Imperial College London

COURSEWORK

IMPERIAL COLLEGE LONDON

DEPARTMENT OF MATHEMATICS

Statistical Learning

1 Introduction

In this report we will investigate the use of regression methods for QSAR modelling of the relationship between acute water toxicity and the fathead minnow (Pimephales promelas) fish. Quantitative Structure-Activity Relationship models are, in our framework, regressions models with physicochemical properties of molecules as its predictors with biological activity as the response.

As environmental issues increasingly dominate headlines and affect our daily lives, investigating the chemical effects on aquatic environments is more important than ever. Pollution from human activities is a pressing concern, and understanding its impact on our waterways is crucial for protecting the health of our planet and all its inhabitants. By studying and addressing the chemical effects on aquatic environments, we can work towards a sustainable future for ourselves and future generations.

This report will be done from the perspective of an environmental researcher working in collaboration with an environmental consultancy company tasked with assessing potential environmental risks associated with the use of new chemicals in an aquatic ecosystem.

Using a dataset on QSAR fish toxicity[1] of 6 molecular descriptors of 908 chemicals alongside the lethal concentration 50 (LC50) value as our response variable. The data includes information ranging from atom-type counts to 2D autocorrelations and will be measuring the toxicity by the amount of substance required to kill 50% of test animals. We are expected to determine the lethal concentration of a chemical, provided its molecular descriptors, for the fathead minnow fish; which will provide insight into the potential risks associated with exposure to chemicals in an aquatic environment.

2 Data Exploration and Preparation

We will explore different regression models for our problem: evaluating how our models perform and what conclusions can be made. We will evaluate our model performs by measuring the mean squared error (MSE) and analysing the residual plots. Before we start it is good practice to perform exploratory data analysis and prepare our data. We start by checking that our data does not include any NA values, variables are numerical and categorical variables are changed to factors using `as.factor()`. Next we will examine if there's any correlation and multicollinearity in our data there are many reasons for doing so such as the effects on the condition number and statistical inferences in our linear regression model, shrinkage of highly correlated features in LASSO regression and using principal component regression to reduce correlation between variables. We use `ggpairs()` to plot the correlation and distribution of our variables looking at Figure 1 we see that there's a moderately strong positive relationship between the response variable and MLOGP and moderately weak negative relationship between `GATS1i` and `MLOGP`. We will then split our data into train and test sets so we are able to evaluate our models. After splitting the data we will examine the data to see if there are any outliers in our training set using box-plots and remove these outliers (figure omitted due to limit). It is important for us to standardise our data, in particular for LASSO and ridge regression so that our coefficients are comparable. We will standardise our training data then use the mean and standard deviation of our training set to 'standardise' our test set. After the exploratory data analysis, based on the relatively low number of variables and low degree of multicollinearity it will not be necessary to use principal component regression. We will therefore be using linear, LASSO and ridge regression in our analysis.

3 Methodology and Analysis

Once our data is prepared we will first fit a linear regression model to our data using the **R** command `lm()`. Upon summarising our linear model we see that *NdsCH* and *NdssC* have high P-values suggesting that these variables are too highly correlated with other variables to provide an independent explanation and therefore we will remove them as part of our variable selection step. Additionally,

using Cook's distance we will identify any further outliers in our linear regression model in the interest of a robust model. We can now re-train our model without the insignificant variables and form predictions using `predict()`.

Now we have fitted a linear regression model to our training data let us run through some diagnostic plots to evaluate whether we can go forward with our regression methods. Let us first analyse Figure 2 the Residuals vs Fitted plot, we see that the residuals are scattered roughly around the Residuals = 0 line, suggesting that a linear relationship is plausible. We can roughly see horizontal bands suggesting our residuals exhibit constant variance. Even after removing outliers using box-plots and Cook's distance, we see some residuals that stand out which could suggest outliers. Next we analyse the Q-Q plot of the residuals in Figure 3 and notice some weak heavy-tailed behaviour suggesting that our residuals are similar to that of being normally distributed. Furthermore, this behaviour could be due to the model not capturing the full variance or additional outliers in the data. The last diagnostic plot we will examine is Figure 4, the Scale-Location plot we see the points randomly scattered around the red line, however, the line does exhibit weak curvature which could be a result of a need to transform a predictor variable or that a new variable is required altogether, but the weak curvature does suggest that the assumption of homoscedasticity is valid to a degree. After analysis of our diagnostic plot we can conclude to a degree that many of our assumptions for a linear model and linear regression are met, that is we have a linear relationship, normally distributed residuals and homoscedasticity. We can now move onto the rest of our regression methods.

