PSTAT126 Project

Linear Regression Analysis on QSAR Fish Toxicity

Shengjia Yu&Zheren Dong

12/11/2019

Introduction

Abstract

In this project, we will explore the relationship of different chemical factors associated with acute aquatic toxicity towards the fish Pimephales promelas (fathead minnow) in QSAR fish toxicity Data Set provided by UCI machine learning repository. Linear regression is the base method we will use statistically, and we will find the best fitted model by using F−tests and Akaike's Information Criterion. Then we will solve the research question along the way of diagnosing a regression model and finding the assoication between the predictors and the response.

Overview of Data

The QSAR fish toxicity Data Set data set contains 908 observations. The description of each variables in the data set are the following:

Response:

LC50 data: the concentration that causes death in 50% of test fish over a test duration of 96 hours

Molecular Descriptors (Predictors):

MLOGP −− molecular properties: expresses the lipophilicity of a molecule, this being the driving force of narcosis

CIC0 −− information indices: encode information regarding heteroatoms

GATS1i −− 2D autocorrelations: encodes information on molecular polarisability

NdssC −− atom−type counts: number of unsaturated sp^2 carbon atoms of the type =C<

NdsCH −− atom−type counts: number of unsaturated sp^2 carbon atoms of the type =CH−

SM1_Dz(Z) −− 2D matrix−based descriptors: the sum of the eigenvalues of the Barysz matrix, whose elements take into account information on both the bond order and the atomic number

Questions of Interest

We consider the following research questions:

- Question 1: Is the regression model containing at least one predictor useful in predicting the concentration of aquatic toxicity towards the fish Pimephales promelas?
- Question 2: What mean response LC50 value do we expect with NdssC and NdsCH having value of 0, CIC0 having value of 3, and with average values of the other predictors?
- Question 3: Does the effect of CIC0 on LC50 depend on any other predictors? It means if with interaction terms involving CIC0 is statistically significant since according to research we find it is a significant factor in its effect on the biological question in general.

Regression Methods

In order to answer the research questions, we firstly need to find an appropriate linear regression model with picked predictors. We firstly remove some variables of multicollinearity to form a regression model. Then we may use transformation to fulfill "LINE" conditions. We may use F—tests and Akaike's Information Criterion to detect the variable for our final model. At last, we may check if there are any influential points.

While doing the above steps, we may use the following method to solve our interested questions:

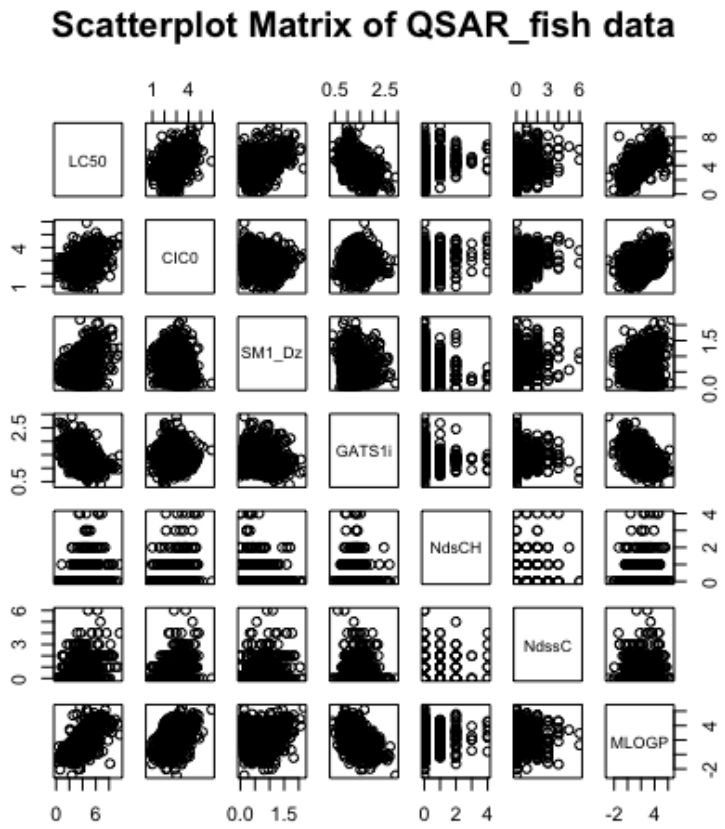
- For Question 1: The research question reduces to testing the hypothesis: H0: $\beta 1 = \beta 2 = \beta 3 = \beta 4 = \beta 5 = \beta 6 = 0$ vs H1 : At least one βk != 0 (k = 0, 1, 2, 3, 4, 5, 6) In this case, we are interested in testing that all six slope parameters are zero. We'll show that the null hypothesis is tested using the analysis of variance F—test.
- For Question 2: We can calculate a 95 percent prediction interval for a new value of the response given the specified values of the predictors.
- For Question 3: It is not hard to solve. What we are trying to see is that given the model with interaction terms involving NdssC, if our final model still significant without interaction terms with CIC0. We should build a model with any potential terms including CIC0 after we determine the interaction term by using the graph of Residuals vs. interaction term. We are only interested in continuous variable's interaction with CIC0 since discrete variables may not generate useful information as the least useful variable and it definitely shows some pattern with continuous variables.

Regression Analysis, Results and Interpretation

Building the model

Removal of Multicollinearity

We firstly plot the scatterplot matrix of the response and all potential predictors:



From the graph, we can see that there seems to be positive linear relationships between LC50 and the predictors CIC0, SM1_Dz and MLOGP; we also have negative linear relationships between LC50 and the predictors GATS1i. For the left two predictors, it seems that there is not a substantial relationship between each of them and the response.

If we use cor() function to check linear associations among the predictors, we didn't see there appears to be a strong linear associations between any of the two since the maximum correlation coefficient is 0.46386714.

	CIC0	SM1_Dz	GATS1i	NdsCH	NdssC	MLOGP
CIC0	1.0000000	-0.2353605	0.14762196	0.12134073	0.24663904	0.46386714
SM1_Dz	-0.2353605	1.0000000	-0.14596719	-0.14140138	0.16317892	0.20066284
GATS1i	0.1476220	-0.1459672	1.00000000	-0.01065656	0.09240977	-0.45073916
NdsCH	0.1213407	-0.1414014	-0.01065656	1.00000000	0.18816360	0.04861995
NdssC	0.2466390	0.1631789	0.09240977	0.18816360	1.00000000	0.02849947
MLOGP	0.4638671	0.2006628	-0.45073916	0.04861995	0.02849947	1.00000000

Using vif() to check the variance Variance Inflation Factor, we can see all of them smaller than 5. Thus, there is no need to clean multicollinearity in the predictors. Thus, we currently have all of the six predictors on the model.

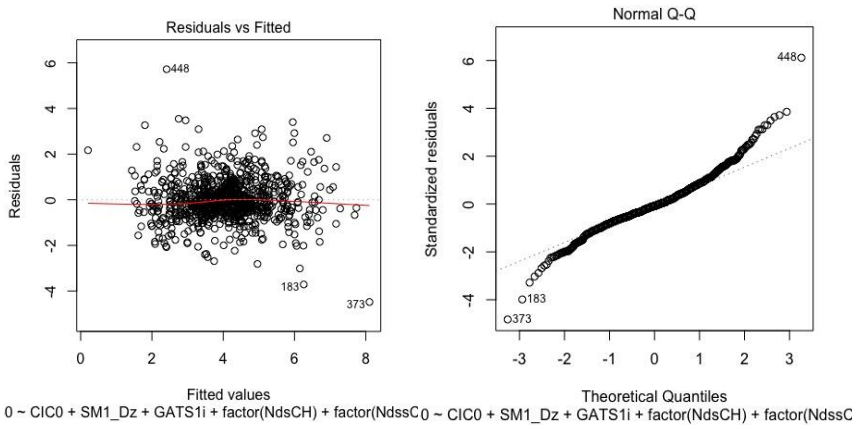
CIC0	SM1_Dz	GATS1i	NdsCH	NdssC	MLOGP
2.131459	1.397732	1.605793	1.078629	1.231891	2.353669

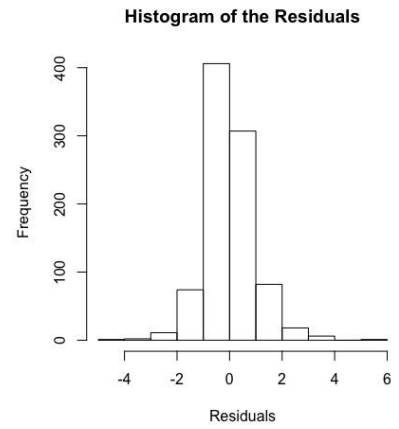
LINE Requirement Checking and Data Transformations

We are going to check the LINE conditions now. Current model is:

$$E(LC50_i) = \beta 0 + \beta 1 * MLOGP_i + \beta 2 * CIC0_i + \beta 3 * GATS1i_i + \beta 4 * NdssC_i + \beta 5 * NdsCH_i + \beta 6 * SM1Dz_i$$

Now we use the residual vs. plotted, normal Q−Q plot, and histogram to check linearity, equal variances, and normality of error terms. Note: Since we do not have any information about time serial information, we will not deal with the independence of error terms here due to the minor effect that temporal dependence may have





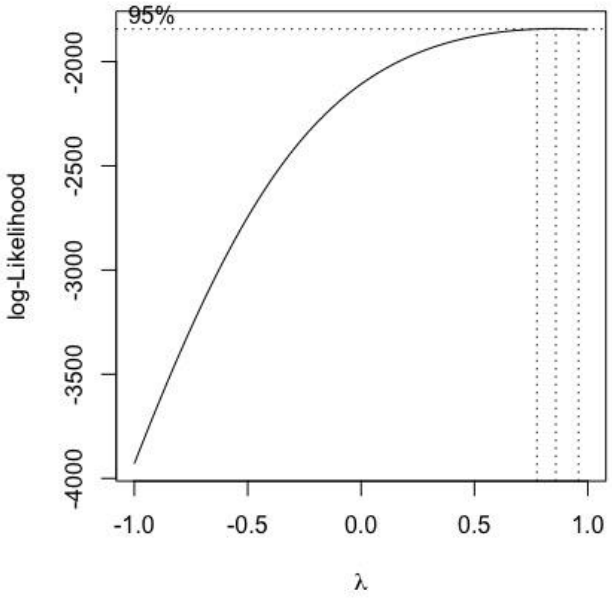
.

We can see that we do not have an obvious non−linearity and non−constant variance pattern from the first plot.

Thus, we now need to deal with non−Normality of the error terms proven by the Normal Q−Q plot and histogram of the residuals. It seems that we are having a "Heavy−Tailed" situation.

Thus, we need to transform the response variable. Here, we are using a Box−Cox plot to determine transformation on response:

Zero is not in the 95 percent confidence interval for the maximum likelihood estimate of lambda, so we use the result from boxcox(0.8585859) to transform. Now we have a final model satisfied "LINE" conditions. The plots now are as following:



The current model after transformation is still:

$$E(LC50_i) = \beta 0 + \beta 1 * MLOGP_i + \beta 2 * CIC0_i + \beta 3 * GATS1i_i + \beta 4 * NdssC_i + \beta 5 * NdsCH_i + \beta 6 * SM1Dz_i$$

Variable Selection

We use two different method to approach the selection of predictors in our model(See detailed code from Appendix). From Stepwise Regression F−test with a significant level of 0.05, the following model is yielded:

$$E(LC50_i) = \beta 0 + \beta 1 * MLOGP_i + \beta 2 * CIC0_i + \beta 3 * GATS1i_i + \beta 4 * NdsCH_i + \beta 5 * SM1Dz_i$$

However, using the AIC will give us the full model:

$$E(LC50_i) = \beta 0 + \beta 1 * MLOGP_i + \beta 2 * CIC0_i + \beta 3 * GATS1i_i + \beta 4 * NdssC_i + \beta 5 * NdsCH_i + \beta 6 * SM1Dz_i$$

Both two models lead to closely similar result. However, while doing Stepwise Regression using F−test, we learned that according to its p−value,$NdsCH_i$ is not eligible for entry into our stepwise model. Therefore, in the following analysis and tests, we we prefer to use the model obtained from Stepwise Regression F−test, which is:

$$E(LC50_i) = \beta 0 + \beta 1 * MLOGP_i + \beta 2 * CIC0_i + \beta 3 * GATS1i_i + \beta 4 * NdsCH_i + \beta 5 * SM1Dz_i$$

Checking Influential Points

Having the model we obtained above, We are not satisfied with the correlation, so we decide to improve it with adjusted data since there were too many influential points. We use Externally Studentized Residuals to identify the outliers and DFFITS to identify the high influential points. We delete the common parts we found in DFFITS and outliers, then we evaluate the model over and over again. Note that we have 908 total instances, deleting several outliers won't significantly impact the entire model.

After deleting all potential the outliers' and high leverage points' effects, we will achieve the following result for the model (See detailed code below on Appendix).

```
Call:
lm(formula = LC50_d2^(0.8585859) ~ MLOGP_d2 + SM1_Dz_d2 + factor(NdsCH_d2) +
    GATS1i_d2 + CIC0_d2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8462 -0.3314 -0.0068  0.3254  1.5806

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.57333    0.10519  14.958  < 2e-16 ***
MLOGP_d2           0.32973    0.01959  16.831  < 2e-16 ***
SM1_Dz_d2          0.90315    0.04843  18.648  < 2e-16 ***
factor(NdsCH_d2)1  0.45836    0.05621   8.155 1.27e-15 ***
factor(NdsCH_d2)2  0.61916    0.10593   5.845 7.25e-09 ***
factor(NdsCH_d2)3  0.49545    0.26088   1.899  0.05789 .
factor(NdsCH_d2)4  0.91660    0.30355   3.020  0.00261 **
GATS1i_d2         -0.43380    0.05868  -7.392 3.50e-13 ***
CIC0_d2            0.31603    0.03424   9.229  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5185 on 838 degrees of freedom
Multiple R-squared:  0.7037,    Adjusted R-squared:  0.7009
F-statistic: 248.8 on 8 and 838 DF,  p-value: < 2.2e-16
```

The assumptions of linearity and constant variance seem to be satisfied. There is a slight “heavy tail” condition, but this effect was similar to the one we originally obtained from the uneditted dataset, which is small and thus it may not have major effect on the assumption of Normality. From the output shown above, we can find Coefficient of Determination is R^2 = 0.7009, which means that about 70.09 percent of the variation in response LC50 is explained by taking into account the variation of the predictors in our model.

Research Questions Answers

Question1:

We need the full model first. The full model is the largest possible model – that is, the model containing all of the possible predictors. $LC50_i = \beta0 + \beta1 * MLOGP_i + \beta2 * CIC0_i + \beta3 * GATS1i_i + \beta4 * NdsCH_i + \beta5 * SM1Dz_i$ The error sum of squares for the full model is SSE(F). Because there are 7 parameters in the full model, df = n – 7. We also need the reduced model. The reduced model is the model that the null hypothesis describes. Because the null hypothesis sets each of the slope parameters in the full model equal to 0, the reduced model is: $LC50_i = \beta0$ The reduced model basically suggests that none of the variation in the response Y is explained by any of

the predictors. The error sum of squares for the reduced model is SSE(R). Because there is only one parameter in the reduced model, df = n − 1. We get the result from Anova function call.

```
Analysis of Variance Table

Model 1: LC50_d2^(0.8585859) ~ 1
Model 2: LC50_d2^(0.8585859) ~ MLOGP_d2 + SM1_Dz_d2 + factor(NdsCH_d2) +
    GATS1i_d2 + CIC0_d2
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    846 760.29
2    838 225.28  8    535.01 248.76 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There is sufficient evidence to conclude that at least one of the slope parameters is not equal to 0.

Question2:

We solve this by calculating a 95 confidence prediction interval for mean response with predictor values NdsCH = 0, NdssC = 0, CIC0 = 3, average values of MLOGP and SM1_Dz(Z). This interval is given by

