance at 5%, we are able to reject the null hypothesis and conclude that at least one of Bi's is not equal to zero.

Significance of the parameters:

```
> summary(Model1)

Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6594 -0.3333  0.0181  0.3531  2.4134 

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)    
(Intercept) -1.831e+02  2.192e+00  -83.532  < 2e-16 ***
X1          -2.998e-01  2.294e-03 -130.682  < 2e-16 ***
X2          -5.815e-02  1.383e-03  -42.047  < 2e-16 ***
X3          -1.569e-02  7.816e-04  -20.072  < 2e-16 ***
X4          -8.549e-01  5.420e-02  -15.773  < 2e-16 ***
X5          -2.519e-02  5.361e-03   -4.698 2.67e-06 ***
X6           6.809e-01  7.638e-03   89.147  < 2e-16 ***
X7          -6.701e-01  1.118e-02  -59.959  < 2e-16 ***
X8           1.583e+00  1.573e-01   10.062  < 2e-16 ***
X9           5.408e-02  6.438e-03    8.400  < 2e-16 ***
X10         -2.678e-02  1.177e-03  -22.753  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.637 on 7373 degrees of freedom
Multiple R-squared:  0.9985,	Adjusted R-squared:  0.9984 
F-statistic: 4.755e+05 on 10 and 7373 DF,  p-value: < 2.2e-16
```

HYPOTHESES TESTING

H0 : The model is not of significant fit. (R2=0)

H1 : The model is of significant fit. (R2≠0)

Adjusted R2 value obtained is 0.9984 which indicates a very good fit. The corresponding p value is less than 0.05. Thus taking significance level at 5%, we are able to reject H0 and conclude the model is of significant fit.

- **Checking for the presence of Multicollinearity in the model**

```
> imcdiag(Model1)

Call:
imcdiag(mod = Model1)


All Individual Multicollinearity Diagnostics Result

      VIF    TOL        Wi        Fi Leamer    CVIF Klein  IND1   IND2
X1 2.0517 0.4874  9186.307 11023.779 0.6981 -2.7462     0 1e-04 1.0849
X2 1.4648 0.6827  4060.258  4872.403 0.8262 -1.9607     0 1e-04 0.6716
X3 1.3258 0.7542  2846.109  3415.396 0.8685 -1.7746     0 1e-04 0.5201
X4 1.9531 0.5120  8325.151  9990.372 0.7155 -2.6142     0 1e-04 1.0328
X5 1.4885 0.6718  4266.962  5120.451 0.8196 -1.9924     0 1e-04 0.6946
X6 3.5015 0.2856 21850.188 26220.726 0.5344 -4.6868     1 0e+00 1.5120
X7 3.3472 0.2988 20502.202 24603.111 0.5466 -4.4802     1 0e+00 1.4841

1 --> COLLINEARITY is detected by the test 
0 --> COLLINEARITY is not detected by the test

* all coefficients have significant t-ratios

R-square of y on all x: 0.6431 
```

- Since VIF for the variables are greater than 5, we consider that Multicollinearity is present.

We can check the correlation matrix, it is clear that X8 has

```
Correlation Matrix
            X1          X2          X3          X4          X5          X6          X7          X8
X1   1.0000000 -0.49309788 -0.46628847  0.46897582  0.19357758  0.33011162  0.20827660  0.20090892
X2  -0.4930979  1.00000000  0.08438144 -0.09414429 -0.04373034 -0.08160451 -0.29014662  0.02942009
X3  -0.4662885  0.08438144  1.00000000 -0.24545608 -0.29770831 -0.26068269  0.02625087 -0.22170633
X4   0.4689758 -0.09414429 -0.24545608  1.00000000  0.84395757  0.91512777 -0.51980671  0.92299064
X5   0.1935776 -0.04373034 -0.29770831  0.84395757  1.00000000  0.89285131 -0.62065201  0.93814162
X6   0.3301116 -0.08160451 -0.26068269  0.91512777  0.89285131  1.00000000 -0.39616119  0.95159003
X7   0.2082766 -0.29014662  0.02625087 -0.51980671 -0.62065201 -0.39616119  1.00000000 -0.65661298
X8   0.2009089  0.02942009 -0.22170633  0.92299064  0.93814162  0.95159003 -0.65661298  1.00000000
X9  -0.3906467  0.20094462  0.15899855 -0.64078864 -0.55717729 -0.73809227  0.02576820 -0.61265262
X10 -0.5935802  0.21423632  0.06535073 -0.58445186 -0.36665515 -0.52008080  0.05445541 -0.44309256
            X9         X10
X1  -0.3906467 -0.59358018
X2   0.2009446  0.21423632
X3   0.1589986  0.06535073
X4  -0.6407886 -0.58445186
X5  -0.5571773 -0.36665515
X6  -0.7380923 -0.52008080
X7   0.0257682  0.05445541
X8  -0.6126526 -0.44309256
X9   1.0000000  0.67839402
X10  0.6783940  1.00000000

====================NOTE====================

X4 and X5 may be collinear as |0.843958|>=0.7
X4 and X6 may be collinear as |0.915128|>=0.7
X5 and X6 may be collinear as |0.892851|>=0.7
X4 and X8 may be collinear as |0.922991|>=0.7
X5 and X8 may be collinear as |0.938142|>=0.7
X6 and X8 may be collinear as |0.951590|>=0.7
X6 and X9 may be collinear as |-0.738092|>=0.7
```

REVISED AND IMPROVED MODEL:

```
> X= cbind(df$X1,df$X2,df$X3,df$X5,df$X7,df$X9,df$X10)
> Y= df$Y
> Model2= lm(Y~X)
> Model2

Call:
lm(formula = Y ~ X)

Coefficients:
(Intercept)          X1          X2          X3          X4          X5          X6          X7
  153.05839    -0.09116     0.14773     0.06102     2.65268    -0.42923    -1.70865    -0.01481
```