Next we will use ridge regression to model our data, using cross-validation to determine the optimal value of our penalty hyper-parameter then build our model with the optimal hyper-parameter. To do so we build our model on the training data using `glmnet()` then perform cross-validation to find the optimal value of our penalty using `cv.glmnet`. For our data, our optimal penalty is $\lambda = 0.094$, using this optimal penalty we form predictions on our test data.

The next regression model we will explore is LASSO regression where, similarly to ridge, we will use cross-validation to determine the optimal value of our penalty hyper-parameter. The difference with LASSO is the added bonus of being able to shrink some coefficients to 0. These are advantageous for us as it performs variable selection which can lead to less variance in our model improving our predictions but also increasing the robustness and interpretability of our model. Using cross-validation, we obtain an optimal value of $\lambda = 0.00388$. Interestingly, our model does not shrink our *NdssC* or *NdsCH* variable unlike in our linear model where we removed the variables due to the high P-values.

4 Conclusion

Now that we have built our three regression models shown that linear models are suitable for our dataset we can evaluate how well they have performed. Figure 5 shows the Residuals vs True Y Test Values for all three of our models and we notice the points seem to be randomly distributed suggesting our models are reasonable for the data and that we don't see extreme variance in the plot suggesting that the models fit the data reasonably too. Using the MSE as our measure of performance we have that the linear model provides the highest MSE of 0.968 then LASSO with 0.94 with the best MSE score of 0.937 coming from the ridge regression model. This suggests that the ridge regression model gives a better fit to the model as opposed to the linear and LASSO models. The bonus of using a ridge regression model is the robustness and stability it provides by shrinking correlated variables towards zero as opposed to selecting one in the case of LASSO. In relation to the task objective, a ridge regression model is better suited for prediction when the goal is to minimise the overall error as opposed to linear and LASSO regression and therefore can use ridge regression to model and make more predictions for our data.

A References

[1] - M. Cassotti, D. Ballabio, R. Todeschini, V. Consonni. A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (Pimephales promelas), SAR and QSAR in Environ-

mental Research (2015), 26, 217-243; doi: 10.1080/1062936X.2015.1018938

B Plots

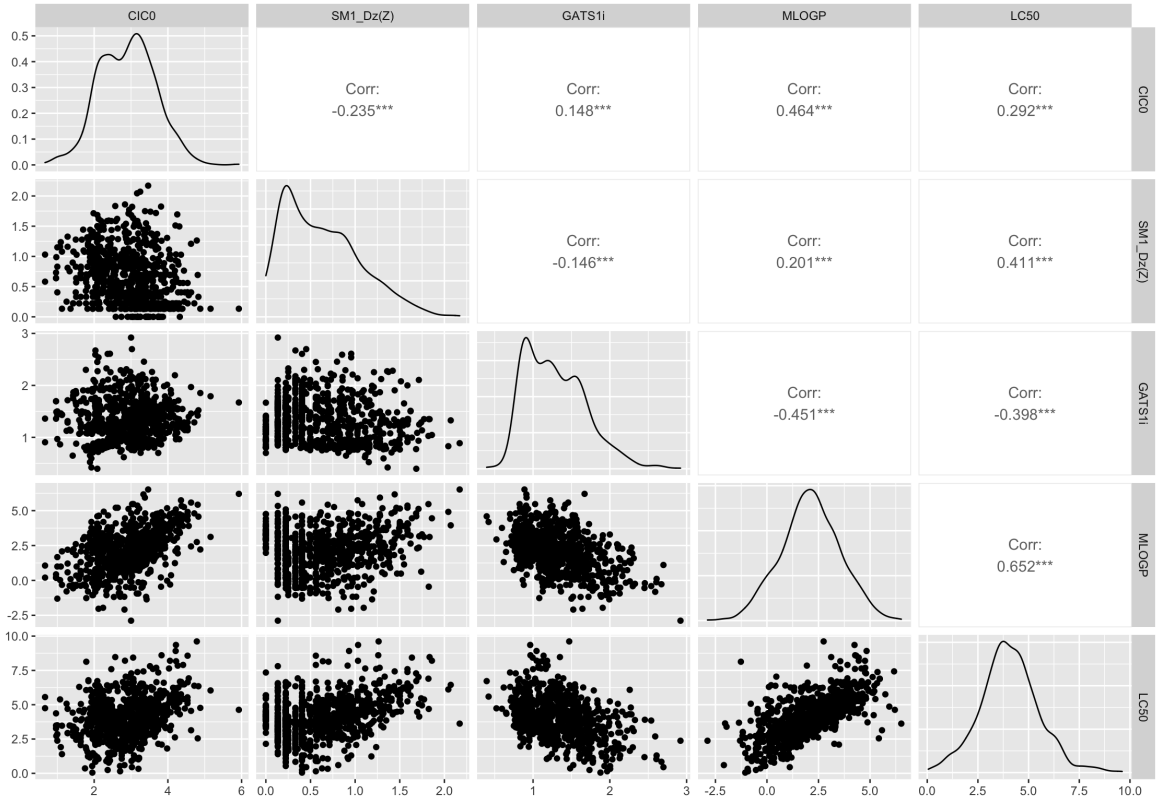

Figure 1: Correlation between variables

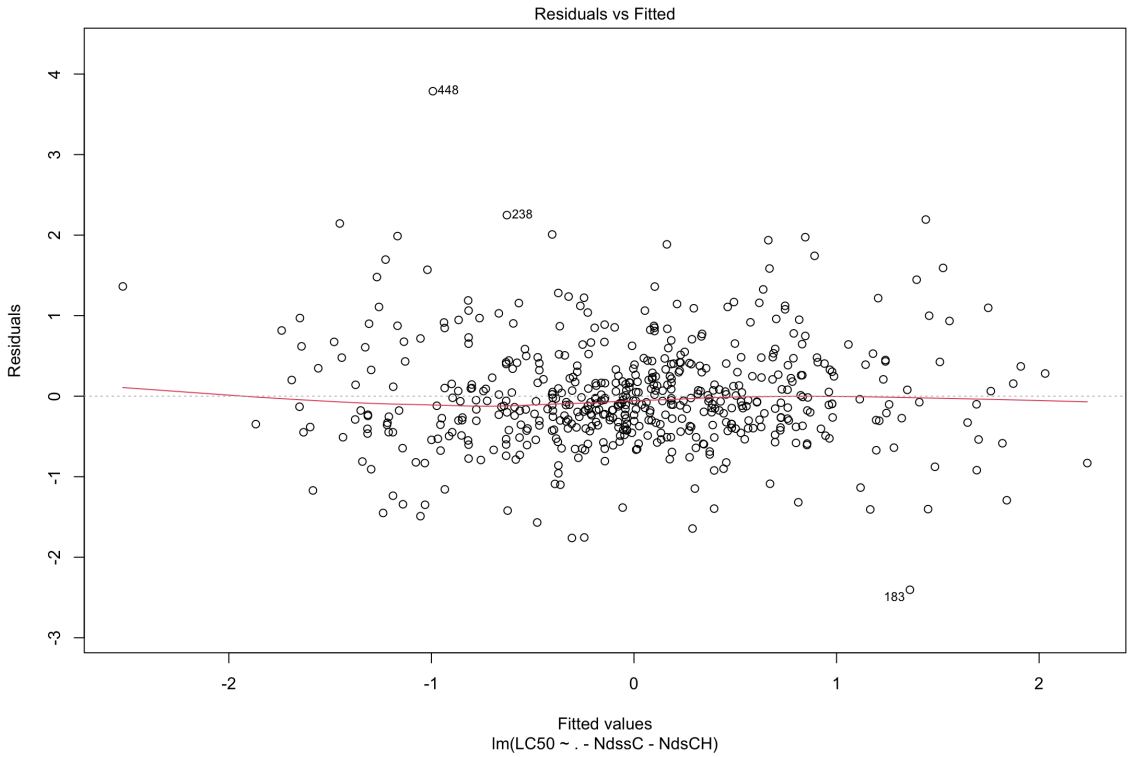

Figure 2: Residuals vs fitted values of linear regression model

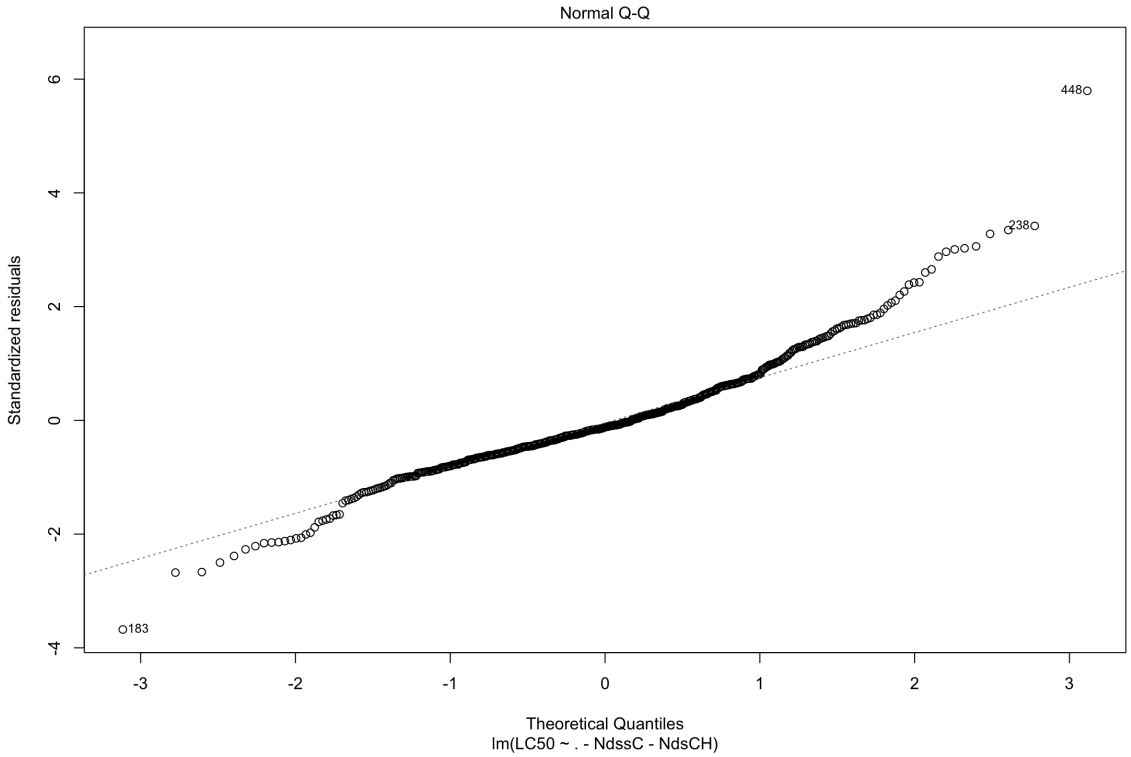

Figure 3: Q-Q plot of residuals of linear regression model

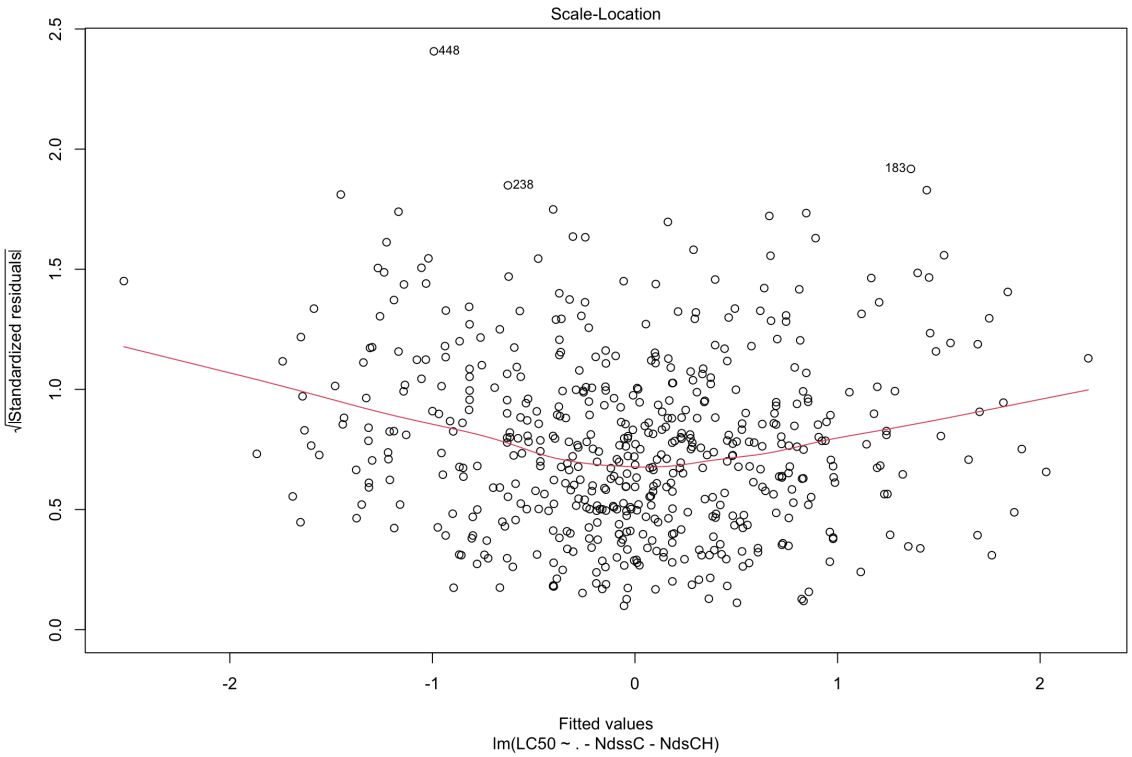

Figure 4: Scale-location plot of linear regression model

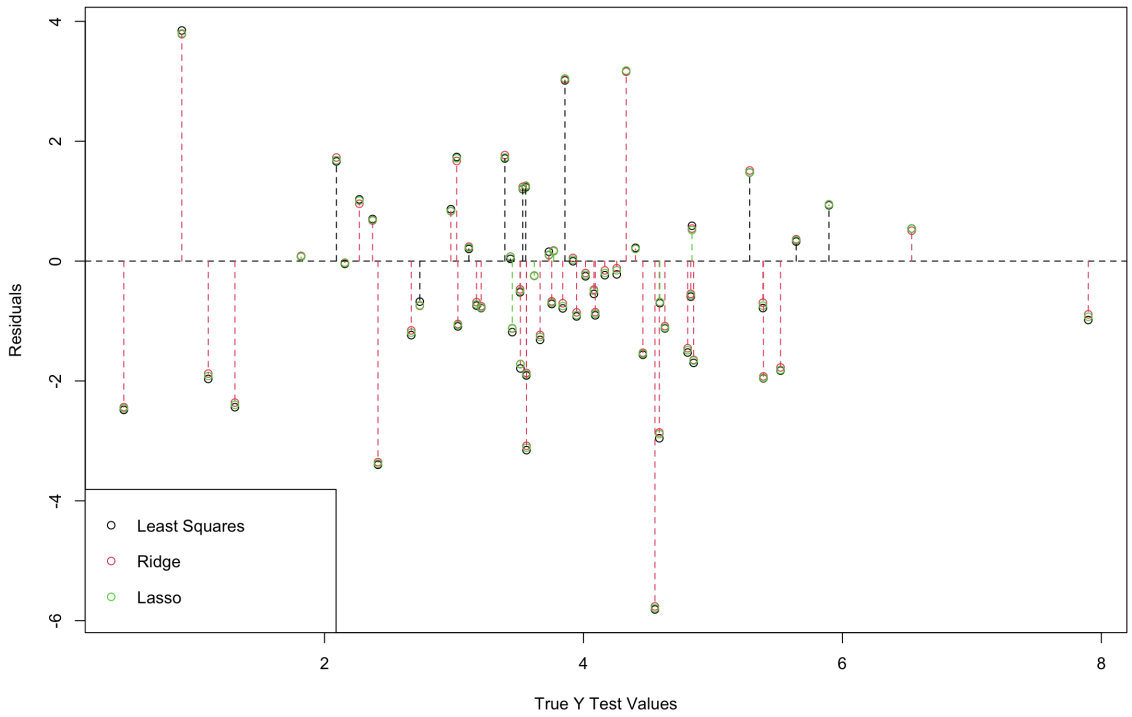

Figure 5: Residuals vs true Y test values of all regression models