Analysis of Dissolved Oxygen to Study Water Quality

By

Ayanabha Jana (18BCE1044)

Saswat Panda (18BCE1281)

Anand Bhalotia (18BCE1121)

# *Abstract*

In this experiment, we have taken a sample dataset with multiple variables and used different regression techniques to come up with appropriate predictions. The main aim of the experiment is to find which regression is most suitable and gives the best prediction results for the sample dataset.

We have taken a dataset that contains combined and cleaned data for historical water quality in certain locations of India. Pollutant measures in each column are the average values measured over a period of time. The dataset contains eight variables that can be used for regression. They are Temp, D.O.(Dissolved Oxygen) (mg/l), PH, CONDUCTIVITY ( mhos/cm), B.O.D.(Biological Oxygen Demand) (mg/l), NITRATE (mg/l), FECAL COLIFORM (MPN/100ml), TOTAL COLIFORM (MPN/100ml)Mean. We have taken D.O. and Nitrate/Nitrite concentration as a measure of water pollution. Lesser the D.O. value, more is the water pollution. Also more the Nitrate concentration, more is the water pollution.

The regression techniques used in the experiment are Linear Regression, Quadratic Regression, Stepwise Regression, Ridge Regression, Lasso Regression and lastly, Elastic Net Regression. The document contains the results of predictions made through different regression types which are later compared. Through these comparisons, we have found the most suitable regression technique which gives the closest results to the actual dataset. The document also contains graphs to visualize the regression techniques.

The regression techniques used are coded in R language. We have used the software "Rstudio" to run the code and get results.

# *Introduction*

In a statistical model, often we are interested in finding the relationship between some given variables. Consider the situation in which you are presented with some data points. Regression analysis comes in handy in deciding how the data points can be related in some assumed manner, be it a linear relation or a quadratic curve. The process of Regression is performed so that we can determine confidently which factors are the most important, what all factors can be ignored, and one factor is influenced by the other. First of all what we need to do is that we select a dependent variable and hypothesize its relation with some factors on which the variable may depend. The objective is to find a line or curve that fits the maximum of the given data points or one that is at a minimum distance from all the data points. Once we frame an equation for the given problem, it can be used to predict the value of the dependent variable for some given values of the independent variables. The various types of regression are as follows:

* *Linear Regression*

It is the most basic form of regression in which the variables are assumed to have a linear dependency i.e. the power of the exponent is 1. This dependency can either be between two variables or multiple variables (multiple linear regression). Here, let us say that the dependent variable is ‘y’ and independent variable is ‘x’, so we frame a linear equation of the form:

y=ax+b

Now, let’s apply summation on both sides of the equation:

∑y=a∑x+nb

Once done we can apply this operation to the set of data points we have to calculate the values of constants ‘a’ and ‘b’.

* *Polynomial Regression*

In this form of regression, we modify the linear regression model to include the independent variable to its nth power. Let us take the example of a quadratic polynomial and frame the required equation:

y=ax^2+bx+c

Now, we apply summation on both sides:

∑y=a∑x^2+b∑x+nc

The next step is the same as the previous regression model. Substituting the data points for the given operation, we calculate the constants ‘a’, ‘b’ and ‘c’.

* *Stepwise Regression*

Consider a multiple regression model in which you correlate the dependent variable with all the other independent variables, for example:

y=ax1+bx2+cx3+dx4+ex5

However, y may not depend on all the variables in this case i.e. the coefficient of a variable is so small that it is almost negligible. This is where stepwise regression comes into play. It consists of a series of steps, where in each step each x-variable is evaluated using the akaike information criterion (AIC) and accordingly choosing a variable that satisfies the stipulated criterion (forward selection) or removing a variable that least satisfies the criterion (backward elimination). The AIC is an estimator of the relative quality of statistical data. If ‘k’ is the number of estimated parameters and ‘L’ is the maximum value of the joint probability distribution of a random sample, then the AIC value would be:

AIC=2k - 2 ln (L)

* *Ridge Regression*

SSERidge = Σ(Y - Ŷ)2 + λ Σβ2

The term with lambda is known as L1 penalty factor. Ridge Regression is used to lessen the errors that were caused due to multicollinearity due to the inclusion of several parameters.

* *Lasso Regression*

SSELasso = Σ(Y - Ŷ)2 + λ ΣIβI

The term with lambda is known as L2 penalty factor. Lasso regression’s full form is ‘least absolute shrinkage and selection operator’. The Lasso Regression uses the concept of shrinkage which means that the data values are move towards the central tendency of the data.

* *Elastic net Regression*

SSEEN = Σ(Y - Ŷ)2 + λ [(1-α)Σβ2 + αΣIβI]

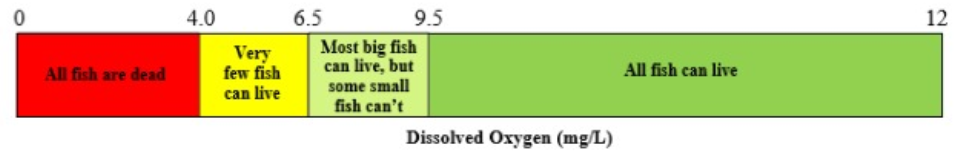
The L1 and L2 penalty factor are combined together to make an efficient regression known as Elastic net regression. This further reduces the error. Note that we can derive

Ridge regression by putting α=0

Lasso regression by putting α=1

*Proposed Work*

Given a dataset pertaining to the levels of water nutrients and properties such as temperature and pH, we are applying the six types of regression in R to draw a clear dependency of the dissolved oxygen in water (D.O.) with the other factors. In linear and polynomial regression, we are also mapping the dependency of conductivity on pH and nitrate on fecal and total coliform. Next in stepwise regression, we have considered D.O. as the dependent variable and then we have iteratively selected independent variables based on Akaike information criterion or AIC (forward selection) and eliminated independent variables to keep the strongest impacting factors (backward selection) to find the best fit line for D.O. In elastic-net regression, we aim at reducing the error incurred while calculating the predictions by bifurcating the data set into test and train data. Next we choose the multiple-variable dependency, in this case, D.O. and minimize the coefficients to reduce the error, once by taking the square of the penalty factors (ridge regression) and once by taking only the penalty factors (lasso regression). Next we have taken one of the rows from the dataset and applied it to the corresponding regression models to find the best fit for the given dataset. This has given us a particular value of D.O. Now, as we have discussed, D.O. acts as an indicator for water quality and the various levels of D.O. are as follows for supporting a marine ecosystem:



By observing the range, we can see assess the water quality measures between the actual data and the predicted data of the best regression model.

# 

# *Experimental Analysis*

1. Linear, Multiple and Quadratic Regression

First we have included the intended data set in report\_r using read.csv. Next, for our calculation requirements, we have only considered the columns 4 to 11 and all the rows and omitted the null values using na.omit.

Now, our variable of importance is DO. In the first model labelled lm1, we have shown DO as being dependent on BOD and temperature. Next, we have predicted the value of DO for temperature=2.9 and BOD=1.5 using predict(). ggPredict() is used to visualize the given model based on the given data points and finds the best fit line. Summary(lm1) gives the summary of the entire data model. We also make a second model labelled as lm\_do where DO is mapped to all other variables (DO~.). Here we predict DO by putting values for all the other variables.

Next to study the conductivity, we create a linear model lm2 where the independent variable is PH. ggplot() is a function which is used to plot single linear models along with their respective data points.

Next we use the concept of polynomial regression to check the dependency of conductivity on PH. We do not go ahead of the quadratic form because we observed for our dataset on increasing the value of n in x^n the residual error increases and adjusted R square decreases. We can easily apply the lm() function in this case by taking PH^2 as the variable x1 and predicting the value of conductivity on the basis of that, using ggplot() in the same manner as the previous model.

We also create a prediction model for nitrate by including f\_coliform and t\_coliform as the independent variables and predicting for values 11 and 27 using the predict() function.

2. Stepwise Regression(Also, forward selection and backward elimination)

In stepwise regression, we start by including the data set by read.csv(), including only columns 4 to 11 and omitting null values. All we have to do is include the lm() part inside the step() function and mention the direction of analysis.