ANOVA technique on the revised model

```
> anova(Model2)
Analysis of Variance Table

Response: Y
            Df  Sum Sq Mean Sq F value    Pr(>F)
X            7 1767802  252543   11302 < 2.2e-16 ***
Residuals 7376  164823      22
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Summary of the given model:

```
> summary(Model2)

Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.940  -1.724   0.751   2.676  60.180 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 153.058393  15.485586   9.884  < 2e-16 ***
X1           -0.091159   0.012192  -7.477 8.49e-14 ***
X2            0.147727   0.009860  14.983  < 2e-16 ***
X3            0.061021   0.005359  11.387  < 2e-16 ***
X4            2.652684   0.023739 111.744  < 2e-16 ***
X5           -0.429228   0.016618 -25.829  < 2e-16 ***
X6           -1.708654   0.043007 -39.729  < 2e-16 ***
X7           -0.014806   0.008725  -1.697   0.0898 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.727 on 7376 degrees of freedom
Multiple R-squared:  0.9147,	Adjusted R-squared:  0.9146 
F-statistic: 1.13e+04 on 7 and 7376 DF,  p-value: < 2.2e-16
```

2. **Checking for the presence of Auto-Correlation in the model:**
   We use Durbin Watson test to check the presence of Auto-Correlation.

```
> lmtest::dwtest(Model2)

	Durbin-Watson test

data:  Model2
DW = 0.43922, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

- **Hypotheses testing**

  H0: there is no presence of autocorrelation

  H1: there is presence of autocorrelation

We see that the obtained value of DW statistic is 0.43922 which indicative of positive autocorrelation. Furthermore, the p value being less than 0.05. Therefore, taking level of significance at 5 %, we are able to reject H0 and conclude that there is autocorrelation present in the model.

Removal of Autocorrelation from the Model:

Cochran Orcutt iterative method is being used for estimating parameters under autocorrelation.

```
> library(orcutt)
> Model3= orcutt::cochrane.orcutt(Model2)
> Model3
Cochrane-orcutt estimation for first order autocorrelation 
 
Call:
lm(formula = Y ~ X)

 number of interaction: 22
 rho 0.954797

Durbin-Watson statistic 
(original):    0.43922 , p-value: 0e+00
(transformed): 1.69285 , p-value: 3.963e-40
 
 coefficients: 
(Intercept)          X1          X2          X3          X4          X5          X6          X7 
-450.338873   -0.669587    0.313311    0.032135    3.905519    0.325359   -0.196805   -0.058278 
```

Revised Model M4 is fitted with Cochran Orcutt iterative procedure.

```
> summary(Model3)
Call:
lm(formula = Y ~ X)

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept) -4.5034e+02  4.9412e+01  -9.114 < 2.2e-16 ***
X1          -6.6959e-01  3.5463e-02 -18.881 < 2.2e-16 ***
X2           3.1331e-01  4.7935e-02   6.536 6.735e-11 ***
X3           3.2135e-02  8.6007e-03   3.736 0.0001882 ***
X4           3.9055e+00  1.7567e-02 222.319 < 2.2e-16 ***
X5           3.2536e-01  1.1591e-02  28.069 < 2.2e-16 ***
X6          -1.9681e-01  1.9051e-02 -10.331 < 2.2e-16 ***
X7          -5.8278e-02  5.9416e-03  -9.808 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.1032 on 7381 degrees of freedom
Multiple R-squared:  0.9359 ,  Adjusted R-squared:  0.9359
F-statistic: 15381 on 1 and 7381 DF,  p-value: < 0e+00

Durbin-Watson statistic
(original):    0.43922 , p-value: 0e+00
(transformed): 1.69285 , p-value: 3.963e-40
```

Hypotheses testing

H0: all the ai’s are equal to zero (i=0,1,2,3,5,7,9,10)

H1: at least one of the ai’s is not zero.

We have adjusted R2 for the model M4 as 0.9359, and the F value is 15381. The corresponding p value is less than 0.05. Thus, taking level of significance at 5%, we are able to reject H0 and conclude that at least one the ai’s is not zero.

3. **Checking for the presence of Heteroscedasticity in the Model:**

The Model 2 is further checked for presence of heteroscedasticity in the error variance. **Goldfield Quandt test** is used for the purpose.

```
> lmtest::gqtest(Model2)

        Goldfeld-Quandt test

data:  Model2
GQ = 0.016957, df1 = 3684, df2 = 3684, p-value = 1
alternative hypothesis: variance increases from segment 1 to 2
```

HYPOTHESES TESTING

H0: There is no presence of heteroscedasticity in the error variance.

H1: There is presence of heteroscedasticity in the error variance.

The GQ value, which follows F distribution, is 0.016957 and the calculated p value> 0.05, and thus we are able to accept H0 and conclude that there is no significant heteroscedasticity present in the error variance.

RESULT:

- The Model shows the presence of Multicollinearity and Autocorrelation. Heteroscedasticity is absent in the model.
- Model1: $\hat{Y}$= -183.12571 -0.669587X1 + 0.313311X2 + 0.032135X3 + 3.905519X4 -0.02519X5 +0.68088X6 -0.67006X7 + 1.58265X8 + 0.05408X9 - 0.02678X10

- Model2(after removal of autocorrelation) $\hat{Y}$= -450.338873+ 1.268642X1+ 4.291919X2+ 0.032135X3 + 3.905519X5 + 0.325359X7 + 0.196805X9 -0.058278X10