```
        fit      lwr      upr
1 3.222931 3.184017 3.261846
```

We then convert to original units of the response using the exponential and have the result:

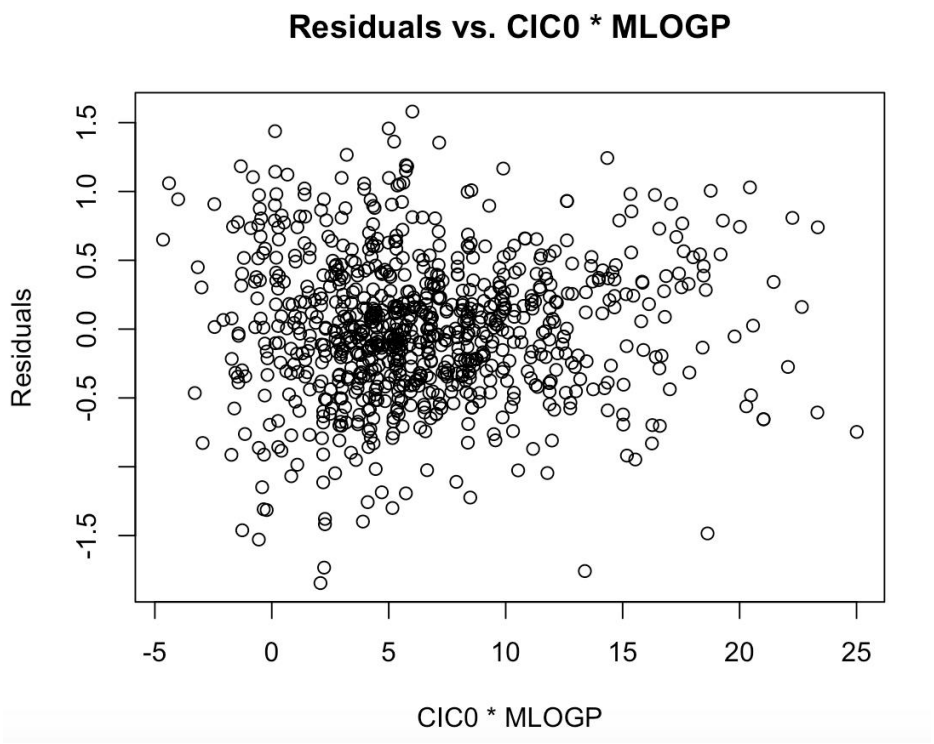
```
        fit      lwr      upr
1 3.908076 3.853171 3.963089
```

Therefore, we are 95% confident that given NdsCH = 0, NdssC = 0, CIC0 = 3, average values of MLOGP and SM1_Dz(Z), the response would be between (3.853171, 3.963089).

Question3:

We used the final model built after deleting outliers and high leverage data points as the original model; then we add interaction terms including CIC0 and other continuous variables.

We firstly add one interaction term CIC0 * MLOGP to our full_model as our new model; then using the residual vs. CIC0 * MLOGP plot and F−test to evaluate its significance.



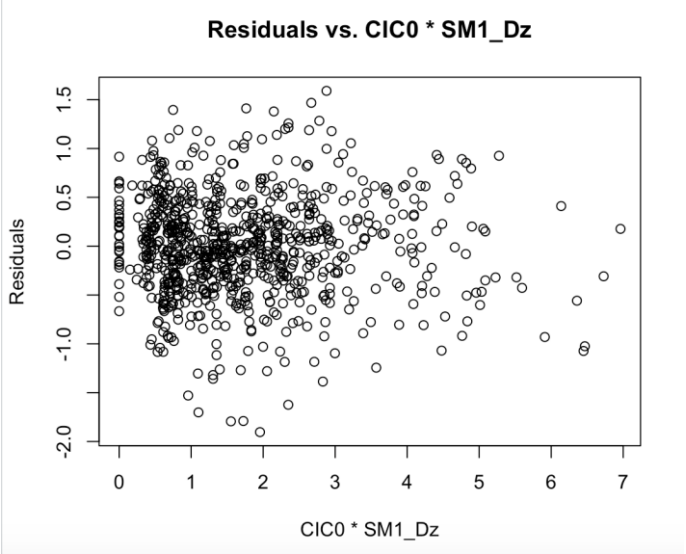

```
Analysis of Variance Table

Model 1: LC50_d3^(0.8585859) ~ MLOGP_d3 + SM1_Dz_d3 + factor(NdsCH_d3) +
    GATS1i_d3 + CIC0_d3
Model 2: LC50_d3^(0.8585859) ~ MLOGP_d3 + SM1_Dz_d3 + factor(NdsCH_d3) +
    GATS1i_d3 + CIC0_d3 + MLOGP_d3:CIC0_d3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    835 225.13
2    834 221.22  1    3.9181 14.772 0.0001306 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As we can see, the f−test gives a smaller p−value than 0.5 and the graph shows a somewhat pattern, so we conclude that there is significance if including this interaction term with CIC0.

Then we are interested in adding another interaction term with other two predictors left by comparing the results from two possible updates: one is with (CIC0 * SM1_Dz) & (CIC0 * MLOGP) and the other is with (CIC0 * MLOGP) and (CIC0 * GATS1i).

