QSAR Fish Toxicity

David Kao
Joanne Chan
Yifan Wei
Oyku Ozer

Introduction: Data

	CIC0	SM1_Dz	GATS1i	NdsCH	NdssC	MLOGP	y
1	3.26	0.829	1.676	0	1	1.453	3.77
2	2.189	0.58	0.863	0	0	1.348	3.115
3	2.125	0.638	0.831	0	0	1.348	3.531
4	3.027	0.331	1.472	1	0	1.807	3.51
5	2.094	0.827	0.86	0	0	1.886	5.39
6	3.222	0.331	2.177	0	0	0.706	1.819
7	3.179	0	1.063	0	0	2.942	3.947
8	3	0	0.938	1	0	2.851	3.513
9	2.62	0.499	0.99	0	0	2.942	4.402
10	2.834	0.134	0.95	0	0	1.591	3.021

Showing 1 to 10 of 908 entries Previous 1 2 3 4 5 … 91 Next

- The data set consists of 908 examples with 7 variables.

- **Explanatory Variables:**
  - Numerical
    - CIC0 (information indices)
    - SM1_DZ (2D matrix-based descriptors)
    - GATS1i (2D autocorrelations)
    - MLOGP (molecular properties)
    - NdsCH (atom-type counts)
    - NdssC (atom-type counts)
- LC50 (median lethal dose) will be the response variable.

Introduction: Hypothesis

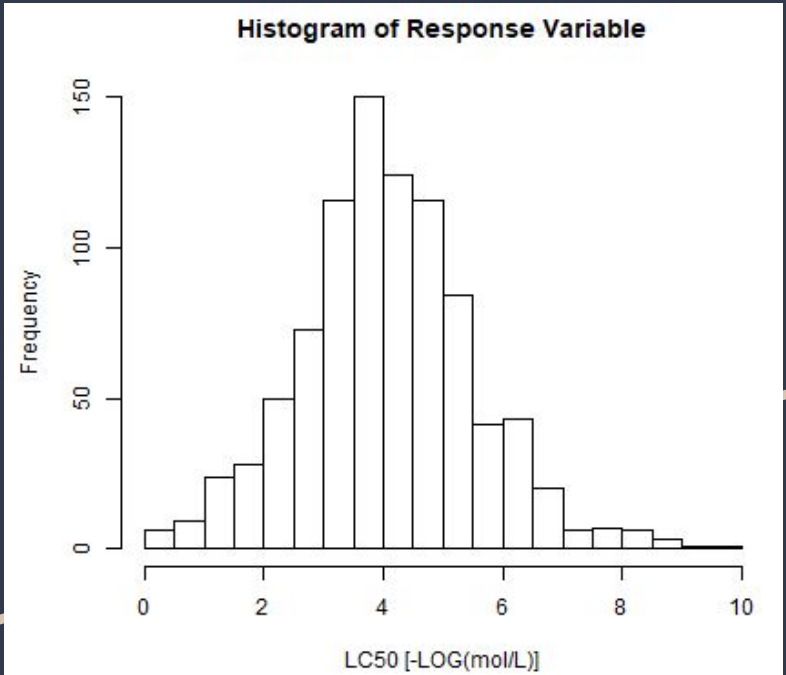
Type of study: This is an experimental study that aims for prediction.

