STA 138 Project I

Xi Cheng, Yi Zhang,

Haozhe Gu, Jiansen Jiang

# Introduction

Car evaluation is a key concept for auto business industry. Car companies need to study and understand what drives the consumers the most to accept or decline a car. However, the level of acceptance of a car is not simply based on the price that is no-doubt an important factor. Our group decided to study the most concerned factors of a car evaluation. The problem that we are interested in is the relationship between buying price, safety levels and the final decision to acceptance of cars.  We are interested in questions like whether cars with lower buying price or higher safety level are more likely to cause an acceptance in final decision than other cars.

# Materials and Methods

Car Evaluation Dataset is obtained from the UCI Machine Learning Data repository. It was derived from a simple hierarchical decision model originally for DEX. It has 6 attributes and 1 response variable (Final Decision); and it has 1728 instances without any missing values. All of the 7 variables are categorical variable with different levels; actually, all of them are ordinal variables. The response variable or final decision’s level varies from “unacc” (Unacceptable), “acc” (Acceptable), “good” (Good) and “v-good” (Very Good).  The 7 variables are buying price, price of maintenance, comfort, number of doors, capacity in terms of person to carry, size of luggage boot and the estimated safety of the car.

Because we emphasis on certain variables, we trimmed the data and did some concatenation.  We combined “acc”, “good” and “v-good” to form a new level “acc” for the response variable so that the levels of response variable are “unacc” and “acc” (with approximate 70% and 30% frequencies) We also merged the levels of buying price from (low, medium, very and very-high) into to levels (low, high).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Safety | | | | | |
|  | Low | | Medium | | High | |
| Acceptance | Price | | Price | | Price | |
|  | Low | High | Low | High | Low | High |
| Acc | 1 | 1 | 154 | 64 | 184 | 115 |
| UnAcc | 288 | 288 | 134 | 223 | 104 | 173 |
|  |  |  |  |  |  |  |

**Table 1: Contingency Table For Dataset Car Evaluation**

Clearly, the odds ratio between the level of acceptance and level of price is 1 given the safety is low. This means, given low safety level, the factor, price, does not affect the level of acceptance.

Based on the contingency table above, we assumed three models to fit:

1. Conditional independent model: *(XY, YZ)* for two factors that are independent given the other factor (we guess the price and level of safety are independent given level of acceptance);
2. Homogeneous association: *(XY, YZ, XZ)* for all three pairwise relations exist;
3. Three-way saturated model: *(XYZ)* for a general log-linear model for three-way table.

Then, by fitting to these models and testing the fitness, we chose the “best” model and got rid of the non-associated terms. After that, by calculating the odds ratios, the final results about the original question can be answered. The analysis may use notations {PR, SA, AC} for {Buying Price, Safety Level, Acceptance} for simplicity.

# Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model |  | **DF** | **P-Value** | **AIC** |
| Conditional Independence Model | 23.632 | 4 | 9.464e-05 | 112.62 |
| Homogeneous Association Model | 3.0837 | 2 | 0.21340 | 96.074 |
| Saturated / Three-way Model | 5.5e-15 | 0 | -------- | 96.99 |

**Figure1:** Model Selection Statistics

From the Statistics of Goodness of fit table above, we choose the Homogeneous Association Model as our primary model to study because of small , large p-value and smallest AIC score. Conditional Independence model fail to pass the goodness of fit test; the saturated/ three-way model has AIC score higher than homogeneous association model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Coefficients |  | **Std.Error** | **Z-Val** | **P-value** | **LowerLimit** | **UpperLimit** |
| PRlow:SAlow | 0.598 | 0.136 | 4.378 | 1.2e-05 | 1.3915 | 2.3776 |
| PRlow:SAmedium | 0.169 | 0.124 | 1.362 | 0.1731 | 0.9284 | 1.511 |
| PRlow:ACUnacc | -1.157 | 0.124 | -9.272 | < 2e-16 | 0.2462 | 0.4015 |
| SAlow:ACUnacc | 5.903 | 0.714 | 8.260 | < 2e-16 | 90.27 | 1487.2 |
| SAmedium:ACUnacc | 0.612 | 0.124 | 4.903 | 9.46e-07 | 1.444 | 2.357 |

**Figure2:** Summary Statistics of Homogeneous Model Coefficients

From the summary of Homogeneous model, there is strong evidence of conditional association in four out of five terms. The only term PRlow: SAmedium lack the evidence of strong association because of it p-value equal to 0.1731.

We construct 95% confidence interval for the odd ratios of these conditional associations. Out of the five terms, four of them confidence interval greater than 1 or less than 1 imply the following infomation.

1. PRlow : Salow -> (1.3915 , 2.3776)

There is positive conditional association between PR and SA scales. To be precise, for any given AC, the cars of “Low” in PR with “Low” in SA is at least 1.4 time more odds than that for cars of “Low” in PR with “High” in SA.

