Analyzing the Toxicity Dataset

By: Alex Djidjev, Kyle Westwood, Hariswar Baburaj, Karthik Thota

Introduction

Chemicals can have immense effects on aquatic systems. There is a type of water flea calles the Daphnia magna which is used to evaluate the aquatic toxicity of chemicals. There is a concentration of chemical called LC50 which kills 50% of test organisms over a duration of 48 hours. So, having a low LC50 means that the chemical is more toxic because a lower concentration of it killed 50% of the animal. Having a high LC50 which means the chemical is less toxic since a higher concentration was needed to kill 50% of the animal which means that it is not as toxic. We have got an overall of 350 chemicals and 8 predictors for this Toxicity Dataset.

Predictors: (molecular descriptors)

- TPSA (tot) – topological polar surface area
- SAacc – Van der Waals surface area of atoms that are acceptors of hydrogen bonds
- H050 – number of hydrogen atoms bonded to heteroatoms
- MLOGP – octanol-water partition coefficient calculated from the Moriguchi model
- RDCHI – topological index with information about molecular size & branching
- GATS1p – encodes information on molecular polarisability
- nN - number of nitrogen atoms present in the molecule
- C040 – number of carbon atoms of a specific type

All of our 8 predictors are quantitative variable types. Our goal of this analysis and report is to predict the concentration of a given chemical that will kill 50% of the water fleas (Daphnia magna) based on molecular descriptors of the chemicals and be able to know which chemicals are more toxic to Daphnia magna than others solely based on the molecular descriptors of the chemicals.

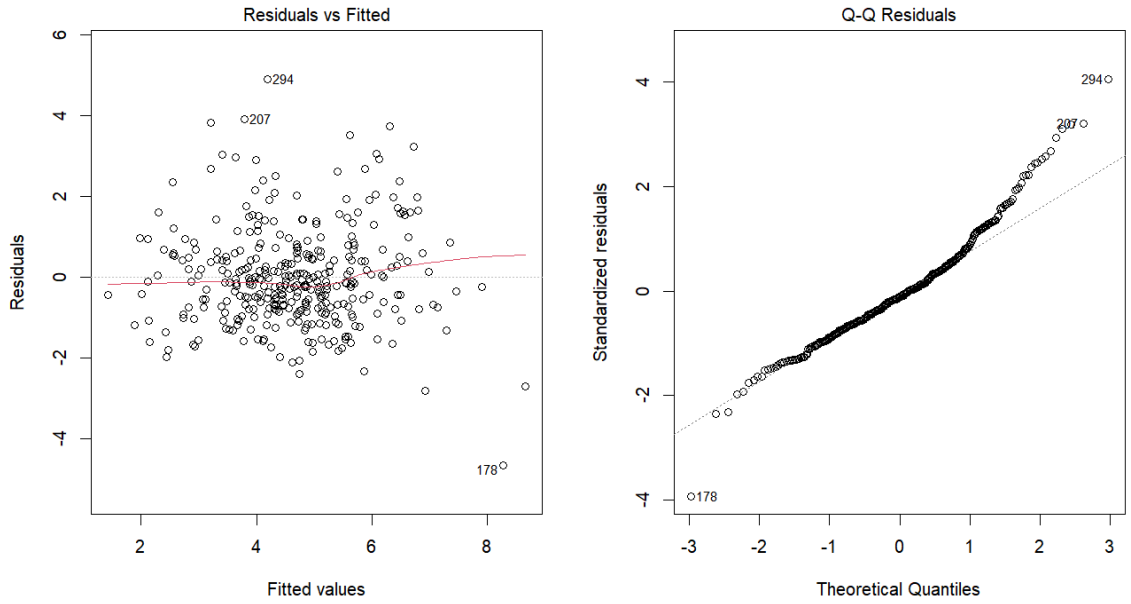
Methodology

We opted to use 7 distinct statistical learning methods to this task of predicting toxicity values based on the molecular descriptors of a chemical. These methods include: Multiple Linear Regression, Polynomial Term Model, Best Subsets Model, Interaction Term Model, Lasso Regression, Principal Component Regression, Partial Least Squares Regression.

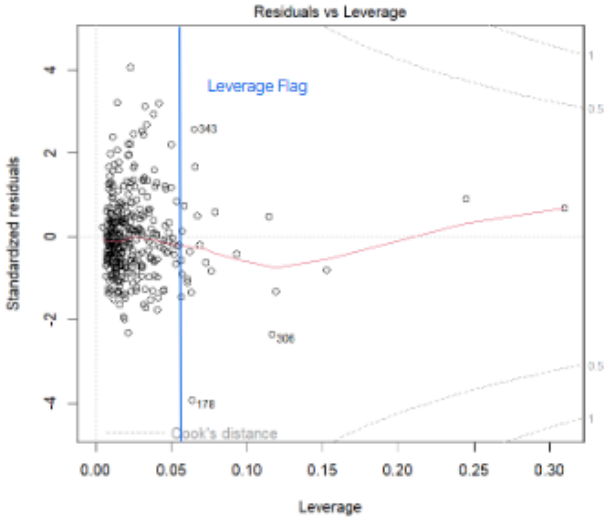
Furthermore, our dataset included 350 observations of chemicals, their LC50 values and their molecular descriptor values. To accurately evaluate each model, we initially split the dataset into a training (80%) and testing set (20%). The testing set was only used to evaluate the final model type in each of our 7 distinct methods we chose. For example, in the Best Subsets Model, the training set was used to determine the best predictors using cross validation (still within the training set). After the best predictors were obtained, then that best model was evaluated on the test set.

Exploratory Data Analysis

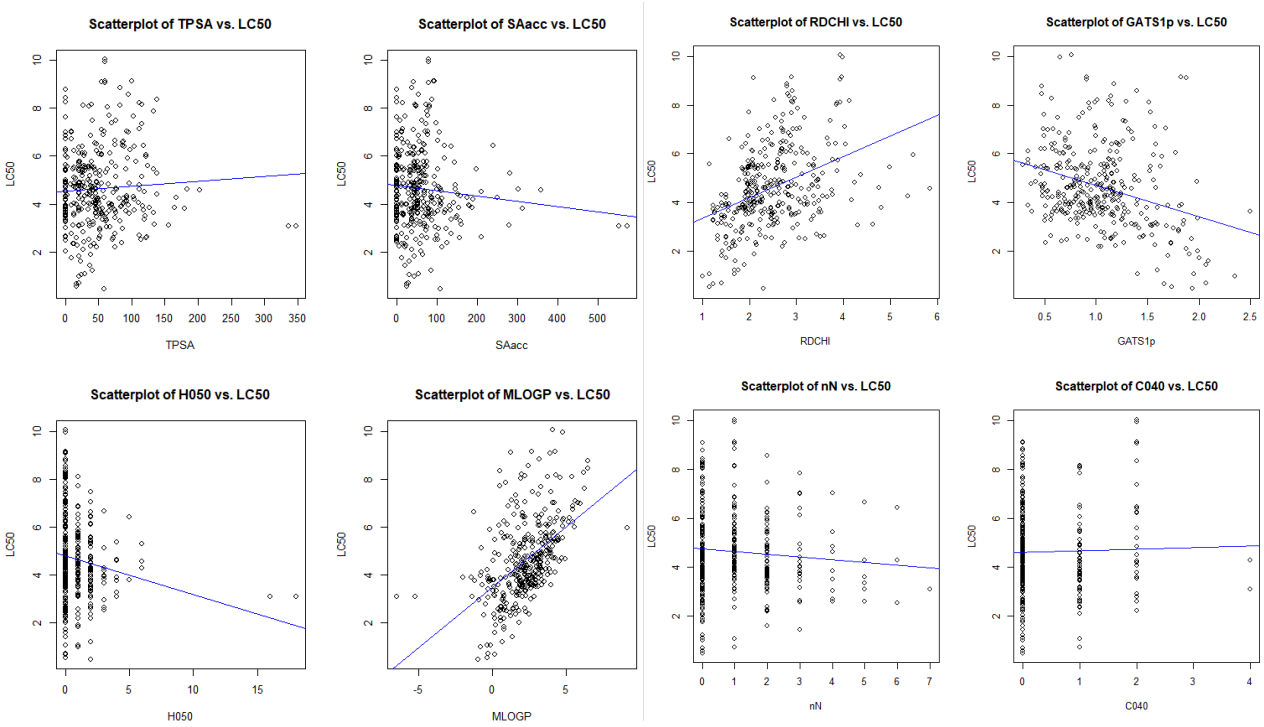
When performing our data analysis, we checked for a few things such as multicollinearity, constant variance, normality, and potential outliers. First, looking at our pairwise correlations, the highest correlation coefficient between a pair of predictors was 0.852 between TPSA and SAacc, as well as 3 separate pairs of predictors having a correlation coefficient greater than 0.6, and 4 other pairs of predictors which had a correlation coefficient between 0.5 and 0.6. We noticed two of our predictors have a VIF value greater than 5 where TPSA had a VIF of 5.7, showing borderline multicollinearity, and SAacc with a VIF of 7.8, showing a much more concerning sign of multicollinearity.