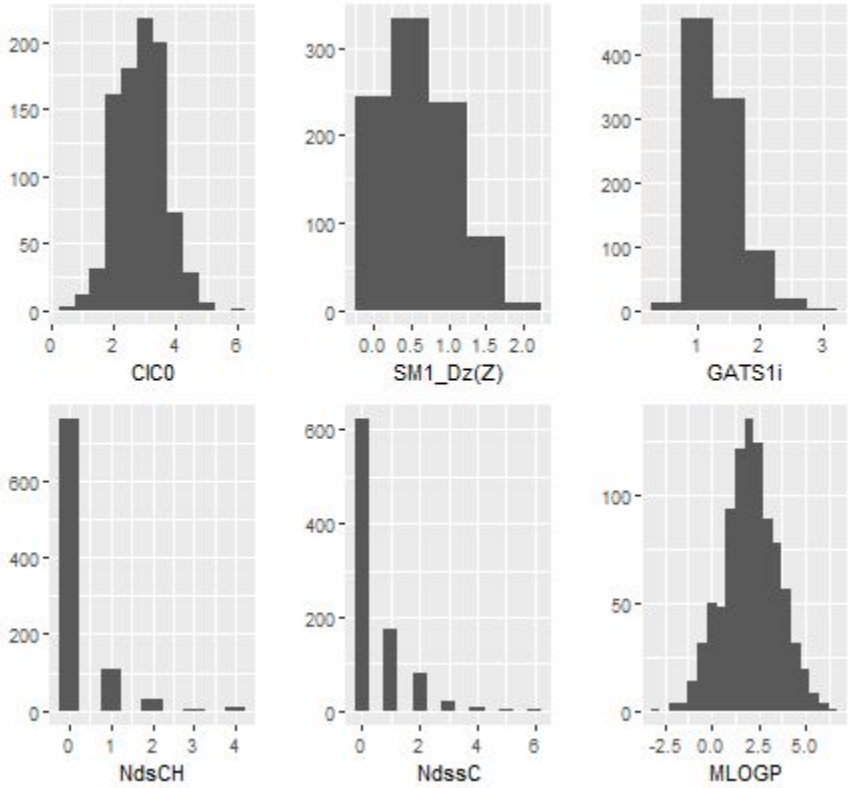
Study paper:
This study aimed to predict the toxicity of chemicals toward a small planktonic species, Daphnia magna using a QSAR model with a data set of 546 molecules. The regression method used was the k-Nearest Neighbour (kNN).

Null Hypothesis: There will be no significant prediction of LC50 (median lethal dose) concentration towards the fathead minnow species by the molecular descriptors: CIC0, SM1_DZ, GATS1i, MLOGP, NdsCH, and NdssC.

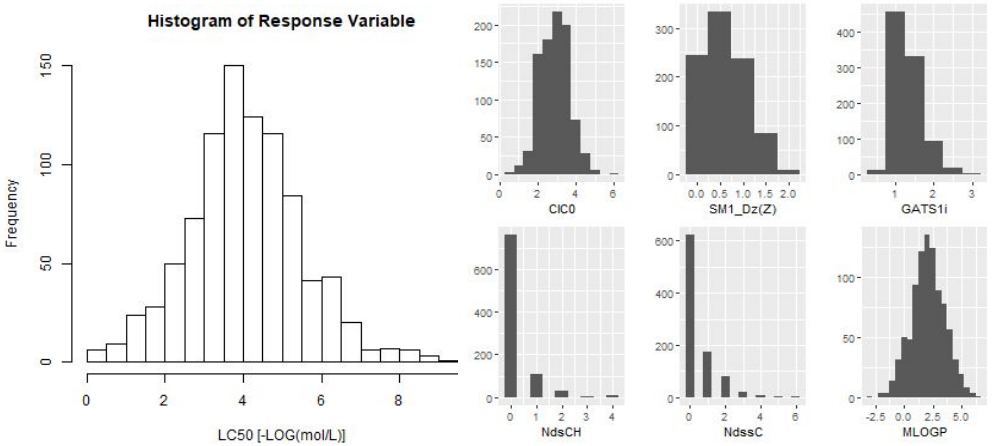
Alternative Hypothesis: There will be a significant prediction of LC50 (median lethal dose) concentration towards the fathead minnow species by the molecular descriptors: CIC0, SM1_DZ, GATS1i, MLOGP, NdsCH, and NdssC.

Data (Histogram)





Data (Histogram) Cont.