When we mention direction=”both”, the model starts with all the other variables in the equation and can both delete and include variables according to the AIC values. The prediction is the same as for other linear models.

The direction=”forward” indicates the forward selection procedure, where we start off with DO~1 and keep on adding the other variables by checking the AIC value after each iteration. When we print this model, we get the AIC values for all the variables involved.

The direction=”backward” indicates the backward elimination process in which we start off with DO~. and keep on eliminating variables according to AIC.

The predict() function gives the value corresponding to a given tuple and this can be used to find the error of prediction.

3. Ridge, Lasso, and Elastic Net Regression

Above we applied various regression and then we determined which one is the best regression for predicting DO, Nitrate and Conductivity.

Ridge, Lasso and Elastic net works on multivariate regression and try to reduce the error. In our experiment we apply these regression to predict DO and Nitrate and to reduce error in their predictions. We are not applying these regressions on conductivity because it is dependent only on ph.

The steps performed for the regression is same.

* Set.seed is used to always select the same set of random values.
* Partitioning the data set into two parts test and train. Test contains small amount of data which is required for finding if predictions performed by using regression on the train data is correct or not.
* For Ridge and Lasso we cv.glmnet to find the suitable lambda that will minimize the cross-validation prediction error. In Elastic net cv.glmnet is used to find the best tune length alpha(α) and lambda(λ).
* To check predictions we compare predicted value with the test data. We find the RMSE and R Square and then accordingly decide which model is the best.

# 

# *Results*

## Linear, Multiple and Quadratic Regression

## 

## 

## 

#DO ~ Temp + BOD

The multiple R-squared value comes out to 0.0003171 which is very close to 0 and hence indicates that this relationship is not strong enough or in other words, the correlation is not 100% linear. The adjusted R-squared value is -0.0001898 which indicates insignificance of the independent variables. The predicted value of DO comes out to be 6.533933. From the first graph, the best fit line is y=0.04x+5.45 when BOD=0.

#DO depends on all other variables

The multiple R-squared value comes out to 0.1577 which exhibits a better relationship than the previous model. The adjusted R-squared value is 0.1547, a positive value which is a better indicator for fit. The predicted value of DO comes out to be 6.655756

#Conductivity ~ PH

The multiple R-squared value is 0.0003171 and the adjusted R-squared value is -0.0001898, which is the same as the first linear model, indicating that the two models are not the best fit for the given set of independent variables. The conductivity value comes out to be 1749.018. The second graph indicates that all the data points are close to the origin, meaning that there is not much difference in the conductivity for a range of PH values.

#Conductivity ~ PH^2 +PH

The multiple R-squared value is 0.0005339, while the adjusted R-squared value is -0.0004803, both of which indicates a slightly better fit than the linear regression model. The conductivity predicted in this case is 1752.713. As it is clear from the third graph, there is minimal change in conductivity values.

#Nitrate ~ F\_Coliform + T\_Coliform

Lastly for the relation of nitrate, the multiple R-squared value is 1.29e-06 which indicates a medium relation but the adjusted R-squared value is negative (-0.001013) which also indicates a medium-level dependency. The predicted value of nitrate from the model is 1.446844.

## Stepwise Regression(Also, forward selection and backward elimination)

Here in stepwise regression, we have considered only one dependent variable DO and created the model as (DO~.), where dot means all other variables. Let us see the results we get by applying the regression in all the three directions.

*Both Directions*

We can see for both the directions, the AIC value comes out to be 1392.94 and given all the dependencies, in the final regression model, the fecal coliform is eliminated.

After applying the model we predicted the DO for a particular row of data Temp=29.2,PH=6.3,CONDUCTIVITY=100,BOD=1.5,NITRATE=0.1,F\_COLIFORM=7942, T\_COLIFORM=13575

DO comes out to be 6.655898.

*Forward Selection*

In this case, we start off with DO~1 and finally all other variables except fecal coliform is included. The AIC value is same as before i.e. 1392.94

After applying the model we predicted the DO for a particular row of data Temp=29.2,PH=6.3,CONDUCTIVITY=100,BOD=1.5,NITRATE=0.1,F\_COLIFORM=7942, T\_COLIFORM=13575

The result is the same i.e. 6.655898

*Backward elimination*

In backward elimination, we get the same regression model with fecal coliform eliminated. The AIC value is 1392.94.

After applying the model we predicted the DO for a particular row of data Temp=29.2,PH=6.3,CONDUCTIVITY=100,BOD=1.5,NITRATE=0.1,F\_COLIFORM=7942, T\_COLIFORM=13575

The predicted DO value is 6.655898.

The predicted value that we got by applying all three procedures instepwise regression on DO and all the variable factors comes out to be the same.

The predicted value 6.655898 is found very close to the actual value.

## Ridge, Lasso, and Elastic Net Regression

Actual value for DO for given data is 7.2

#DO with BOD and temp

We applied ridge model and got the predicted value to be

6.543149

We applied Lasso model and predicted the result to be

6.558347

We applied Elastic net model and predicted the result to be

6.538152

From above we found Lasso was most suitable because the predicted value was close to actual value.

Then we compared the three models further using the test data.

> #comparision

> data.frame(RMSE = RMSE(result\_rid, test$DO),Rsquare = R2(result\_rid, test$DO))

RMSE Rsquare

1 1.444762 0.1365825

> data.frame(RMSE = RMSE(result\_las, test$DO),Rsquare = R2(result\_las, test$DO))

RMSE Rsquare

1 1.435013 0.1400156

> data.frame(RMSE = RMSE(result\_elas, test$DO),Rsquare = R2(result\_elas, test$DO))

RMSE Rsquare

1 1.43582 0.1380681

We noted RMSE(root mean squared error) was minimum for Lasso as well as R-squared value was minimum for lasso.

#DO with all variables

On checking the predictions of these regression model that includes all the variables we found that the predicted values were not too close to the actual value.

#Nitrate

> #comparison

> data.frame(RMSE = RMSE(result\_rid, test$NITRATE),Rsquare = R2(result\_rid, test$NITRATE))

RMSE Rsquare

1 7.627076 6.829169e-07

> data.frame(RMSE = RMSE(result\_las, test$NITRATE),Rsquare = R2(result\_las, test$NITRATE))

RMSE Rsquare

1 7.627138 7.008429e-07

> data.frame(RMSE = RMSE(result\_elas, test$NITRATE), R-square = R2(result\_elas, test$NITRATE))

RMSE Rsquare

1 7.627405 6.968522e-07

We noted Rsquare was minimum for lasso.

Predicted value that we got for the best is 6.292096

But the actual value is 0.1 so these regressions are not preferred for nitrate estimation.

*Conclusion and Future Word*

**DO**

For predicting DO using BOD and Temperature best is LASSO(6.558347)

Overall best is step wise (6.655898)

By analysing and applying regression we found although theoretically DO(dissolved oxygen) depends upon BOD(biological oxygen demand) and temperature but practically it depends on a lot of other factors.

**Nitrate**

Multiple Linear Regression is best(1.446844)

**Ph**

Best is Linear Regression.(Since adjusted R square is more and residual standard error is less)

From the inferences made, we can conclude that the D.O., when considered overall fall in the range of 6.5 to 9.5, which indicates small levels of surface level pollution and is sustainable to support large to medium aquatic populations. The level of nitrate is low enough to harm any aquatic species but can lead to medium levels of eutrophication.

The inference and conclusion drawn from the data set can be used to reduce pollution level in water. We can control the level of pollutants in water by controlling various factors such as nitrate concentration. By controlling pollutants, we can maintain optimum levels of ph as well as dissolved oxygen and hence protect a lot of fishes as well as the marine life as a whole from becoming extinct **in future**. Also, this study will be helpful to determine the concentration of nitrate in water if we know the TOTAL COLIFORM in advance and thereby decide whether the given sample of water can be used for drinking or not.

# *References*

[1] <https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/>

[2] <https://rpubs.com/bigcat/258879>

[3] <http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net/>