Assessing the Residuals vs. Fitted and Q-Q Residual plot, we determined that constant variance holds as the fitted values are scattered up and below zero with no real pattern, and normality holds even though the tail end of our points leave the line or normality slightly towards the tail end.



Analyzing the Residuals vs. Leverage plot, we noticed there are a couple potential outliers to the right of our leverage flag such as points 343 and 178. Finally, we have provided scatterplots of each predictor vs. LC50 with fitted regression lines.



MLR (8 predictors)

We developed a Multiple Linear Regression model on the Toxicity dataset in order to interpret the relationship between the response which is the LC50 and the predictors which has 8 molecular descriptors. In order to find clear information about the data, we made an MLR on the full training dataset with 80:20 ratio split.

```
Call:
lm(formula = LC50 ~ TPSA + SAacc + H050 + MLOGP + RDCHI + GATS1p +
    nN + C040, data = toxicity_3_1.train)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5556 -0.7900 -0.1196  0.5831  4.8468

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.559159   0.369458   6.927 3.13e-11 ***
TPSA         0.030196   0.004017   7.517 8.24e-13 ***
SAacc       -0.015602   0.003176  -4.912 1.56e-06 ***
H050         0.062558   0.093545   0.669   0.5042
MLOGP        0.478190   0.094331   5.069 7.41e-07 ***
RDCHI        0.461655   0.195653   2.360   0.0190 *
GATS1p      -0.465845   0.246701  -1.888   0.0601 .
nN          -0.299422   0.075282  -3.977 8.95e-05 ***
C040         0.110158   0.133681   0.824   0.4106
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.28 on 271 degrees of freedom
Multiple R-squared:  0.459,     Adjusted R-squared:  0.443
F-statistic: 28.74 on 8 and 271 DF,  p-value: < 2.2e-16
```

Based on the coefficients and the model, we were able to find the MSE on a full model with 8 predictors which was 1.015384 squares units. The adjusted R^2 value was 0.443 which means there is an approximate 44.3% of variation in LC50 as explained by the model. The RSE was 1.28 and the F-statistic was 28.74.

The Fitted Regression of our Model:

$$\hat{Y} = 2.559159 + 0.030196 * TPSA - 0.015602 * SAacc + 0.062558 * H050$$

$$0.478190 * MLOGP + 0.461655 * RDCHI - 0.465845 * GATS1p - 0.299422 * nN +$$

$$0.110158 * C040.$$

The model interpretation of the Multiple Linear Regression:

- Null Hypothesis (H_o):

  $B_{TPSA} = B_{SAacc} = B_{H050} = B_{MLOGP} = B_{RDCHI} = B_{GATS1p} = B_{nN} = B_{C040} = 0$

  (At least one them has no significant linear relationship with LC50)