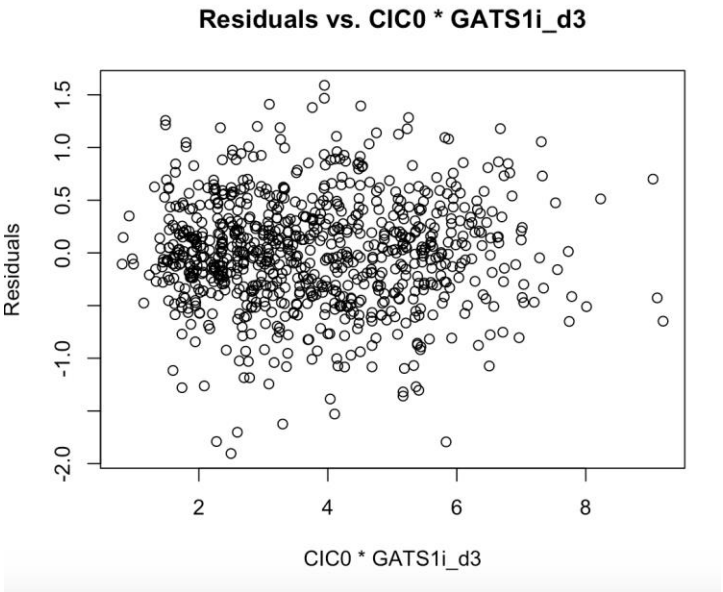
Model with CIC0 * SM1_Dz + CIC0 * MLOGP:




```
Analysis of Variance Table

Model 1: LC50_d3^(0.8585859) ~ MLOGP_d3 + SM1_Dz_d3 + factor(NdsCH_d3) +
    GATS1i_d3 + CIC0_d3 + MLOGP_d3:CIC0_d3
Model 2: LC50_d3^(0.8585859) ~ MLOGP_d3 + SM1_Dz_d3 + factor(NdsCH_d3) +
    GATS1i_d3 + CIC0_d3 + MLOGP_d3:CIC0_d3 + SM1_Dz_d3:CIC0_d3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    834 221.22
2    833 217.44  1    3.7741 14.458 0.0001538 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

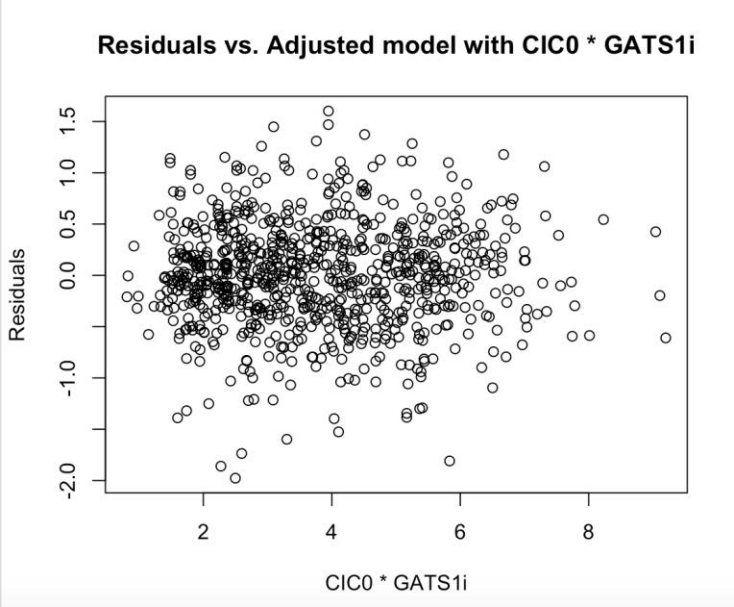
Model with CIC0 * GATS1i + CIC0 * MLOGP:



```
Analysis of Variance Table

Model 1: LC50_d3^(0.8585859) ~ MLOGP_d3 + SM1_Dz_d3 + factor(NdsCH_d3) +
    GATS1i_d3 + CIC0_d3 + MLOGP_d3:CIC0_d3
Model 2: LC50_d3^(0.8585859) ~ MLOGP_d3 + SM1_Dz_d3 + factor(NdsCH_d3) +
    GATS1i_d3 + CIC0_d3 + MLOGP_d3:CIC0_d3 + GATS1i_d3:CIC0_d3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    834 221.22
2    833 220.91  1   0.31247 1.1783  0.278
```

From the graph and f—test, we conclude that the model should be updates with both CIC0 * SM1_Dz and CIC0 * MLOGP. After this, we may test whether we need to include the last interaction term or not, which is CIC0 * GATS1i.




```
Analysis of Variance Table

Model 1: LC50_d3^(0.8585859) ~ MLOGP_d3 + SM1_Dz_d3 + factor(NdsCH_d3) +
    GATS1i_d3 + CIC0_d3 + MLOGP_d3:CIC0_d3 + SM1_Dz_d3:CIC0_d3
Model 2: LC50_d3^(0.8585859) ~ MLOGP_d3 + SM1_Dz_d3 + factor(NdsCH_d3) +
    GATS1i_d3 + CIC0_d3 + MLOGP_d3:CIC0_d3 + SM1_Dz_d3:CIC0_d3 +
    GATS1i_d3:CIC0_d3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    833 217.44
2    832 217.38  1  0.063877 0.2445 0.6211
```

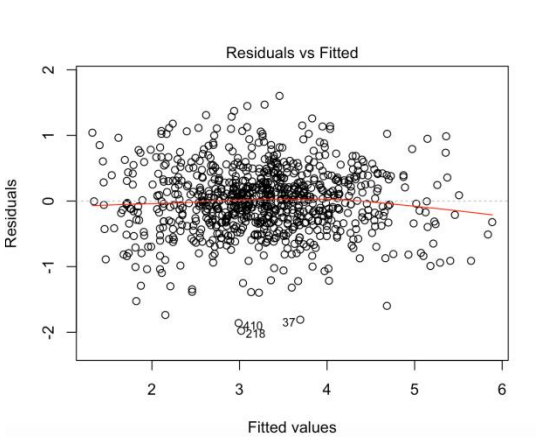
The answer is no, so we will use the original model updated with two interaction terms. The answer to our third research question is that there is a significance with the model including the interaction terms involving CIC0.

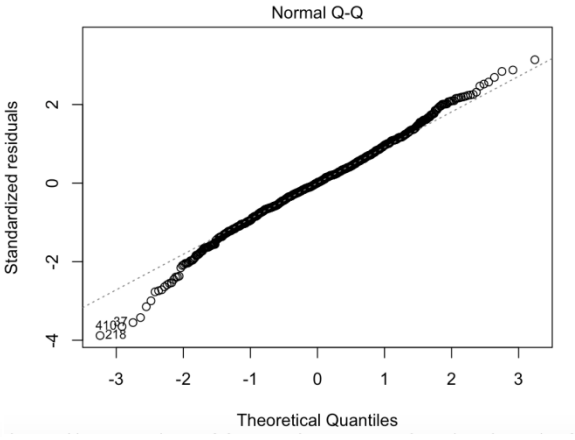
Updated Final Model

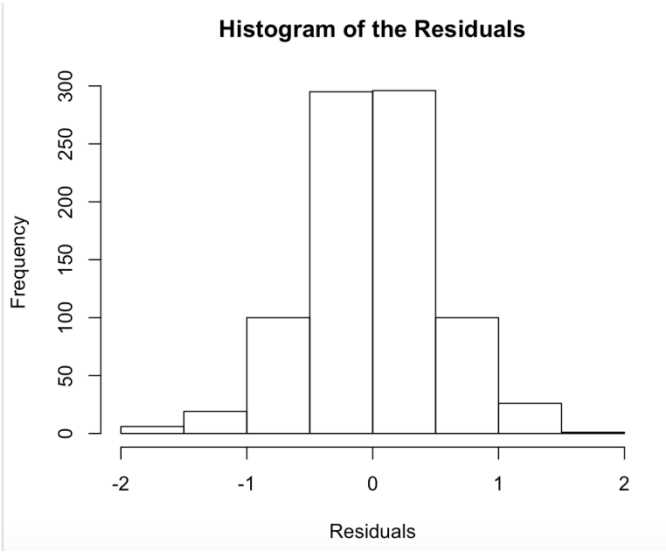
After answering our research question, we comes up a potentially better model than the previous one. Now we have:

$$E(LC50_i) = \beta 0 + \beta 1 * MLOGP_i + \beta 2 * CIC0_i + \beta 3 * GATS1i_i + \beta 4 * NdsCH_i + \beta 5 * SM1_Dz_i + \beta 6 * CIC0_i * MLOGP_i + \beta 7 * CIC0_i * SM1Dz$$

We then need to check if the new_model still fulfills the "LINE" conditions.







Then, we can use summary function on the model and conclude that we indeed achieve a better model with a better R^2 value. Now, we can confirm that our final model is reasonable.

```
Call:
lm(formula = LC50_d3^(0.8585859) ~ MLOGP_d3 + SM1_Dz_d3 + factor(NdsCH_d3) +
    GATS1i_d3 + CIC0_d3 + MLOGP_d3:CIC0_d3 + SM1_Dz_d3:CIC0_d3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.97693 -0.31037  0.00325  0.31125  1.60189

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.52076    0.19138   7.946 6.24e-15 ***
MLOGP_d3            0.05028    0.05960   0.844 0.399125
SM1_Dz_d3           1.78368    0.22554   7.909 8.26e-15 ***
factor(NdsCH_d3)1   0.45854    0.05557   8.251 6.09e-16 ***
factor(NdsCH_d3)2   0.56802    0.10481   5.419 7.84e-08 ***
factor(NdsCH_d3)4   0.56912    0.30596   1.860 0.063224 .
GATS1i_d3          -0.52731    0.06055  -8.708  < 2e-16 ***
CIC0_d3             0.37231    0.06980   5.334 1.24e-07 ***
MLOGP_d3:CIC0_d3    0.08007    0.01763   4.541 6.42e-06 ***
SM1_Dz_d3:CIC0_d3  -0.27129    0.07135  -3.802 0.000154 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5109 on 833 degrees of freedom
Multiple R-squared:  0.7136,    Adjusted R-squared:  0.7105
F-statistic: 230.6 on 9 and 833 DF,  p-value: < 2.2e-16
```

Final model is:

$$\mathbf{E}(LC50_i) = \beta 0 + \beta 1 * MLOGP_i + \beta 2 * CIC0_i + \beta 3 * GATS1i_i + \beta 4 * NdsCH_i + \beta 5 * SM1_Dz_i + \beta 6 * CIC0_i * MLOGP_i + \beta 7 * CIC0_i * SM1Dz$$

Conclusion

In conclusion, we build a linear regression model for the multivariate data set QSAR fish toxicity regarding the concentration of deadly toxicity associated to the six predictors: MLOGP, CIC0, GATS1i, NDssC, NDsCH and SM1_Dz. Along the way of solving research questions, we found we can be 95 percent confident that given NdsCH = 0, NdssC = 0, CIC0 = 3, average values of MLOGP and SM1_Dz(Z), the response would be between (3.853171, 3.963089). Another interesting finding is that the influential points' effect is huge on this dataset since the we can see the R^2 is improved by over 0.1 after taking off the effect from outliers and high leverage points.

We also find a better model along the studying our third research questions. The new model does not have an overfit condition since it does not include every interaction terms, achieves better look graphs on "LINE" conditions and last but not least, it improves the coefficient of determination value by 0.01.

Still, there is something needed to be improved for this data set analysis. Firstly, we don't have a handle information of serial correlation in the data set, so we have to assume the condition of independence of error terms has been met. Additionally, two of the predictors are discrete and fixed variables, which represent the atom of the chemicals, but they actually may have a huge effect on the model, that is why we think our adjusted model and data still have 0.711 on R^2. The data was also based on one kind of fish so what we found in this report can only be applied to such fish kind. Last but not least, it may be interesting to categorize the atomic number for one or two of the variables (NdssC & NdsCH) or including other potential interaction term so there might be a better model. However, we need to be cautious on overfitting while doing this.

Reference

Chun, Christina. "How Does Tumor Size Relate to Breast Cancer Stage?" Medical News Today, Sep 23, 2019, https://archive.ics.uci.edu/ml/datasets/QSAR+fish+toxicity. Accessed 23 Nov. 2019

M. Cassotti, D. Ballabio, R. Todeschini, V. Consonni. (2015) A similarity−based QSAR model for predicting acute toxicity towards the fathead minnow (Pimephales promelas). SAR and QSAR in Environmental Research 26:3, pages 217$-$243.

Appendix

