Predicting the Mileage per Gallon of Cars

Bharat Mallala; Graduate student, School of Informatics and Computing;
Indiana University Bloomington, Indiana; bmallala@umail.iu.edu

ABSTRACT

This project aims to predict the mileage of cars based on various attributes of the car. The data set for this project is taken form the UCI machine leaning data set repository. The problem is a regression problem and will be converted to a classification problem with the response variable mpg having 5 classes. This project summarizes various models that can be fitted to the data such as Multi category logistic regression with ordinal response variable type, Baseline category model. Horsepower, year of manufacture, weight etc. are some of the explanatory variables that influence the mileage of a car. This project provides the results from the models while using different variables of interest from the data set and the interaction among these variables, along with the interpretation of the results with the goodness of fit statistics.

Index Terms— Classification, Logistic regression, Multi category, Poisson log linear, Explanatory, goodness of fit.

1. INTRODUCTION

Mileage of a car is one of the most important feature that most people look out for when they purchase a car. With increase is fuel costs all around the globe, a car which provides good mileage is good selling point for any manufacturer in the market. Numerous factors influence the mileage of a car in multiple ways. Some of them include weight, displacement, year of purchase etc.

The goal of this is paper is to fit models to the data such that best explains the data. The type of model to fit depends on the type of data and the attribute we are trying to predict. Since the data available to us has multiple attributes it is not feasible to use the generally used models for 2-way tables. For large data sets it is generally recommended to use Multicategory models. Generally, there are two types of multicategory logit models; one with nominal response variable and the other with ordinal response variable.

Since the response variable in this data set is ordinal with multiple categories it is ideal to use the Multi category logit model with the categorical response variable type. The data originally available in the repository has a nominal response variable ‘mpg’ making it a regression type of problem which is beyond the scope of this course. For the sake of simplicity and to fit the multinomial models in this project the goal is to convert the response variable into an ordinal response variable containing rating.

This project explains in detail how the importance of Multicategory models to deal with response variable having multiple categories and having an order. This project aims to fit multiple models on the multicategory model with and without interaction of the explanatory variables and to check for the goodness of fit of these models based on some metrics and to provide a model that best fits the data.

2. DATA DESCRIPTION

The data for this project is the ‘mileage’ data set taken from the UCI Machine learning repository. The data set has a total of 3 multivalued discrete and 6 continuous attributes. The attributes are mpg (the response variable), cylinders, displacement, horse power, weight, acceleration, year of purchase, origin and car name. The attribute car name has too many levels i.e. almost equal to the length of the data set. This attribute will not be feasible to use for predicting the ‘mpg’. Hence, we discard this attribute and use the remaining variables for the fitting the model. In this project, we will fit multiple models with and without interactions to come up with a model that best fits the data available

Most of these attributes are multivalued discrete or continuous attributes. The response variable ‘mpg’ is continuous variable, but for the commencement of this project this attribute is converted into an ordinal variable having rating from 1 to 5. For this process, the response variable is first bucketed in to multiple bins. These bins are then given rating based on intrusion. This process converts ‘mpg’ to a discrete ordinal response to which we can fit multicategory models.

3. PROPOSED MODEL

With data, here being a multicategory ordinal response variable data, the model that would best fit would be logit models for ordinal responses.

5.1. Problems with Baseline category models

The base line category model is generally used in the case where there are multiple categories in response variables and the response should be nominal. In this model, the goal is to choose a baseline categories from the various categories available we find the odds of every variable to fall in a category given it falls in the baseline category. This model works well in the case of nominal response variable. But since in this project we are dealing with ordinal response variable this project may not provide satisfactory results.

5.2 Cumulative logit model

When the response variables are ordered, the best model would be the one which takes into consideration this ordering into fitting the model. The Cumulative logit model of an attribute is the probability that attribute falls on or below a category of the response variable.

Considering Y as the response variable. The probability that Y falls on belong a category j is given by

$$P(Y \leq j) = \pi_1 + \cdots + \pi_j, \quad j = 1, \ldots, J$$

Now the all the probabilities of all the categories add to 1.

$$\text{logit}[P(Y \leq j)] = \log\left[\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right] = \log\left[\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J}\right], \quad j = 1, \ldots, J - 1$$

These probabilities are called cumulative logits.

6. EXPERIMENT AND RESULTS

Now that the model to be fitted has been decided and the final data set after all the data cleaning; the first step is to select the subset of the data set and reduce few explanatory variables. Using all the explanatory variables will not be feasible solution as it would increase the values of out residual deviance i.e. the G^2 value.

For feature section in this project we first fit the complete independence model using all the variables in the data set and then perform a stepwise backward elimination of the variables in the data set and check for the AIC (Akaike information criterion) value; which is the measure of the relative quality of statistical models for a given set of data. Using a R programming language to obtain these results we get the best independence model with a subset of the data set. Using this method, we have reduced the variables to 4 comprising of weight, horsepower, origin and year of manufacture which has an AIC value of 514.9662 and G^2 value of 498.9662.