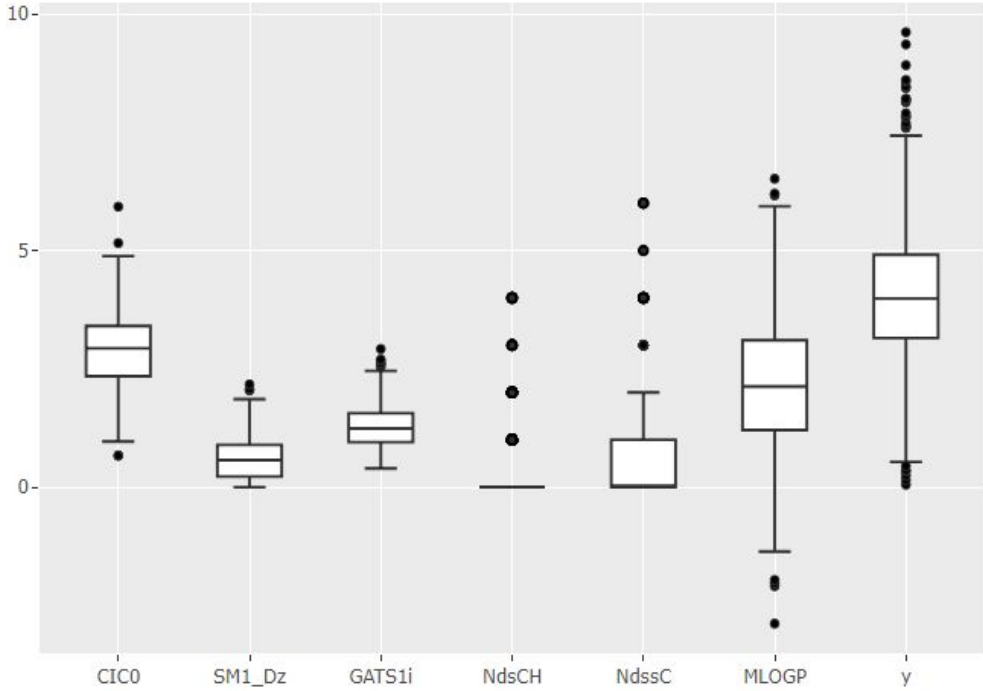
- Histogram shows us that:
  - CIC0 is normal
  - SM1_DZ is right skewed
  - GATS1i is right skewed
  - MLOGP is normal
  - LC50 (Response) is normal
  - NdsCH has mostly 0's and 1's
  - NdssC has mostly 0's and 1's
- Since data are mostly normal, we decided to use the original data instead of transformed data

Data (Box Plot)

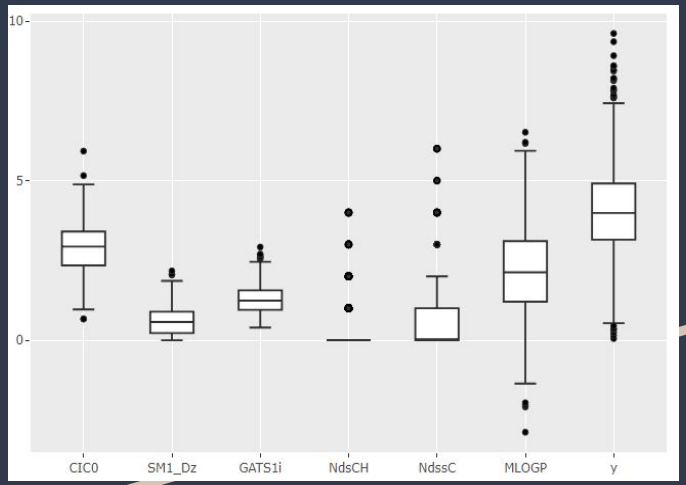
Statistics summary:

```
> summarydt
                CIC0       SM1_Dz    GATS1i
Min.     0.667000 0.0000000 0.396000
1st Qu.  2.347000 0.2230000 0.950750
Median   2.934000 0.5700000 1.240500
Mean     2.898129 0.6284681 1.293591
3rd Qu.  3.407000 0.8927500 1.562250
Max.     5.926000 2.1710000 2.920000
                NdsCH      NdssC      MLOGP
Min.     0.0000000 0.0000000 -2.884000
1st Qu.  0.0000000 0.0000000  1.209000
Median   0.0000000 0.0000000  2.127000
Mean     0.2290749 0.4856828  2.109285
3rd Qu.  0.0000000 1.0000000  3.105000
Max.     4.0000000 6.0000000  6.515000
                y
Min.     0.053000
1st Qu.  3.151750
Median   3.987500
Mean     4.064431
3rd Qu.  4.907500
Max.     9.612000
```