```
library(base)
library(car)
library(carData)
library(corrplot)
library(datasets)
library(forcats)
library(ggplot2)
library(graphics)
library(methods)
library(readr)
library(dplyr)
library(leaps)
library(purrr)
library(MASS)
library(tibble)
library(tidyr)
library(utils)

#load file into datagram
QSAR_fish <- read.csv("qsar_fish_toxicity.csv")
names(QSAR_fish) <- c("CIC0", "SM1_Dz", "GATS1i", "NdsCH", "NdssC",
"MLOGP","LC50")
```

```
#initialize variables
CIC0 <- QSAR_fish$CIC0
SM1_Dz <- QSAR_fish$SM1_Dz
GATS1i <- QSAR_fish$GATS1i
NdsCH <- QSAR_fish$NdsCH
NdssC <- QSAR_fish$NdssC
MLOGP <- QSAR_fish$MLOGP

#Response
LC50 <- QSAR_fish$LC50

#Scatterplot
pairs(LC50~CIC0+SM1_Dz+GATS1i+NdsCH+NdssC+MLOGP,QSAR_fish,main =
"Scatterplot Matrix of QSAR_fish data")
#pairs(LC50^(0.8585859)~MLOGP)

#correlation
cor(subset(QSAR_fish,select = -c(LC50)))
#Variance Inflation Factor
vif(subset(QSAR_fish,select = -c(LC50)))

#First model
full.mod <- lm(LC50~CIC0+SM1_Dz+GATS1i+factor(NdsCH)+factor(NdssC)+MLOGP)
#plots, observe Non-constant error variance
plot(full.mod)
#histgram
hist(resid(full.mod),main = "Histogram of the Residuals", xlab = "Residuals")

#plot(CIC0, resid(full.mod), main = "Residuals vs. CIC0", xlab = "CIC0", ylab =
"residuals")

#plot(SM1_Dz, resid(full.mod), main = "Residuals vs. SM1_Dz", xlab = "SM1_Dz", ylab
= "residuals")
#plot(GATS1i, resid(full.mod), main = "Residuals vs. GATS1i", xlab = "GATS1i", ylab
= "residuals")
#plot(factor(NdsCH), resid(full.mod), main = "Residuals vs. NdsCH", xlab = "NdsCH",
ylab = "residuals")
#plot(factor(NdssC), resid(full.mod), main = "Residuals vs. NdssC", xlab = "NdssC",
ylab = "residuals")
#plot(MLOGP, resid(full.mod), main = "Residuals vs. MLOGP", xlab = "MLOGP", ylab
= "residuals")

#boccox
potential.trans <- boxcox(full.mod, lambda=seq(-1,1,length=10))
potential.trans$x[which.max(potential.trans$y)]
```

```
#Second model dealt with Non-constant error variance with Transformation on
Response
full.mod2 <-
lm(LC50^(0.8585859)~CIC0+SM1_Dz+GATS1i+factor(NdsCH)+factor(NdssC)+MLO
GP)
plot(full.mod2)
hist(resid(full.mod2),main = "Histogram of the Residuals", xlab = "Residuals")

#Stepwise Regression using AIC
mod0 = lm(LC50^(0.8585859)~1)
mod.upper =
lm(LC50^(0.8585859)~CIC0+SM1_Dz+GATS1i+NdsCH+NdssC+MLOGP)
summary(mod.upper)
step(mod0,scope = list(lower=mod0,upper=mod.upper))

#Best subset regression
mod.bsubsets <-
regsubsets(cbind(CIC0,SM1_Dz,GATS1i,NdsCH,NdssC,MLOGP),LC50^(0.8585859))
summary.bsubsets <- summary(mod.bsubsets)
#with adjusted R^2 criteria
summary.bsubsets$adjr2
summary.bsubsets$which
#Mallows CP
summary.bsubsets$cp