1. PRlow : ACUnacc -> (0.2462,0.4015)

There is negative conditional association between PR and AC scales. To be precise, for any given SA, the cars of “Low” in PR with “Unacc” in AC is at least less than 0.4015 the odds for cars of “Low” in PR with “Acc” in AC.

1. SAlow:ACUnacc -> ( 90.27 , 1487.2)

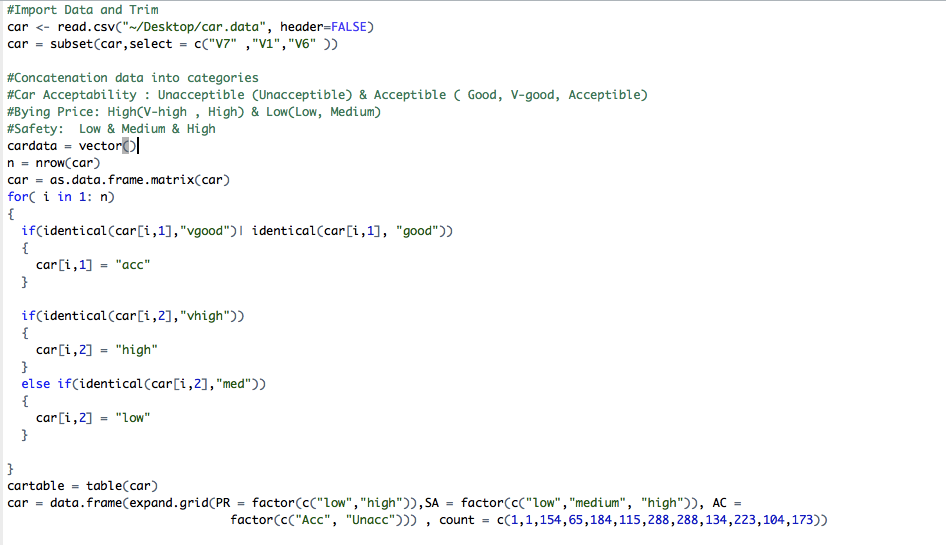
There is positive conditional association between SA and AC scales. To be precise, for any PR, the cars of “low” in SA with “Unacc” in AC is at lest 90.27 times more odds than that for cars of “low” in SA with “Acc” in AC.

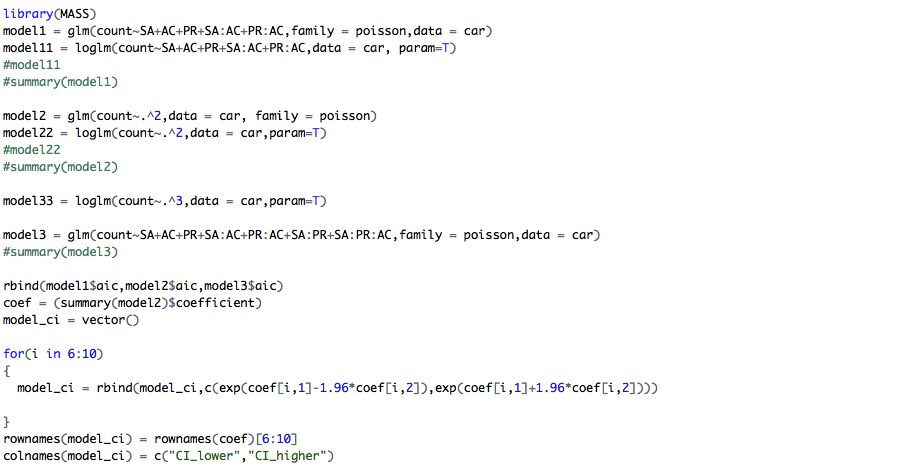
1. **Conclusion and Discussion**

From Figure1 statistics, we found that the homogeneous association model (pair-wise association) fit the dataset best, that is, all three variables are conditionally dependent. Based on Figure2 p-values, we conclude that buying price at level low and safety level at medium lack evidence to prove strong conditional association. The three odd ratio confidence interval from Figure2 shows that I) No matter the decision is acceptance or not, when a car has low buying price; this car is more likely to be at low safety level than other levels. II) At any given safety level, cars with low buying price are less likely to lead to a final decision of un-acceptance than acceptance. In other words, lower buying price cars are more likely to be accept than higher price cars with any given safety level. III) Given any buying price level, low safety level are much more likely to be unaccepted than to be accepted. Additionally, when safety level is low, the partial 2 by 2 table of buying price and final decision has an odd ratio of one, implies that they are independent.

However, there are still some statistics that cannot be explained reasonably by these three factors. This might because we arbitrarily trimmed many features of the dataset. Some of the other variables like capacity in terms of person and luggage boot size might also be an important factors contributing to the final decision.

1. **Appendix**

****

****