Box Plot for variables

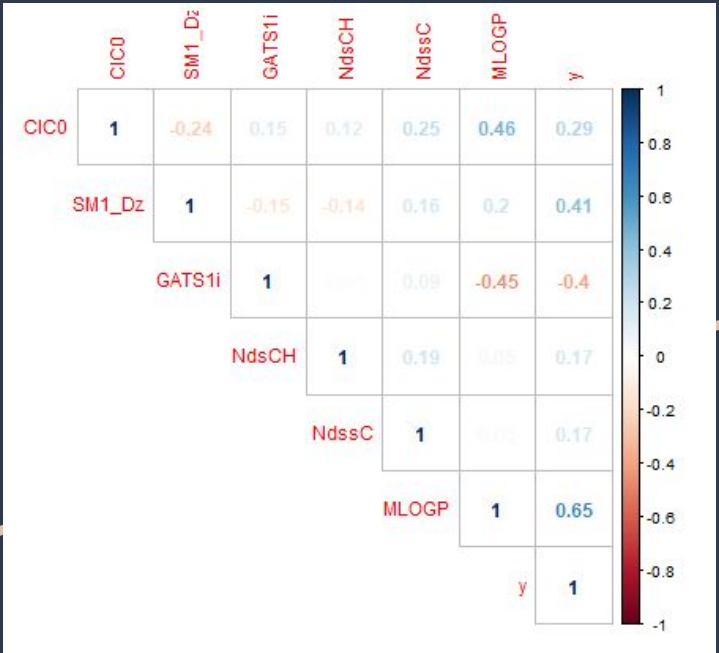


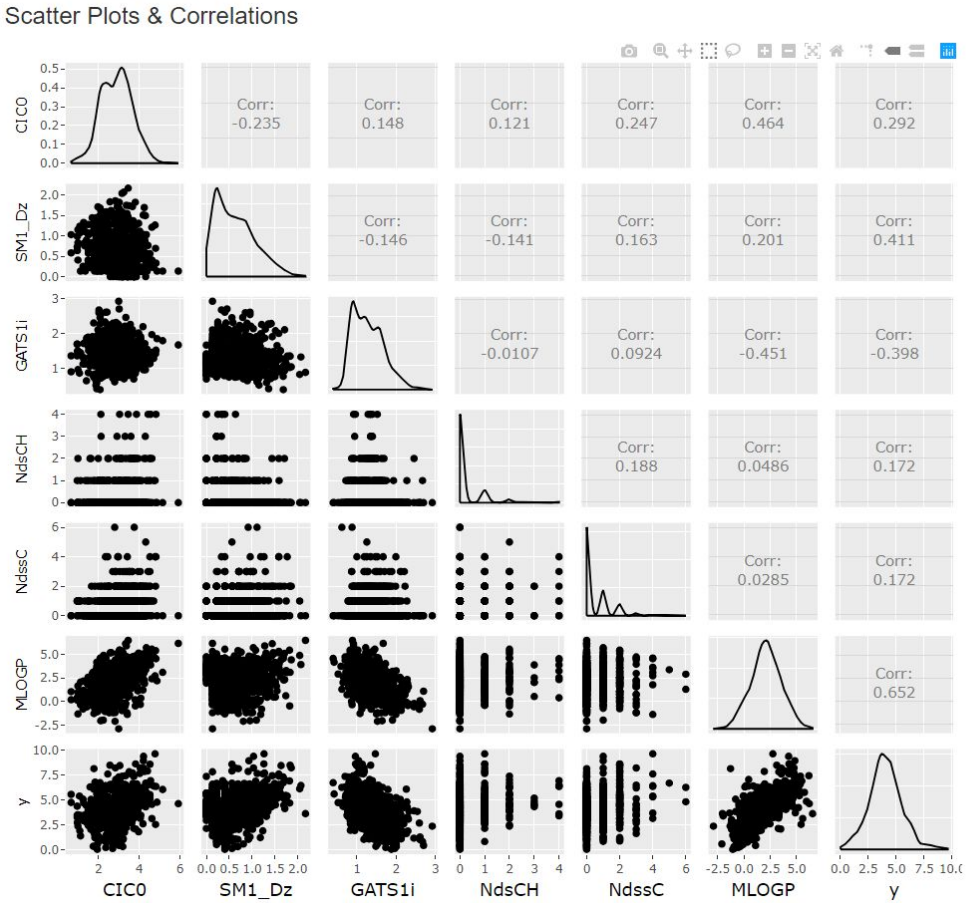
Data (Box Plot) cont.