#Stepwise regression using F-tests
add1(mod0,~.+CIC0+SM1_Dz+GATS1i+NdsCH+NdssC+MLOGP,test="F")
mod1 = update(mod0, ~.+MLOGP)
add1(mod1,~.+CIC0+SM1_Dz+GATS1i+NdsCH+NdssC,test="F")
mod2 = update(mod1, ~.+SM1_Dz)
summary(mod2)
add1(mod2,~.+CIC0+GATS1i+NdsCH+NdssC,test="F")
mod3 = update(mod2, ~.+NdsCH)
summary(mod3)
add1(mod3,~.+CIC0+GATS1i+NdssC,test="F")
mod4 = update(mod3, ~.+GATS1i)
summary(mod4)
add1(mod4,~.+CIC0+NdssC,test="F")
mod5 = update(mod4, ~.+CIC0)
summary(mod5)
add1(mod5,~.+NdssC,test="F")
#not adding NdssC

#Delete influential points using Studentized Deleted Residuals
which(abs(rstudent(full.mod2)) > 3)
```

```
Data_deleted <- QSAR_fish[-c(66,121,170,183,220,238,260,288,373,448,681),]
CIC0_d <- Data_deleted$CIC0
SM1_Dz_d <- Data_deleted$SM1_Dz
GATS1i_d <- Data_deleted$GATS1i
NdsCH_d <- Data_deleted$NdsCH
NdssC_d <- Data_deleted$NdssC
MLOGP_d <- Data_deleted$MLOGP

LC50_d <- Data_deleted$LC50

full.mod4 <-
lm(LC50_d^(0.8585859)~MLOGP_d+SM1_Dz_d+factor(NdsCH_d)+GATS1i_d+CIC0_
d)

#deleted base on DFFITS value
dffits(full.mod2)
dim(QSAR_fish)
num = 2 * sqrt((7 + 1)/(908-7-1))
which(dffits(full.mod2) > num)
Data_deleted_2 <- QSAR_fish[-
c(19,31,45,62,66,73,75,76,77,78,84,85,121,152,166,170,178,183,220,231,238,250,2
55,257,260,282,288,303,348,350,364,373,417,418,429,443,448,455,463,483,503,5
05,551,552,563,578,681,692,695,709,724,726,761,768,784,826,890,893,901,905,9
06),]
CIC0_d2 <- Data_deleted_2$CIC0
SM1_Dz_d2 <- Data_deleted_2$SM1_Dz
GATS1i_d2 <- Data_deleted_2$GATS1i
NdsCH_d2 <- Data_deleted_2$NdsCH
NdssC_d2 <- Data_deleted_2$NdssC
MLOGP_d2 <- Data_deleted_2$MLOGP

LC50_d2 <- Data_deleted_2$LC50
full.mod5 <-
lm(LC50_d2^(0.8585859)~MLOGP_d2+SM1_Dz_d2+factor(NdsCH_d2)+GATS1i_d2+
CIC0_d2)

#Question1
reduced.mod <- lm(LC50_d2^(0.8585859)~1)
full.mod6 <-
lm(LC50_d2^(0.8585859)~MLOGP_d2+SM1_Dz_d2+factor(NdsCH_d2)+GATS1i_d2+
CIC0_d2)
anova(reduced.mod, full.mod6)

#Question2
```

```
new.df = data.frame(MLOGP_d2 = mean(MLOGP_d2), SM1_Dz_d2 =
mean(SM1_Dz_d2), CIC0_d2 = 3, GATS1i_d2=mean(GATS1i_d2), NdsCH_d2 =
0,NdssC_d2 = 0)
predicted = predict(full.mod8, new.df, se.fit = TRUE, interval = "confidence", level =
0.95, type = "response")
predicted$fit

#Question3
interaction_mod1 = update(full.mod5, ~. + CIC0_d3*MLOGP_d3)
plot(CIC0_d3*MLOGP_d3, resid(full.mod5),main = "Residuals vs. CIC0 * MLOGP",
xlab = "CIC0 * MLOGP",
    ylab = "Residuals")
anova(full.mod5,interaction_mod1)

interaction_mod2 = update(interaction_mod1, ~. + CIC0_d3*SM1_Dz_d3)
plot(CIC0_d3*SM1_Dz_d3, resid(interaction_mod1),main = "Residuals vs. CIC0 *
SM1_Dz", xlab = "CIC0 * SM1_Dz",
    ylab = "Residuals")
anova(interaction_mod1,interaction_mod2)

interaction_mod2.1 = update(interaction_mod1, ~. + CIC0_d3*GATS1i_d3)
plot(CIC0_d3*GATS1i_d3, resid(interaction_mod1),main = "Residuals vs. CIC0 *
GATS1i_d3", xlab = "CIC0 * GATS1i_d3",
    ylab = "Residuals")
anova(interaction_mod1,interaction_mod2.1)

interaction_mod3 = update(interaction_mod2, ~. + CIC0_d3*GATS1i_d3)
plot(CIC0_d3*GATS1i_d3, resid(interaction_mod2),main = "Residuals vs. Adjusted
model with CIC0 * GATS1i", xlab = "CIC0 * GATS1i",
    ylab = "Residuals")
anova(interaction_mod2,interaction_mod3)

plot(interaction_mod2)

hist(resid(interaction_mod2), main = "Histogram of the Residuals", xlab =
"Residuals")

summary(interaction_mod2)
```