Bonus Assignment #1: Cluster-Wise Regression

PREDICT 422 Section 58

Winter 2018

Scott M. Morgan

INTRODUCTION

The purpose of this report is to perform cluster-wise (or latent class) regression to predict class membership using the car data set in RStudio. Cluster-wise regression is a model-based clustering technique which allows for model comparison using various criteria (AIC, BIC, etc.) (Grun & Leisch, 2008).

RESULTS

In the subsequent sections, a series of cluster-wise regression models are generated and evaluated. I first use the functionality within RStudio to perform a brief exploratory data analysis (EDA). Following this, the data set is cleaned to generate 5 cluster-wise models. Finally, the models are evaluated and the best is chosen.

Exploratory Data Analysis (EDA). The 'car.test.frame' data set is first loaded from the *rpart* package. The data set contains 60 observations and 8 variables. Each represents information on the makes of cars taken from the April, 1990 issue of *Consumer Reports*. All variables in the data set are non-negative with 6 of the variables appearing continuous (including the response). The remaining 2 are categorical.

While the naming convention of the variables is relatively effective, the data dictionary is a helpful resource. Table 1 below provides an alphabetic list of the variables with their respective data types and descriptions, which are also available in RStudio using the *?car.test.frame* function.

Table 1: Alphabetic List of Variables and Data Dictionary

Variable	Type	Label
Price	Int	a numeric vector giving the list price in US dollars of a standard model
Country	Factor	Country of origin, a factor with levels France, Germany, Japan , Japan/USA, Korea, Mexico, Sweden and USA
Reliability	Int	A numeric vector coded 1 to 5.
Mileage	Int	Fuel consumption miles per US gallon, as tested.
Type	Factor	A factor with levels (Compact, Large, Medium, Small, Sporty, Van)
Weight	Int	Weight in pounds.
Disp.	Int	The engine capacity (displacement) in liters.
HP	Int	The net horsepower of the vehicle.

Using the *corrplot* package in RStudio, Exhibit 1 below which presents the correlations of the continuous variables. Weight, Disp., and HP appear relatively strongly positively correlated with Price while Mileage is negatively correlated with the response. Reliability has no correlation with Price.

Exhibit 1: Correlations of Continuous Variables

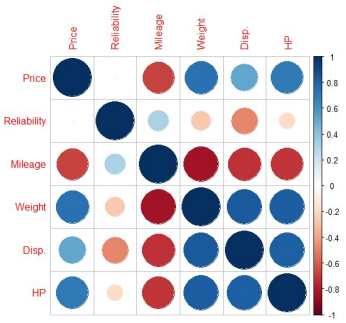


Exhibit 2 shows additional exploratory analysis of the continuous variables being considered relative to Price. Using *ggplot*, a LOESS curve (i.e. local regression) is overlaid with upper and lower confidence intervals around the mean. Each point is colored by Country and the shape is representative of Type. All observed relationships are generally linear in nature. Mileage appears to have a negative relationship with Price which makes sense intuitively as someone would generally not pay more money for a car that has higher mileage. The remaining variables have positive relationships with the response. The inclusion of the categorical variables as coloring and shape is interesting but noisy. Overall it is difficult to ascertain much insight from them. The majority of vehicles appear to be either from the USA or Japan, and small cars get much more inexpensive than other Types at higher mileage values and lower values of Weight, Disp. And HP. HP seems to be the only continuous variable that could benefit from transformation.