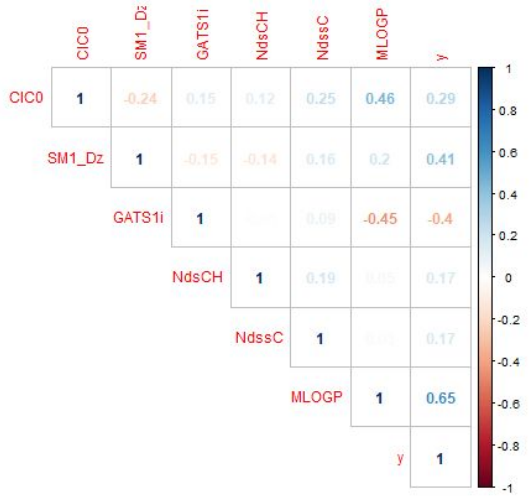
- IQR
  - MLOGP> LC50 > CIC0 > SM1_Dz > GATS1i
- Mean
  - LC50 > CIC > MLOGP > GATS1i > SM1_Dz
- Outliers
  - 3 for CIC0
  - 2 for SM1_Dz
  - 3~4 for GATS1i
  - 5~6 for MLOGP
  - About 20 for LC50

Data (Scatter Plot/Correlation Heat Map)





Data (Scatter Plot/Correlation Heat Map) cont.



- CIC0 and MLOGP has a strong positive correlation
- MLOGP and LC50 has a strong positive correlation
- GATS1i and MLOGP has a moderate negative correlation
- GATS1i and LC50 has a moderate negative correlation

Variable Selection

- Forward selection
- Selection criteria:
  - Cp Statistics
  - Adjusted R^2

```
Call: regsubsets.formula(dt$y ~ ., data = dt, method = "forward")
6 Variables  (and intercept)
       Forced in Forced out
CIC0       FALSE      FALSE
SM1_Dz     FALSE      FALSE
GATS1i     FALSE      FALSE
NdsCH      FALSE      FALSE
NdssC      FALSE      FALSE
MLOGP      FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: forward
         CIC0 SM1_Dz GATS1i NdsCH NdssC MLOGP
1  ( 1 ) " "  " "    " "    " "   " "   "*"
2  ( 1 ) " "  "*"    " "    " "   " "   "*"
3  ( 1 ) " "  "*"    " "    "*"   " "   "*"
4  ( 1 ) " "  "*"    "*"    "*"   " "   "*"
5  ( 1 ) "*"  "*"    "*"    "*"   " "   "*"
6  ( 1 ) "*"  "*"    "*"    "*"   "*"   "*"
```

```
> models_res <- data.frame(
+   Adj.R2 = which.max(models_summary$adjr2),
+   CP = which.min(models_summary$cp)
+ ); models_res
  Adj.R2 CP
1      6  6
> models_summary$adjr2
[1] 0.4240310 0.5053395 0.5398406 0.5492448 0.5736370 0.5743478
> models_summary$cp
[1] 321.949063 149.721899  77.286447  58.254638   7.506151   7.000000
```

Model 1

```
Call:
lm(formula = y ~ CIC0 + SM1_Dz + GATS1i + NdsCH + NdssC + MLOGP)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4921 -0.5287 -0.0712  0.4861  5.6876

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.17456    0.18122  12.000  < 2e-16 ***
CIC0         0.38563    0.06089   6.333 3.79e-10 ***
SM1_Dz       1.25562    0.08702  14.430  < 2e-16 ***
GATS1i      -0.74641    0.10135  -7.365 4.00e-13 ***
NdsCH        0.41355    0.05410   7.644 5.41e-14 ***
NdssC        0.06433    0.04064   1.583    0.114
MLOGP        0.39005    0.03376  11.555  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9497 on 901 degrees of freedom
Multiple R-squared:  0.5772,    Adjusted R-squared:  0.5743
F-statistic:   205 on 6 and 901 DF,  p-value: < 2.2e-16
```

Model 1 includes all 6 variables:

```
model1 <- lm(y ~ CIC0 + SM1_Dz + GATS1i + NdsCH
+ NdssC + MLOGP)
```

