Regression Analysis

Group 16

Project objective:

To apply regression models to analyse a given data set

Data Set: QSAR fish toxicity

Data Set Characteristics:	Multivariate	**Number of Instances:**	908	**Area:**	Physical
Attribute Characteristics:	Real	**Number of Attributes:**	7	**Date Donated**	2019-09-23
Associated Tasks:	Regression	**Missing Values?**	N/A	**Number of Web Hits:**	40630

Understanding the Data set

Data Set information :

Was used to develop quantitative regression to predict acute aquatic toxicity towards the fish Pimephales promelas (fathead minnow) on a set of 908 chemicals.

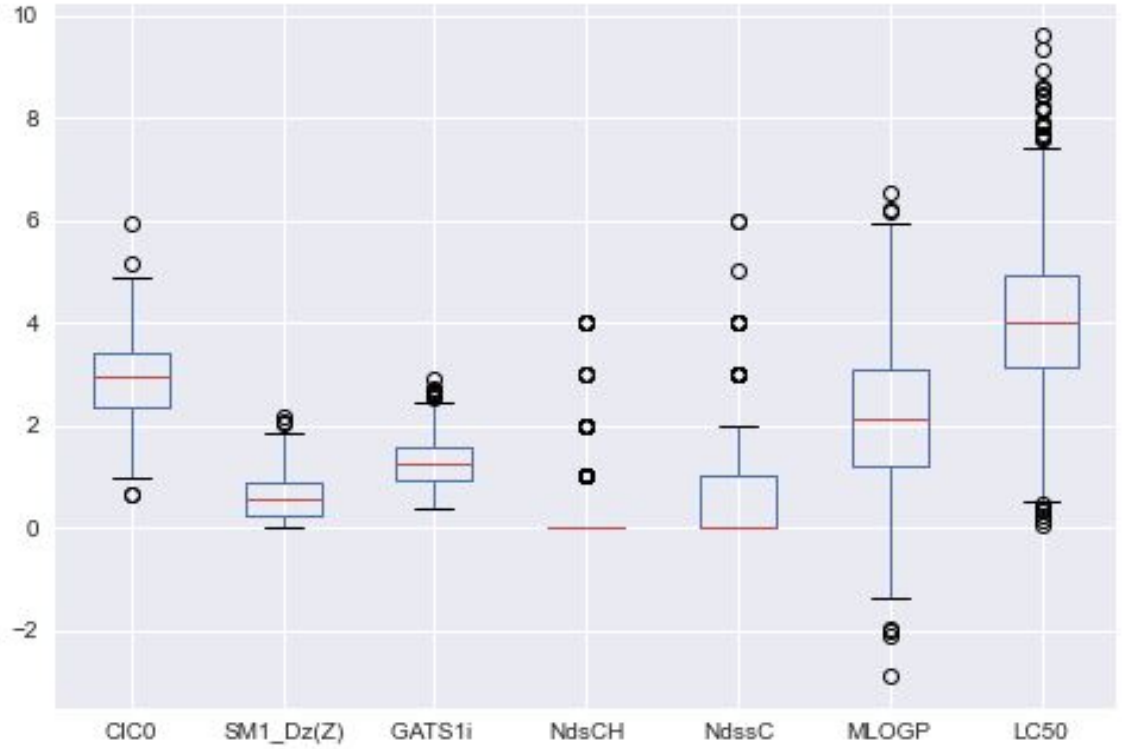
LC50 data, the conc. causing death in 50% of test fish over a test duration of 96 hrs, was used as model response.

Attribute Information:

Contains values for 6 attributes (molecular descriptors) of 908 chemicals used to predict quantitative acute aquatic toxicity towards the fish Pimephales promelas (fathead minnow).

Variable plots

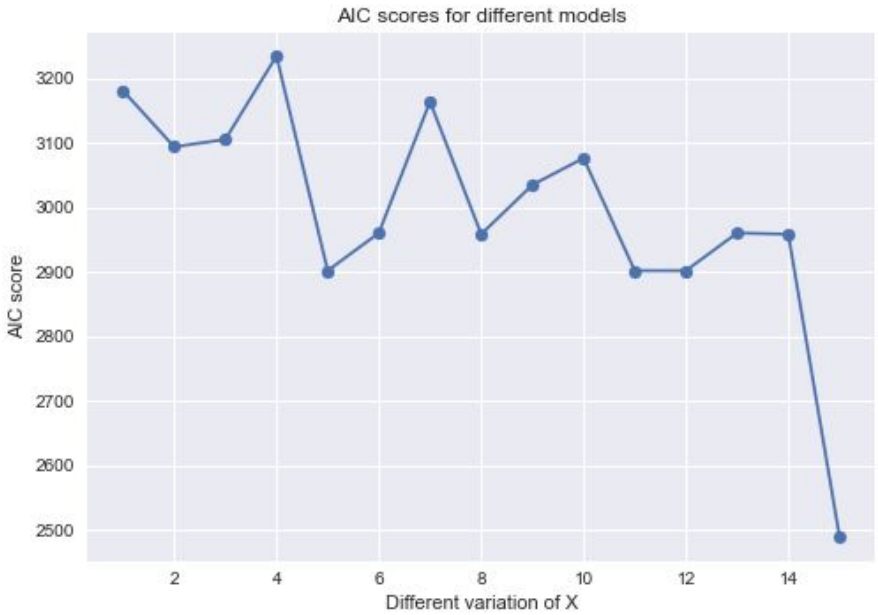
Box plot of the variables:



Forward Selection plots:

AIC

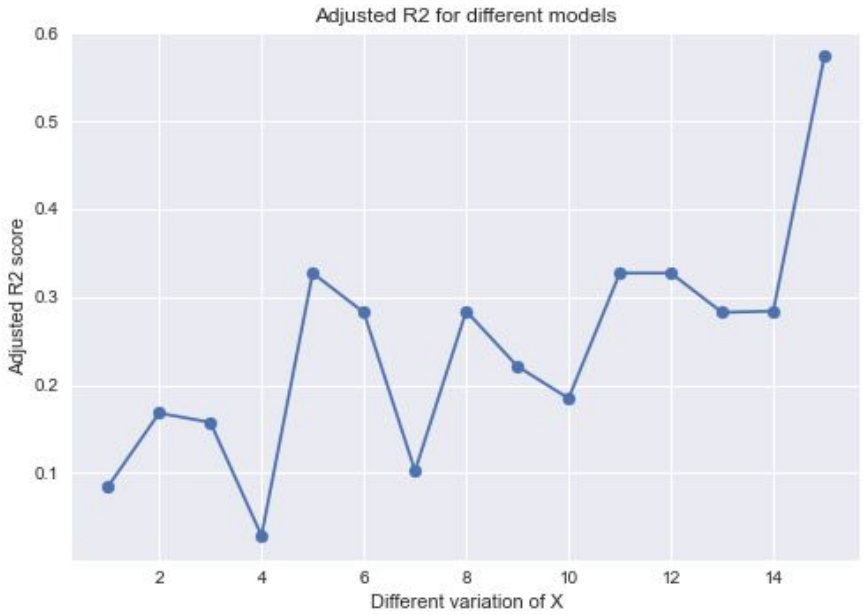
- Minima : 2490.095



Forward Selection plot:

Adjusted R2

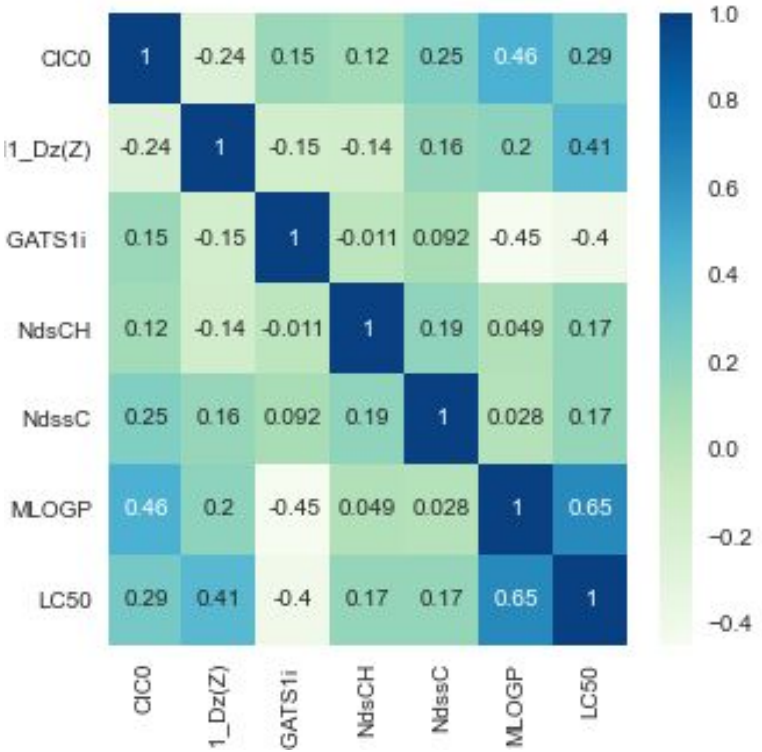
- Maxima : 0.5743



Tests run:

- **Multicollinearity check**
  - VIF Values
- **Autocorrelation check**
  - Durbin-Watson Test
- **Normality check**
  - Shapiro-Wilk Test
- **Homoscedasticity check**
  - Goldfeld-Quandt Test

Multicollinearity Check

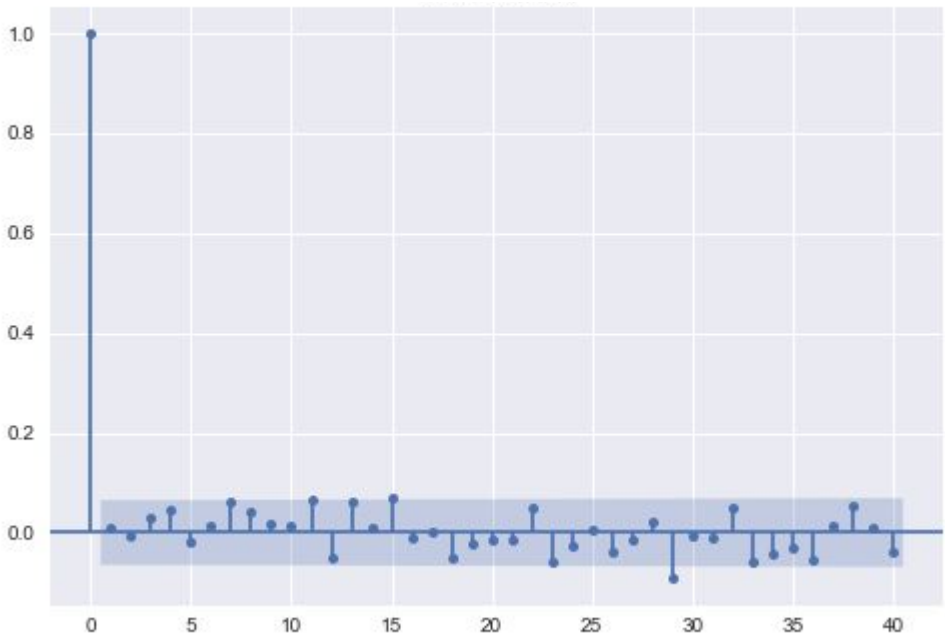


	Features	VIF Factor
0	ClC0	25.598575
1	SM1_Dz(Z)	3.464257
2	GATS1i	14.480146
3	NdsCH	1.220603
4	NdssC	1.544874
5	MLOGP	7.428264

Removing the columns

	Features	VIF Factor
0	SM1_Dz(Z)	2.489933
1	NdsCH	1.184511
2	NdssC	1.406001
3	MLOGP	2.405152

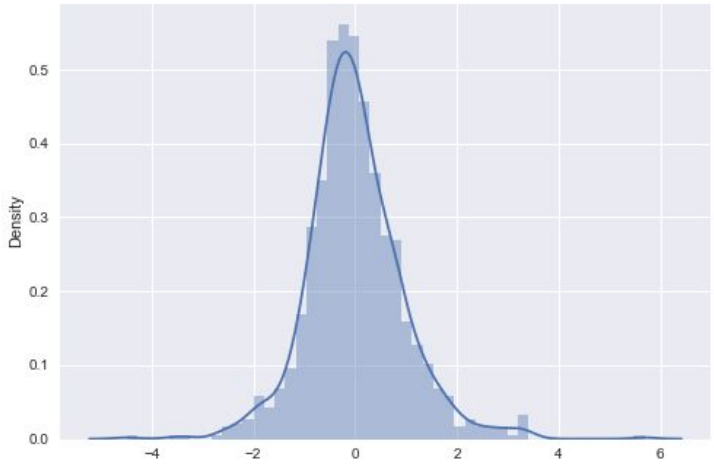
Autocorrelation Check

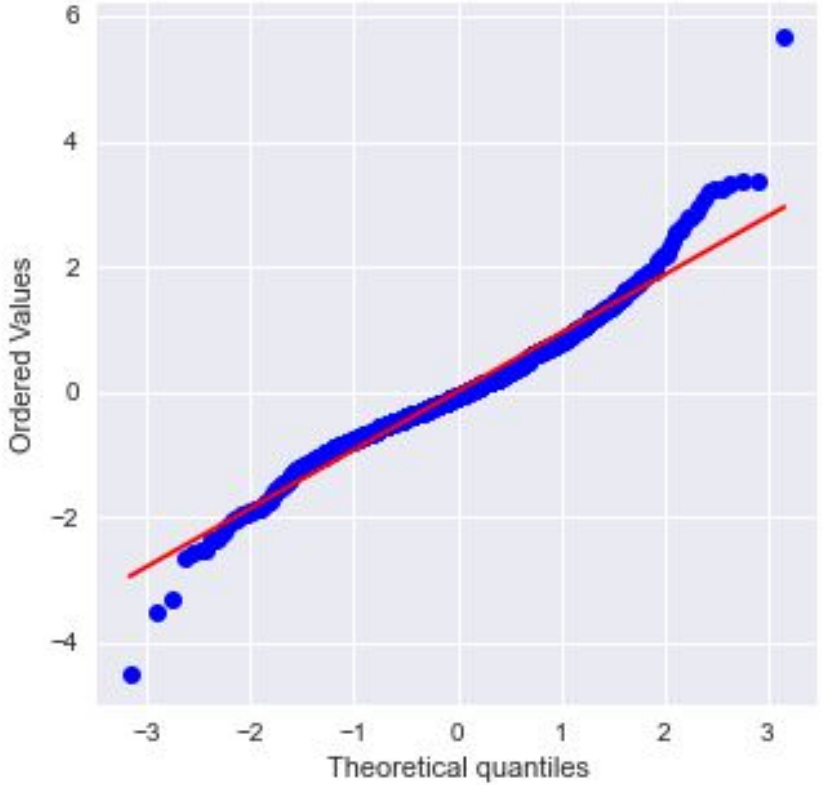


Durbin-Watson Test Result: 1.9785048067212314

Normality check

It's a normal distribution!!!

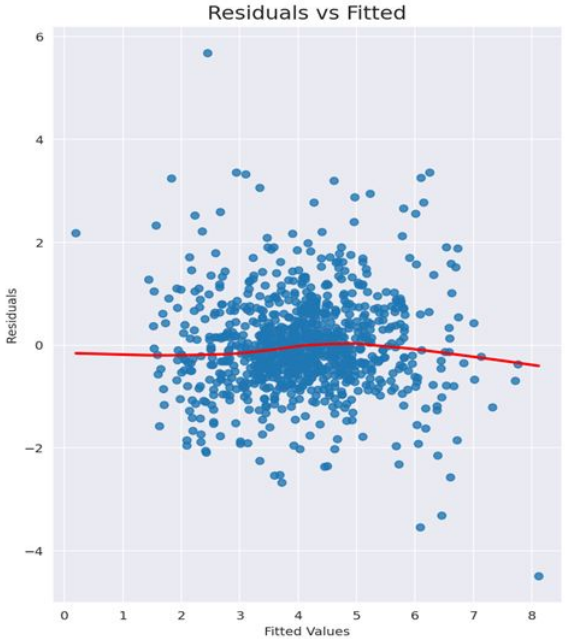


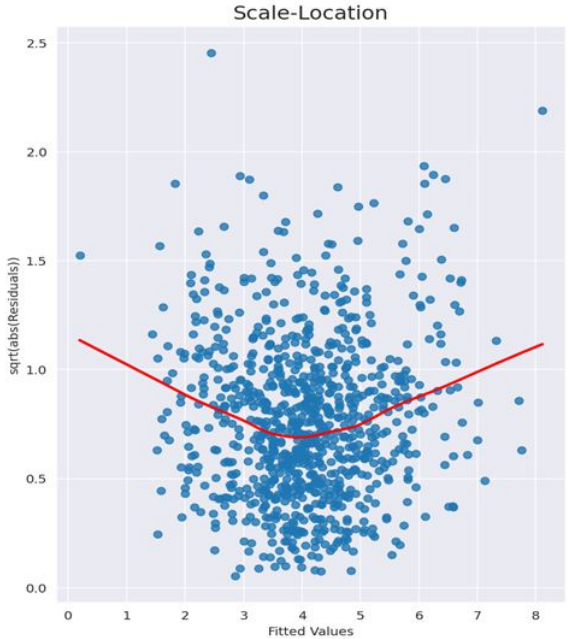


Shapiro-Wilk test ----
statistic: 0.9630, p-value: 0.0512

Homoscedasticity test

- Since the p-value is above 5%, so the null hypothesis is True.
- Thus the data is indeed Homoscedastic.
- Variance remains more or less constant.





Goldfeld-Quandt test ----

	value
F statistic	0.664172
p-value	0.999992

P value 0.99

Results:

- End points for both plots are reached when these 4 columns are taken into account : [SM1_Dz(Z) , NdsCH , NdssC , MLOGP]

  SM1_Dz(Z) -> 1.2556

  NdsCH -> 0.4136

  NdssC -> 0.0643

  MLOGP -> 0.3901

- If Multicollinearity is taken into account, a significant drop in AIC score is observed.
- Taking the columns to be independant gives better result.

References :

1. UCI Machine Learning repository - Dataset