- Alternative Hypothesis (H_a):

  $B_{TPSA}, B_{SAacc}, B_{H050}, B_{MLOGP}, B_{RDCHI}, B_{GATS1p}, B_{nN}, B_{C040} \neq 0$ (At least one them has a significant linear relationship with LC50)

- F-statistic = 28.74
- Numerator DF = 8, Denominator DF = 271
- P-value = 2.2e-16 < 0.05 (Reject H_o)

There is sufficient evidence to conclude at least one of the predictors has a significant relationship with LC50 (2.2e-16 < 0.05 - Reject H0)

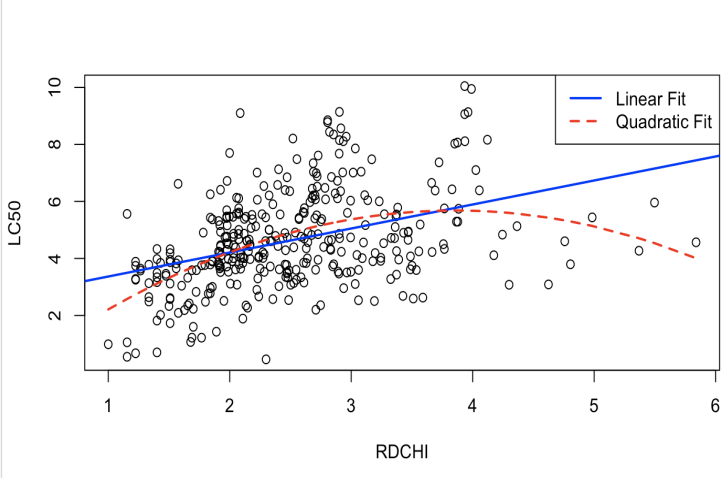
Polynomial Term Model

The polynomial term model is a model that has an added term to the model. The added term is a predictor that is a power of one of the original predictors and is used to capture non-linear relationships. These relationships can be quadratic, cubic, or higher-order powers.

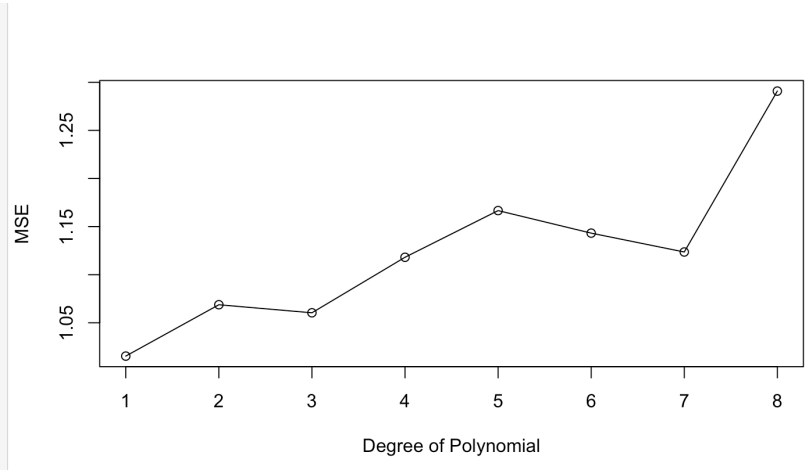
We added polynomial terms, specifically quadratic, to the model one by one to see if any of the polynomial terms were statistically significant to the model when added. Only one polynomial term was statistically significant and that was RDCHI, the topological index with information about molecular size and branching. It is the only predictor which is anywhere near close to having a quadratic/polynomial relationship.

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.5564 -0.7779 -0.1351  0.6010  4.8640

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.307723   0.617702   2.117  0.03498 *
TPSA         0.029227   0.003455   8.458 8.16e-16 ***
SAacc       -0.015751   0.002713  -5.806 1.47e-08 ***
H050         0.105016   0.076614   1.371  0.17137
MLOGP        0.461590   0.082105   5.622 3.94e-08 ***
RDCHI        1.414772   0.450638   3.139  0.00184 **
I(RDCHI^2)  -0.155210   0.070827  -2.191  0.02910 *
GATS1p      -0.465289   0.200362  -2.322  0.02081 *
nN          -0.300722   0.063861  -4.709 3.63e-06 ***
C040         0.005409   0.114243   0.047  0.96226
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.219 on 340 degrees of freedom
Multiple R-squared:  0.4974,    Adjusted R-squared:  0.4841
F-statistic: 37.39 on 9 and 340 DF,  p-value: < 2.2e-16
```

The p-value of I(RDCHI^2) is 0.02910, which is < 0.05, so it is statistically significant. It was the only polynomial term that was statistically significant. The RSE is 1.219, the F-statistic is 37.39, the P-value is < 2.2e-16, and the adjusted R^2 is 0.4841. We also plotted the predictor RDCHI against the response LC50 on a scatter plot and used best fit linear and quadratic lines to see which relationship the data best fitted.



The scatter plot shows that even though I(RDCHI^2) is statistically significant when added to the model, the data still seems to follow closer to a linear relationship and isn't exactly quadratic. This can be seen further when looking at the graph of the test MSE vs the degree of the polynomial for RDCHI. This was done with a validation set approach.

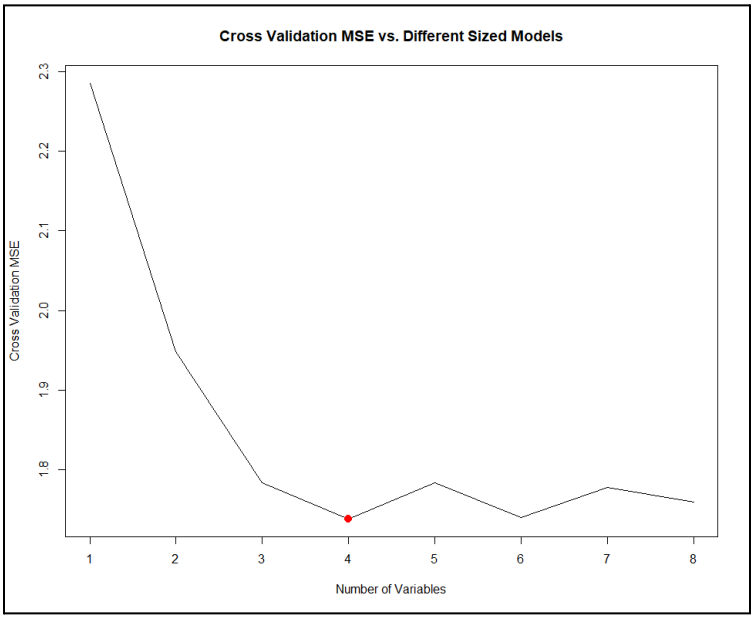


This graph shows the MSE values for each degree of polynomial for the predictor RDCHI. We want a lower test MSE. When adding I(RDCHI^2) to the model, we get a test MSE of 1.068782 squared units of LC50 as shown on the graph at degree of polynomial 2. But we have a lower test MSE when there is no polynomial term for RDCHI, when the degree of the

polynomial is 1. So therefore, it is best not to use a polynomial term in our model, as we get the lowest test MSE when the degree of the polynomial for RDCHI is 1.

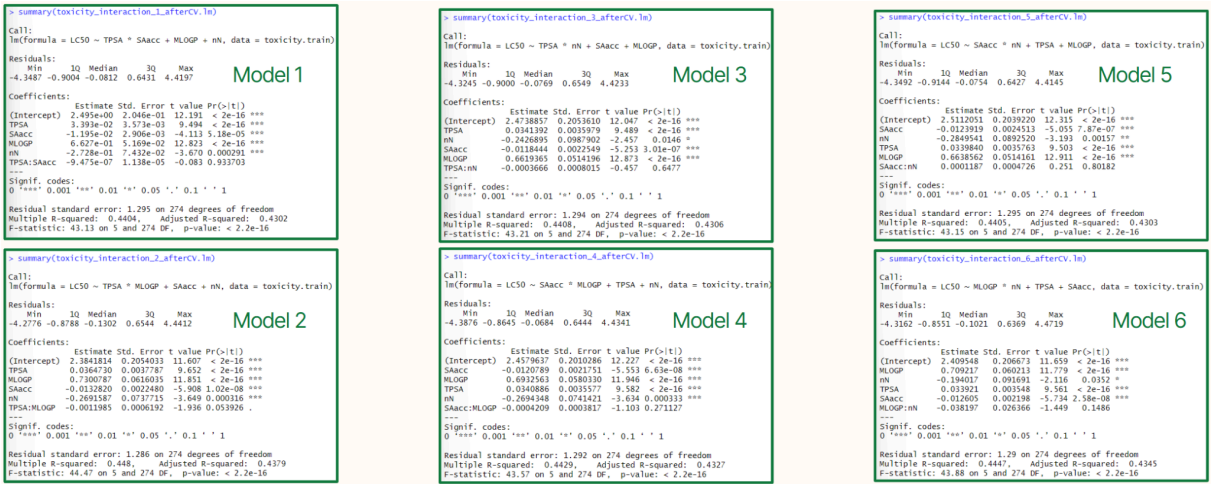
Best Subsets Model

After the MLR with all 8 predictors was used, we decided to perform the best subsets variable selection method. We used 5-fold CV within the training set to conduct this best subsets selection procedure. The graph below shows the mean CV MSE values for each sized model.



As it can be seen the 4-variable model had the lowest mean CV MSE value at 1.738 squared units of LC50. Once tested on the testing set, this best subset 4-variable model obtained a Test MSE of 1.102 squared units of LC50. We do note that this is a higher value than the simple full MLR model. We believe that this can be attributed to the fact that we used only the training set to perform best subsets selection and furthermore used 5-fold cross validation within this training set to evaluate each differently-sized model.

Interaction Term Model



After performing our Best Subsets 5-fold cross validation, we took that model and made 6 new models with 6 different interactions using our 4 remaining predictors. The goal for doing this was to see if adding an interaction term would benefit the Test MSE compared to our original Best Subsets model. First thing we noticed is that none of the interaction terms are significant, while Model 4 is borderline, and Model 4 also provided us our highest adjusted R^2. After calculating each of the Test MSEs though, it seems the addition to an interaction term lowered prediction error. Models 2, 4, 5, and 6 all returned lower Test MSEs than our original Best Subsets model, with Model 6 providing the lowest of 1.042868, which is an improvement of our Best Subsets Test MSE 1.102706. So, although none of our interaction terms were significant, it proved that adding them to our model improves our Best Subsets model in terms of prediction accuracy.

Lasso Regression

The Lasso Regression model is a commonly used shrinkage method that also has the ability to perform variable selection due the fact that a predictor's coefficient value is able to take on the value of zero, unlike the Ridge Regression model. This artifact of the Lasso method was the primary reason for choosing it as part of our analysis.

When running the model, 5-fold CV was used to determine the best value of lambda, a hyperparameter used to control the strength of the Lasso penalty term. This value was found to be 0.0184.

Once the optimal lambda value was found, the Lasso Regression model was fit on the entire training set and the following coefficients values were found for each predictor:

(Intercept)	TPSA	SAacc	H050	MLOGP	RDCHI	GATS1p	nN	C040
2.685	0.0261	-0.0121	0.00	0.456	0.466	-0.546	-0.234	-0.018

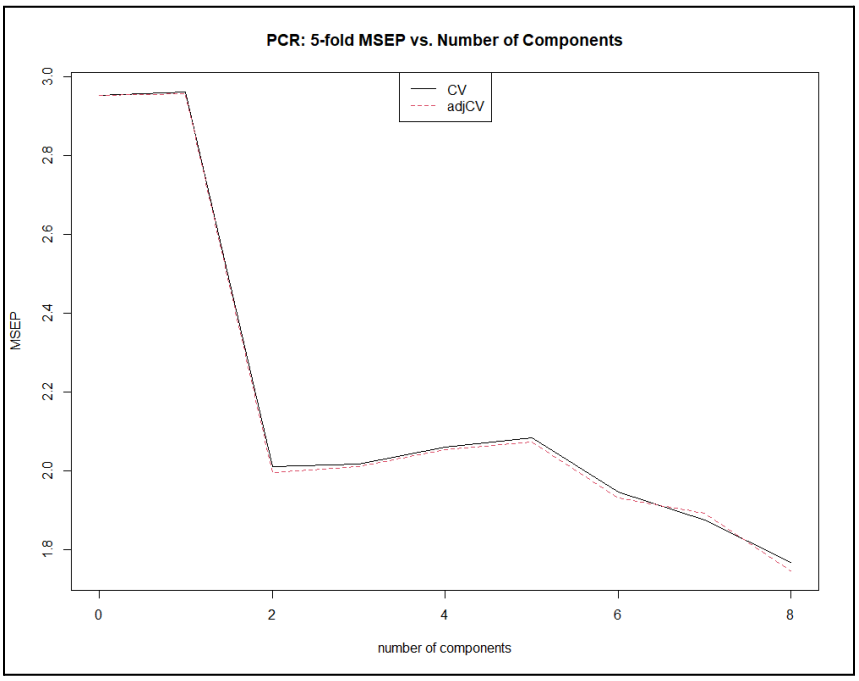
As it can be seen, the H050 (number of hydrogen atoms bonded to heteroatoms) predictor was forced out of the model.

Lastly, when evaluating on the testing set, this model produced a Test MSE of 1.026 squared units of LC50.

Principal Component Regression (PCR)

The Principal Component Regression model is considered a dimension-reduction model. It is based on a popular dimension-reducing visualization method called Principal Components Analysis (PCA). Principal Component Regression works based on the idea that a smaller number of predictors (now called principal components) will be able to explain more of the variability in the data and the relationship with the response variable. This method relies on the critical assumption that the directions in which the original predictor variables show the most variation are also the directions associated with the response variable.

When running the model, the training set was first standardized and 5-fold CV was used to determine the best number of components to use. The MSE of prediction value can be seen in the graph below.



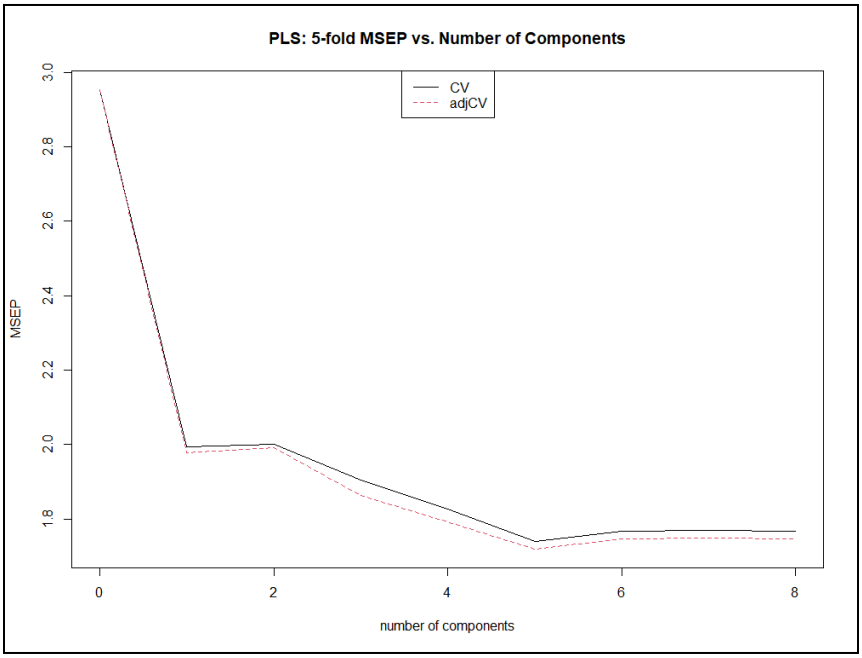
It can be seen that the MSE of prediction value was the lowest at 8 components which indicates that this dataset is not benefiting from the Principal Component dimension reduction technique. Therefore, the 8-component Principal Component Regression model has the exact same coefficients as the MLR model with all 8 predictors.

Lastly, when evaluated on the testing set, the Test MSE value was 1.015 squared units of LC50.

Partial Least Squares Regression (PLS)

After running the Principal Component Regression model and seeing unsatisfactory results, we decided to try the Partial Least Squares Regression model as another dimension-reduction technique. PLS is a supervised alternative to PCR as it obtains transformed predictors by making use of the response variable, unlike PCR.

When running the model, the training set predictor values were standardized first and 5-fold CV was once again used to determine the best number of components. As it can be seen in the graph comparing the 5-fold MSEP to the number of components, the optimal number of components was 5. This is different from the number we obtained in the PCR model and it is interesting to see that the inclusion of the response variable in calculating the new components does yield a different result.



Furthermore, as we can see from the PLS summary R code output, the variation explained in the response variable (within the training set) with 5 components is 45.86%, which is not the highest, but close. The 7 and 8 component models have the highest at 45.90%.

```
Data:   X dimension: 280 8
        Y dimension: 280 1
Fit method: kernelpls
Number of components considered: 8

VALIDATION: RMSEP
Cross-validated using 5 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
CV           1.719    1.412    1.414    1.381    1.352    1.320    1.329    1.331    1.329
adjCV        1.719    1.407    1.411    1.366    1.339    1.311    1.322    1.322    1.322

TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X    19.55    60.48    67.38    73.64    78.87    90.33    97.36    100.0
y    35.13    37.09    42.83    44.60    45.86    45.89    45.90     45.9
```

Lastly, when evaluated on the testing set, the PLS model achieved a Test MSE of 1.000 squared units of LC50, notably lower than the PCR model.

Conclusion

The test MSE (in squared units of LC50) varied across the different models. Multiple Linear Regression (MLR) had a test MSE of 1.015384, the Best Subsets model had a test MSE of 1.102706, the Interaction Term model had a test MSE of 1.042868, the Polynomial Term model had a test MSE of 1.068782, Lasso Regression had a test MSE of 1.026755, the PCR model had a test MSE of 1.015384, same as MLR, and the PLS model had a test MSE of 1.000126.

In conclusion, our lowest test MSE model was Partial Least Squares (PLS) at 1.000126. Our chosen model for interpretability is Multiple Linear Regression (MLR) (all 8 variables). The MLR model had the 2nd lowest test MSE.

Looking back at the MLR model coefficients, we found that the TPSA, H050, MLOGP, RDCHI, and C040 predictors had positive coefficients with MLOGP being the highest at +0.478190. This indicates that keeping all other variables constant, a one unit increase in MLOGP results in the highest increase (0.478) in LC50 value out of all the 8 predictors. Therefore, chemicals with higher MLOGP values can be said to be less toxic. Conversely, we found that the SAacc, GATS1p and nN predictors had negative coefficients with the GATS1p coefficient being the most negative at -0.465845. Similarly, this indicates that keeping all other variables constant, a one unit increase in GATS1p results in the greatest decrease (0.465) in LC50 value out of all the 8 predictors. Thus, chemicals with higher GATS1p values can be said to be more toxic.

This analysis serves as an excellent starting point for toxicologists to further study molecular descriptors and their effects on the toxicity of various chemicals.

References

1. James, Gareth, et al. An Introduction to Statistical Learning: With Applications in R. 2nd ed., Springer, 2021.

2. Cornell University, Environment, Health and Safety. "Laboratory Safety Manual." Cornell University Environment, Health and Safety, Accessed 6 May 2025, https://ehs.cornell.edu/book/export/html/1390