Exhibit 2: Line Plot with LOESS of Continuous Variables with Response

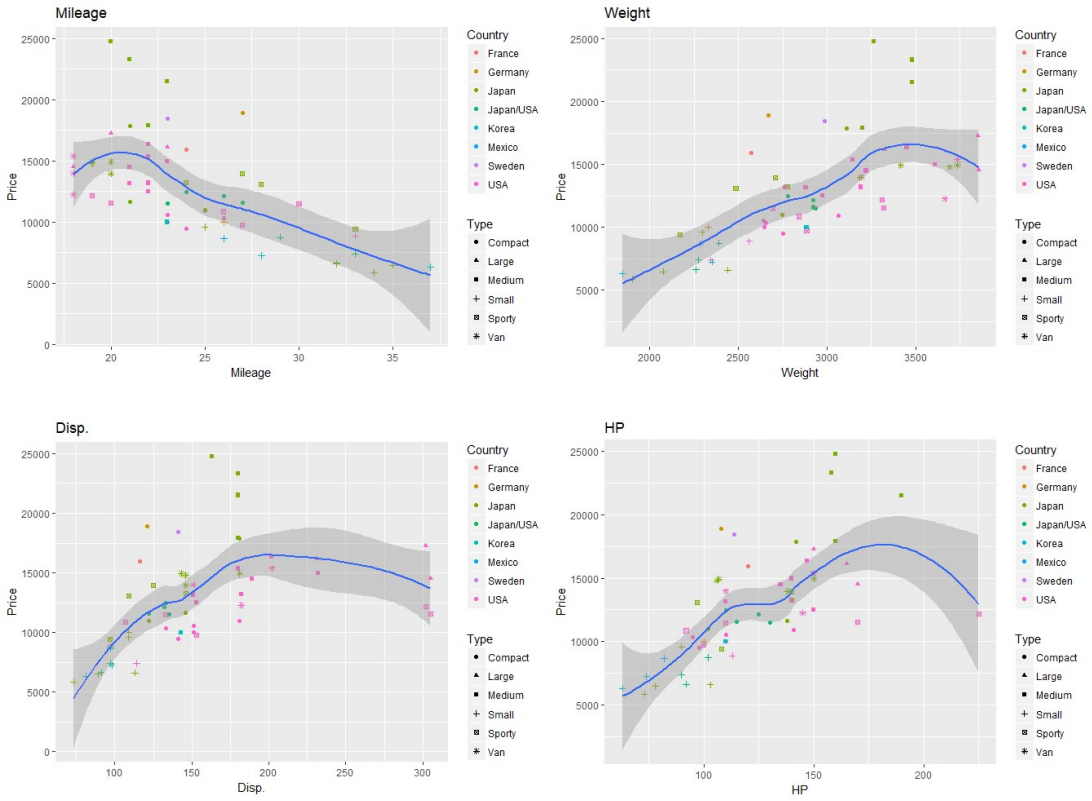
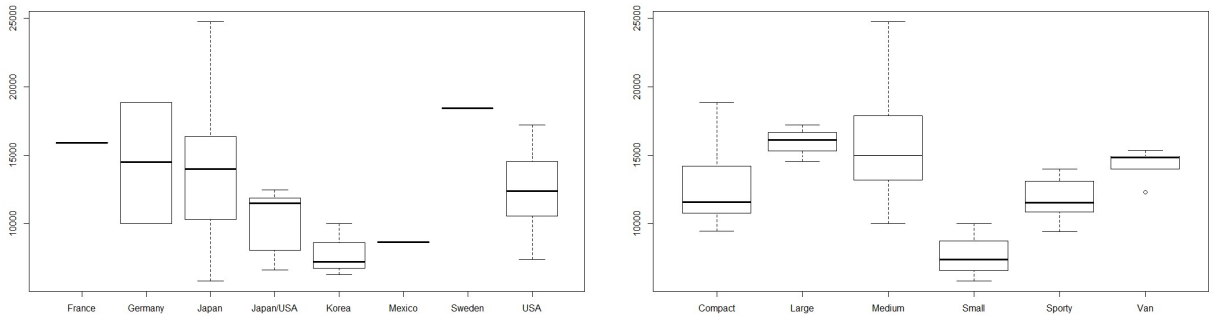


Exhibit 3 displays boxplots of the categorical variables with Price. The Price range is particularly wide with vehicles manufactured in Japan and USA. There also appears to be a relatively wide price range of Medium Type cars as well as a possible outlier in the Van category.

Exhibit 3 Categorical Variables versus Price



DATA PREPARATION. In this section, the data is prepared before moving on to the model construction portion of the project.

Missing Variables. Using the *na.omit*() function, any rows with *NA* in the original set are removed. The resultant data frame has 49 entries.

Variable Transformations. Ultimately, I decided not to transform any variables as all the predictor relationships appear generally linear in nature with the response.

Categorical Variables. While interesting, *flexmix* does not support categorical variables so these columns are removed from the data set. The final training data set has 6 variables including the response.

MODEL PRODUCTION. In this section, 5 models are generated. The first 3 models use the *flexmix* package using 1, 2 and 5 clusters, respectively. Model 4 uses the *flexmix* package with a Zero Inflation Poisson distribution specified with 10 clusters via the *FLXMRziglm()* component-specific driver. While the data set might violate several assumptions of the Zero Inflation Poisson distribution, I include the model purely for experimentation purposes. The final model uses the component-specific model driver *FLXMRglm()* which allows fitting of finite mixtures of GLMs with 100 clusters. Weight is used as the concomitant variable to hopefully improve precision of the estimate via the model driver *FLXPmultinom().* Model 5 likely violates several statistical assumptions but, similar to Model 4, is included for experimentation purposes. Rootograms of the posterior probabilities and the estimated coefficients for each model are presented in the following sections.

Variable Selection. Before presenting the models, Table 2 below shows the variables selected by the *randomForest* package. Variables with a large mean decrease in accuracy are more important for classification of the data. According to this metric, Weight and HP are the two most important variables from the cleaned data set. However, as all of the variables seem relatively correlated with the response and the relationships generally make sense intuitively, all 5 variables are included in the final models as predictors.

Table 2: Variable Importance

Variable	Mean Decrease Gini
Reliability	5.20
Mileage	9.49
Weight	11.59
Disp.	9.60
HP	10.38

Model Results. Using the *flexmix* package, 5 models are generated. Exhibit 4 provides the rootograms of the posterior probabilities to illustrate how well observations are clustered by mixture. The posteriors of the different components are considered well-separated if the modes are at 0 and 1. This is the case with Models 2 and 5.

Exhibit 4: Rootogram of the posterior probabilities

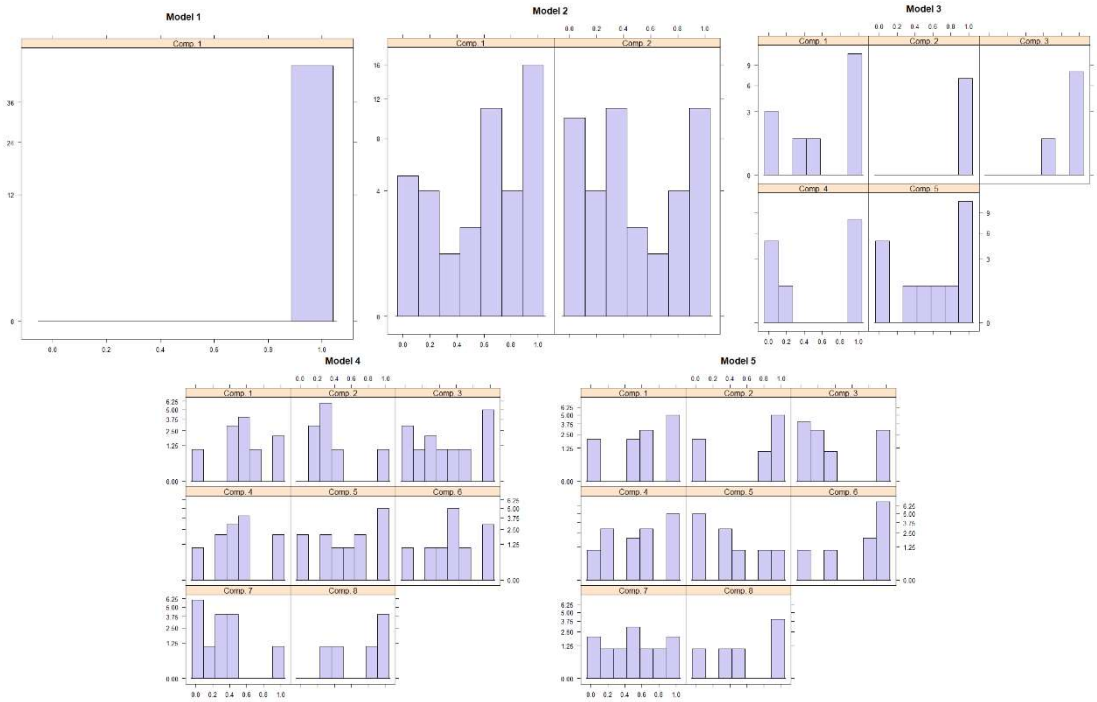
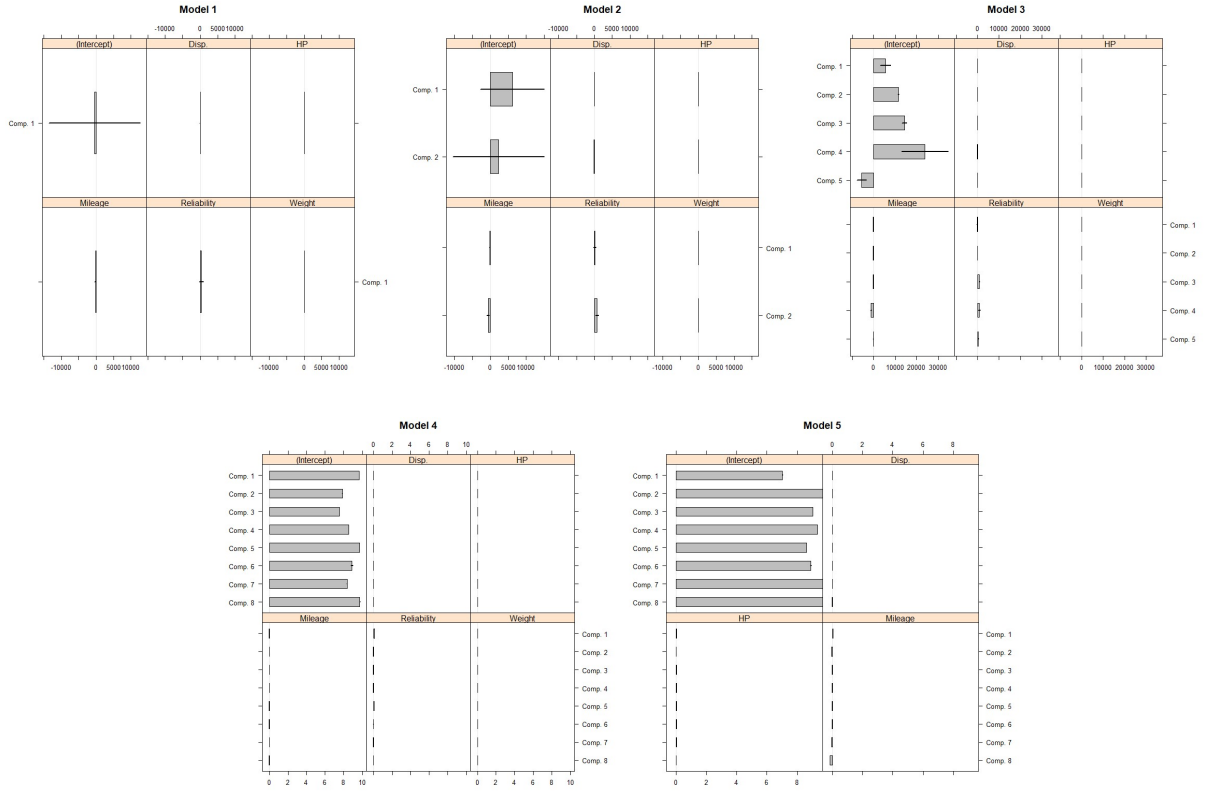


Exhibit 5 provides the estimated coefficients. The dash black lines indicate the 95% confidence intervals. These figures were generated using the *plot()* and *refit()* functions. The plots suggest that the estimated coefficients do not vary much across the different models and components.