For the independence model with the subset of variables we have,

```
Coefficients:
      Value Std. Error t value
h -0.019664  0.0085775  -2.292
y  0.478525  0.0149690  31.968
o  0.291455  0.1640935   1.776
w -0.004319  0.0003797 -11.374

Intercepts:
    Value     Std. Error t value
1|2   17.7871    0.0031  5677.6669
2|3   23.6481    0.4474    52.8573
3|4   27.3075    0.5540    49.2914
4|5   30.7276    0.6632    46.3344

Residual Deviance: 498.9662
AIC: 514.9662
```

The next step is to check if the interaction of these subset of variables will improve the accuracy of the model. Since in the 4 variables of interest in the subset, the aim is to fit all models from complete dependence to complete independence which includes, joint associations, conditional association and the homogenous association models. Since the complete dependence model will not be good fit to the data due to degree of freedom (df) becoming zero, we discard this model.

To fit these models, we will be using the 'polr' function in R to fit all the above stated models and then use step AIC function to get all the combinations of models that best suits the model. The model fits all the combination of variables until there is an increase in the AIC value for the model. Since we have 4 variables we will have 3 way and 2 way interactions. Some of these interactions may have a good impact on the logit model considerably decreasing the G^2 value.

Fitting the cumulative logit model with inclusion of the interaction terms. We get the model (yh , oy, yw, oh) as the best model that fits the data.

Running the model we get,

```
Coefficients:
       Value Std. Error    t value
y    0.499431   0.018480  2.702e+01
h    0.294539   0.010594  2.780e+01
```

```
w   -81.812816   0.000295 -2.773e+05
o    -5.386045   0.002815 -1.913e+03
y:o   0.075547   0.008507  8.880e+00
y:w   0.868873   0.016577  5.241e+01
h:o  -0.002007   0.007616 -2.635e-01
y:h  -0.004129        NaN        NaN

Intercepts:
    Value       Std. Error   t value
1|2     25.9084     0.0015   16828.0500
2|3     31.9942     0.4967      64.4109
3|4     35.6828     0.6014      59.3375
4|5     39.2246     0.7200      54.4774

Residual Deviance: 489.8577
AIC: 513.8577
```

From the results of the model we get the intercept value for every variable and the interaction terms along with the standard error and the t-value.

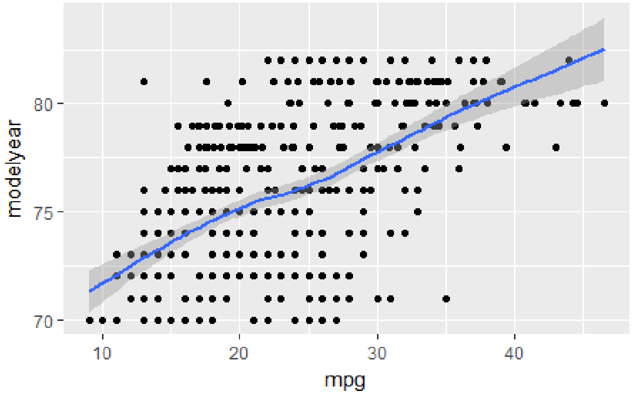
For every logit only the intercept value changes and but the coefficients remain the same. The summary of the model gives us the intercept value for every combination of logits. For example, from the results 25.9084 is the intercept of logit of probability of getting a rating of 1 to the probability of getting a rating of 2.

The model has got a G^2 value of 489.8577 which is much lower compared to the model with completed independence and the P-value is large enough. Hence, we can say that this model better fits the data.

Model	G^2	AIC
y+h+o+w	498.9662	514.9662
Complete independence with all variables	495.821	517.821
(y,h,,w) , (y,o,w),(h,o,w)	502.4917	528.4917
(yh , oy, yw, oh)	489.8577	513.8577

From the results in the above table the model (yh , oy, yw, oh) gives the best results with low AIC and also G^2 value and also having a high P-value. Hence this model best suits the data.

The goodness of fit statistics has been performed on this model and the results look promising giving a P-value of 0.9 for the chi squared test.



The graph above show how the mpg is distributed according to the year of purchase. We can see that there is a positive linear relationship. The model also proves this as the we have a positive coefficient value for the y variable.

7. CONCLUSION

In this project, we have analyzed and fitted various models to the selected data set and came up with a model that could best explain rating given to cars for the 'mpg' they offer based on the weight, horsepower, origin and year of purchase. In this project we have explained the importance of Cumulative logit model and how it can be applied to data with Multi category ordinal response variable data. Different models from completed independence to complete dependence have been performed on the "mileage" data and the best model has been selected. The ideal model explains most of the variance in the data. Goodness if fit along with the model statistics have outputted satisfactory results for the selected model.

12. REFERENCES

[1] "An Introduction to Categorical Data Analysis" by Alan Agresti, Second edition, 2007.

[2] http://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression.