|  |  |
| --- | --- |
| Algorithm: Poisson Regression | |
| USN :1MS17CS148 | NAME :SATHVIK B |
| USN :1MS17CS143 | NAME : SATHVIK K P |

**Description of the Algorithm:**

Poisson regression is similar to regular multiple regression except that the dependent (*Y*) variable is an observed count that follows the Poisson distribution. Thus, the possible values of *Y* are the nonnegative integers: 0, 1, 2, 3, and so on. It is assumed that large counts are rare. Hence, Poisson regression is similar to logistic regression, which also has a discrete response variable. However, the response is not limited to specific values as it is in logistic regression.

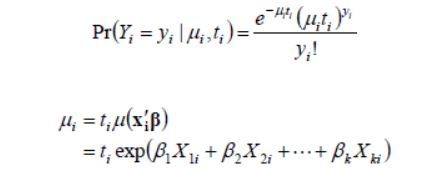
One example of an appropriate application of Poisson regression is a study of how the colony counts of bacteria are related to various environmental conditions and dilutions. Another example is the number of failures for a certain machine at various operating conditions. Still another example is vital statistics concerning infant mortality or cancer incidence among groups with different demographics.

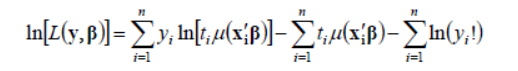
Poisson Distribution :



Here, u = mean, t = time, y = value of random variable. The Poisson distribution has the property that its mean and variance are equal.

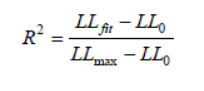
Poisson Regression Model :

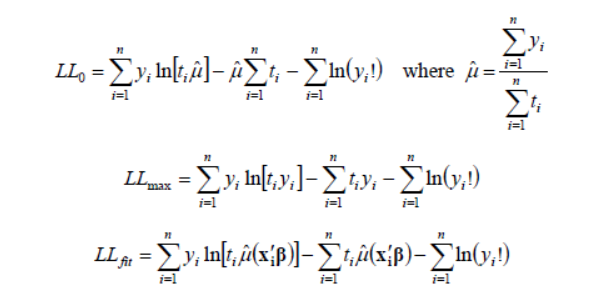


Maximum Likelihood estimation is used for estimating regression co-efficients. The logarithm of this function is given by :   


Pseudo R squared measures: The *R*-squared statistic does not extend to Poisson regression models. Various pseudo *R*-squared tests have been proposed. the most popular pseudo *R*-squared measure is

function of the log-likelihoods of three models





**Algorithm Pseudocode:**

1. Make sure the target values are integer counts.

2. Include only those variables which influence the observed counts (Choosing only required features)

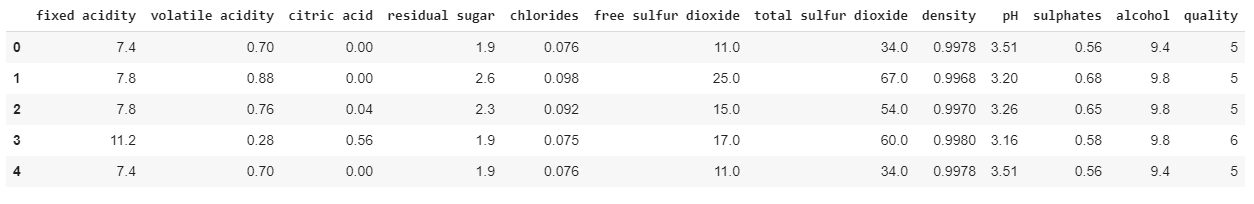
3. Split the train and test data in a ratio of 8:2

4. Use statsmodel.api.Poisson(y\_train,x\_train) to fit the data.

5. Test the performance of the model by running on test data.

6. Evaluate the goodness of fit measure using pseudo R -squared measures.

**Data set Used: (Attach Screen shot of the few rows)**

The dataset used was ‘winequality.csv’ which was obtained from kaggle. This dataset was chosen for the optimal amount of data on which regression could be performed(12 columns and 1600 rows). This dataset was also preprocessed, and hence there were no NaN values, or outliers.

The dependent variable is ‘quality’ and the rest of the columns are the independent variables.

This dataset was split into training and testing dataset in the 80:20 ratio

**Challenges faced during the implementation of the program:**

Poisson regression is an algorithm that is not present in sklearn, which is generally used to implement most machine learning algorithms. Thus, we had to look for other modules, and we found a library called ‘statmodels’ that could be used to perform poisson regression

Another challenge faced was the values of the Root Mean Squared Error(RMSE) during the first round of training the model(it was a bit high). This was overcome by normalizing the dataset, which resulted in much better values of RMSE.

**Code:**

df=pd.read\_csv('winequality.csv')

X = df[['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates','alcohol']].values

y = df['quality'].values

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

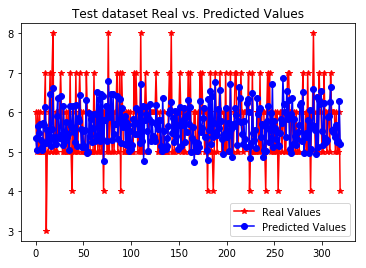
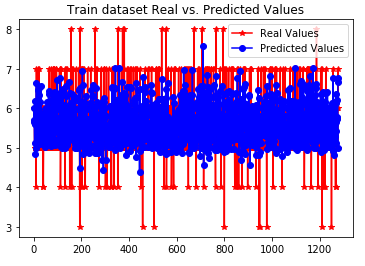
import statsmodels.api as sm

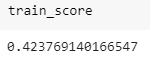
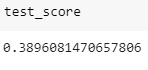
poisson\_mod = sm.Poisson(y\_train, X\_train)

poisson\_res = poisson\_mod.fit(method="newton")

print(poisson\_res.summary())

predVals1 = poisson\_res.predict(X\_test)

**Output: (Screen shots)**

**References:**

1. <https://www.statsmodels.org/stable/generated/statsmodels.discrete.discrete_model.Poisson.html>
2. <https://en.wikipedia.org/wiki/Poisson_regression>
3. <https://towardsdatascience.com/an-illustrated-guide-to-the-poisson-regression-model-50cccba15958>
4. NCSS Statistical Software : Chapter 325 Poisson Regression