Adjusted R^2 = 0.5734

Other models

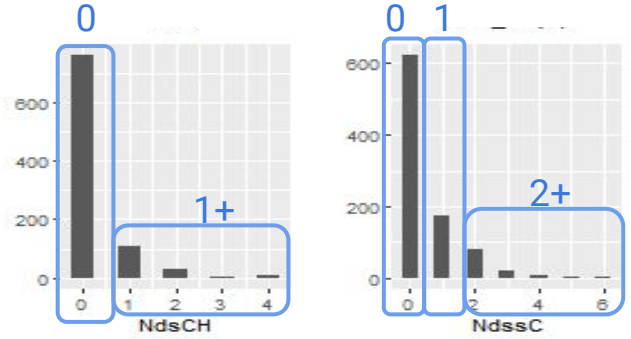
```
# Introducing some interaction terms and see how adj R2 changes:
reg3 <- lm(y ~ CIC0 + SM1_Dz + GATS1i + NdsCH + NdssC + MLOGP + I(MLOGP*CIC0)); reg3
summary(reg3) # 0.5827, I(MLOGP*CIC0)) is significant

reg3 <- lm(y ~ CIC0 + SM1_Dz + GATS1i + NdsCH + NdssC + MLOGP + I(MLOGP*CIC0) + I(MLOGP*GATS1i)); reg3
summary(reg3) # 0.5826, I(MLOGP*GATS1i) is NOT significant

reg3 <- lm(y ~ CIC0 + SM1_Dz + GATS1i + NdsCH + NdssC + MLOGP + I(MLOGP*CIC0) + I(CIC0*NdssC)); reg3
summary(reg3) # 0.5822, I(CIC0*NdssC) is NOT significant

reg3 <- lm(y ~ CIC0 + SM1_Dz + GATS1i + NdsCH + NdssC + MLOGP + I(MLOGP*CIC0) + I(NdsCH*NdssC)); reg3
summary(reg3) # 0.5828, I(NdsCH*NdssC) is NOT significant

reg3 <- lm(y ~ CIC0 + SM1_Dz + GATS1i + NdsCH + NdssC+ I(MLOGP*CIC0) + I(SM1_Dz*CIC0)); reg3
summary(reg3) # 0.5852, I(SM1_Dz* CIC0) is significant

# Introducing some quadratic/ cubic terms
reg3 <- lm(y ~ I(CIC0^2) + CIC0 + SM1_Dz + GATS1i + NdsCH + NdssC + I(MLOGP*CIC0) + I(SM1_Dz*CIC0)); reg3
summary(reg3) # 0.5891, I(CIC0^2) is significant

reg3 <- lm(y ~ I(CIC0^2) + I(SM1_Dz^2) + SM1_Dz + GATS1i + NdsCH + NdssC + I(MLOGP*CIC0) + I(SM1_Dz*CIC0)); reg3
summary(reg3) # 0.5870, I(SM1_Dz^2) is significant (note, we took out CIC0 in this model)

reg3 <- lm(y ~ I(CIC0^2) + I(SM1_Dz^2) + I(MLOGP^2) + SM1_Dz + GATS1i + NdsCH + NdssC + I(MLOGP * CIC0) + I(SM1_Dz* CIC0)); reg3
summary(reg3) # 0.5875, I(MLOGP^2) is NOT sigfinicant

reg3 <- lm(y ~ I(CIC0^2) + I(SM1_Dz^2) + I(MLOGP^3) + SM1_Dz + GATS1i + NdsCH + NdssC + I(MLOGP * CIC0) + I(SM1_Dz* CIC0)); reg3
summary(reg3) # 0.5943, I(MLOGP^3) is sigfinicant
```

Model 2

```
lm(formula = y ~ I(CIC0^2) + I(SM1_Dz^2) + I(MLOGP^3) + SM1_Dz +
    GATS1i + FactNdsCH + FactNdssC + I(MLOGP * CIC0) + I(SM1_Dz *
    CIC0) + I(MLOGP * SM1_Dz))

Residuals:
    Min      1Q  Median      3Q     Max
-3.9528 -0.4917 -0.0285  0.4683  5.2531

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.970376   0.188228  15.781  < 2e-16 ***
I(CIC0^2)         0.025886   0.022670   1.142 0.253802
I(SM1_Dz^2)       0.452090   0.155254   2.912 0.003681 **
I(MLOGP^3)       -0.007808   0.002087  -3.740 0.000195 ***
SM1_Dz            1.491453   0.376461   3.962 8.03e-05 ***
GATS1i           -0.808063   0.099128  -8.152 1.20e-15 ***
FactNdsCH1+       0.673093   0.084958   7.923 6.88e-15 ***
FactNdssC1       -0.116537   0.079828  -1.460 0.144681
FactNdssC2+       0.229353   0.105198   2.180 0.029502 *
I(MLOGP * CIC0)   0.182168   0.020526   8.875  < 2e-16 ***
I(SM1_Dz * CIC0) -0.190622   0.103846  -1.836 0.066745 .
I(MLOGP * SM1_Dz) -0.072544   0.051774  -1.401 0.161511
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9154 on 896 degrees of freedom
Multiple R-squared:  0.6093,    Adjusted R-squared:  0.6045
F-statistic: 127.1 on 11 and 896 DF,  p-value: < 2.2e-16
```