Exhibit 5: Estimated coefficients of the component specific models with corresponding 95% confidence intervals



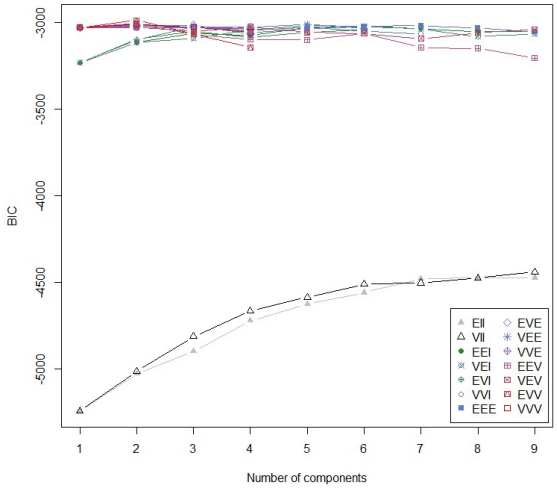
MODEL SELECTION. For purposes of this exercise I use primarily quantitative measures to assess the best model. Table 3 below provides metrics for each of the 5 models. The main quantitative criteria I evaluate on are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Our guiding rule is to minimize both the AIC and BIC.

Table 3: Model Comparison Statistics

Criteria	Model 1	Model 2	Model 3	Model 4	Model 5
Method	*K = 1*	*K=2*	*K=5*	*K=10 /ZIP*	*K=10 / Poisson / Concomitant*
AIC	916.476	890.8245	815.3693	832.1703	829.0061
BIC	929.7187	919.2018	889.1502	936.2204	916.0298

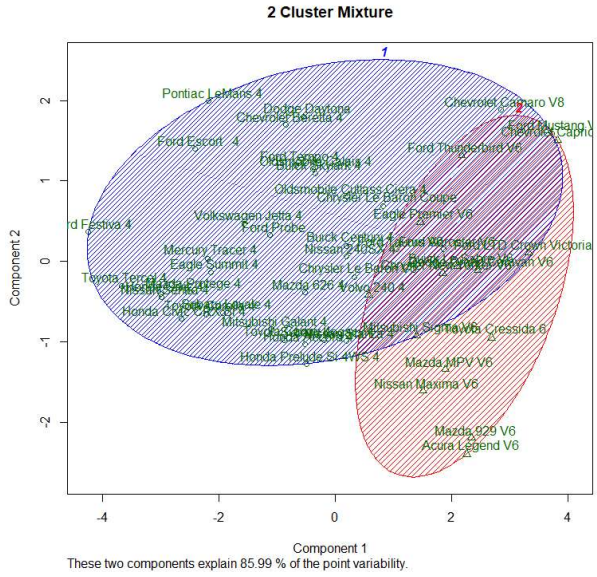
The different techniques produce relatively similar results across the models. In terms of the Poisson models (4 and 5), these outperformed the majority of alternatives by AIC except for Model 3. Using the *mclust()* package, I also check which mixture is considered the optimal model according to BIC for parameterized Gaussian mixture models. Exhibit 6 provides a graph of all the proposed model options given various multivariate mixtures by BIC. According to this algorithm, the BIC suggests the model with 2 groups, or clusters, is the best choice.

Exhibit 6: Model Selection by BIC using *mclust()* package



Given that Models 4 and 5 possibly violate several assumptions given the structure of this data set and Models 1 and 3 are mediocre from a metrics standpoint, the recommendation is to use Model 2. This model shows well separated components (Exhibit 4), has relatively strong performance metrics (Table 3) and is the optimal mixture given the *mclus)* package (Exhibit 6). Using the *fpc* and *cluster* packages, we see that the two components explain 85.99% of the point variability (Exhibit 7).

Exhibit 7: 2 Component Mixture (Model 2)



CONCLUSION

Overall this was an interesting exercise to illustrate that increasing the number of clusters can improve the cluster-wise regression model performance but at a diminishing rate. It also was an excellent introduction into the functionality RStudio has for latent-class regression and clustering visualizations.

REFERENCE(S)

1. Consumer Reports, April, 1990, pp. 235–288 quoted in John M. Chambers and Trevor J. Hastie eds. (1992) *Statistical Models in S*, Wadsworth and Brooks/Cole, Pacific Grove, CA, pp. 46–47.

2.Grün, B., & Leisch, F. (2008). FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software,* 28(4). doi:10.18637/jss.v028.i04