Changing NdsCH and NdssC to factors:

- FactNdsCH has categories “0” and “1+”
- FactNdssC has categories “0”, “1” and “2+”

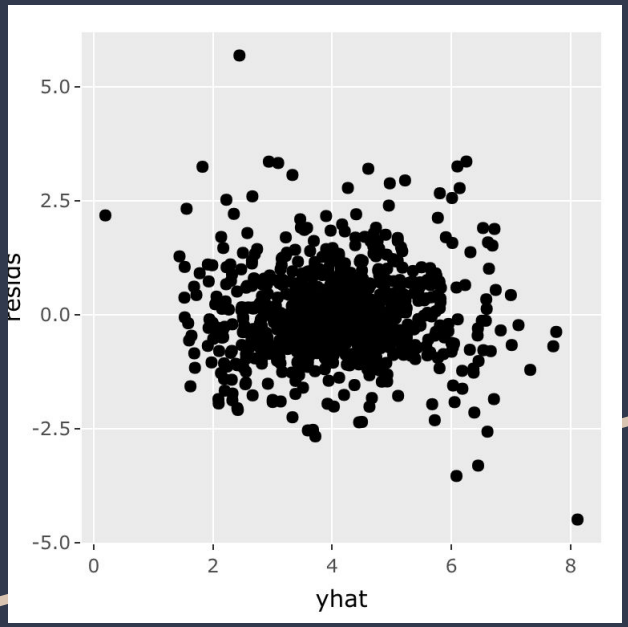


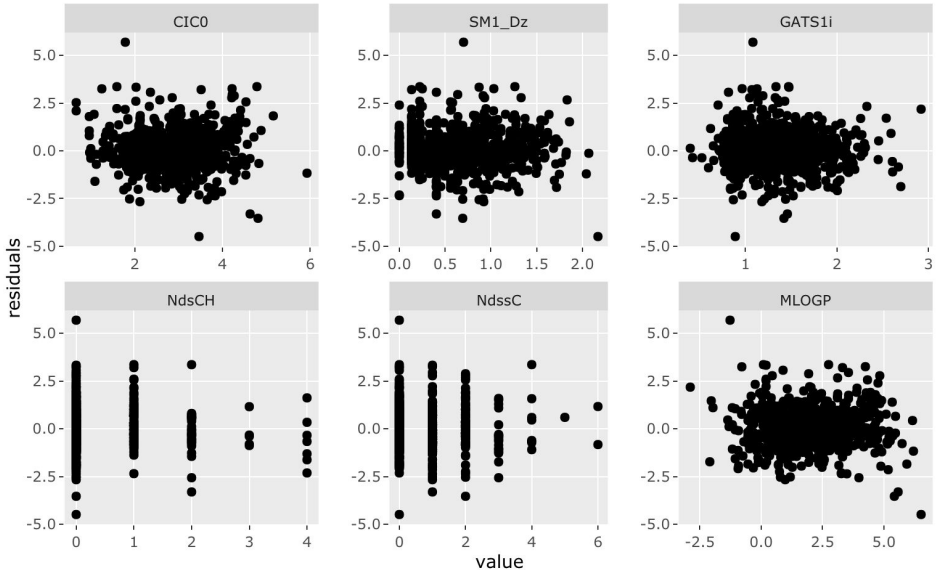
Model 2:

```
lm(formula = y ~ I(CIC0^2) + I(SM1_Dz^2) +
I(MLOGP^3) + SM1_Dz + GATS1i + FactNdsCH +
FactNdssC + I(MLOGP * CIC0) + I(SM1_Dz * CIC0)
+ I(MLOGP * SM1_Dz))
```

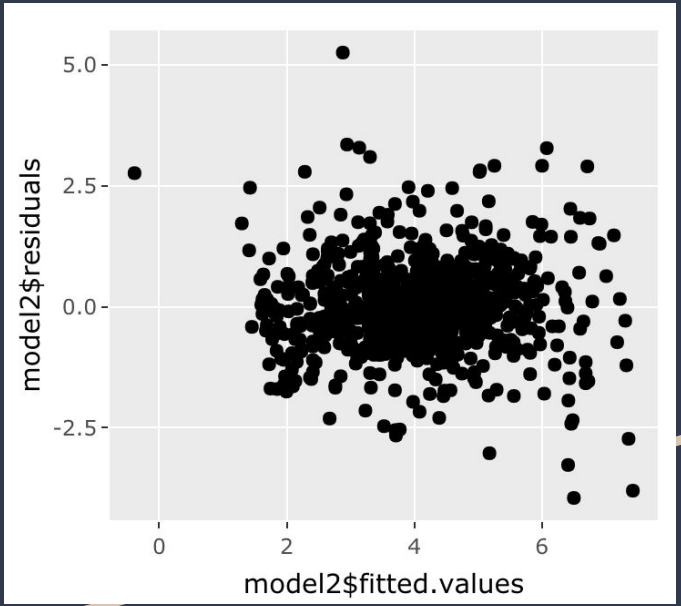
Adjusted $R^2 = 0.6045$

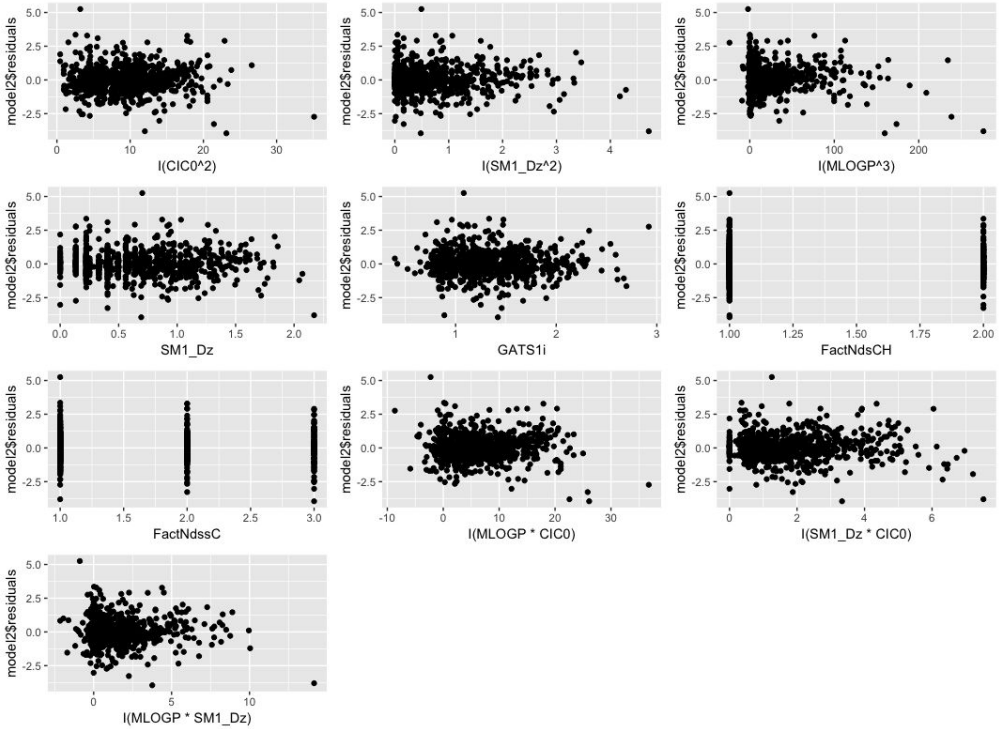
Model 1 – Residual Plots





Model 2 - Residual Plots





Interpretation of result

- Residual plots

For both residual plots from model 1 and 2, there is no obvious pattern. Thus, the plot is unbiased and homoscedastic, which means in both model the average value of residuals is 0 and sd of residuals is same in any thin rectangle in

- Check adj R^2

For model one , the adjusted R^2 is 0.5734 and that for model 2 is 0.6054.

Conclusion

- Conclusion
  - We can not make a decision for model selection only based on the residual plots, because they show the same pattern . So we check the adjusted R square to get the conclusion.
  - Since model 2 has 3% higher adjusted R^2 value, we choose model 2 as the best model.
  - Although model 2 is very ideal, we can still reject the null hypothesis .

    *Note that model 2 is a much complicated model, so it may not be the best model if there are too many examples in the data set
- Reservation:
  - Try other kind of models like K-nearest neighbour, Naive Bayes to achieve an even higher adjusted R^2(Original paper used KNN model, which had a R^2 of 0.78)
  - Try transforming numerical data, since some data may still not be normal.