MA678 Midterm Project

Yuli Jin

2021/11/11

Abstract

I conduct a multilevel linear regression model to find the relationship between the used car price and used car specification. In the process of modelling, I check the marginal Model Plots of predictor and add polynomial form on one predictor. After that, I use manufacture brand and body type as group-level predictors to construct varying-intercept model. The residuals of the fitted model don't have obvious correlation with variables but have heteroskedasticity and don't follow normal distribution. The coefficients of the fitted model shows valid interpretation.

Introduction

Cars are convenient means of transportation. However, selling and buying a used car isn't necessarily convenient. If people are planning to purchase or sell a used car, it always takes them a long time to search similar cars to compare and try to estimate a price for their own ones. Since manual comparison cannot reflect the precise relationship between price and car feature, the seemingly fair price isn't usually accurate. Moreover, when it comes to another car, they have to repeat such time consuming but rough estimation again. It is true that some websites have built the model to give a price for reference as long as customers type in their car conditions. Nonetheless, people still can't preciously figure out what factors on earth decide the value of used car.

In this report, I conduct a multilevel linear regression model to analyse how these factors influence the used car price. The analysis consists of three sections. In Method section, I first introduce the data source used in this report and conduct some data processing. Then I conduct exploratory data analysis and fit the model. After that, I check the residuals of the fitted model. In Result section, I interpret the coefficient of fitted model and display the random effect of multilevel model. In Discussion section, I summarize what I have done and found in this report. Also, I include limitation and next step work at last.

Method

Data Processing

The data is from Kaggle-usedcarscatalog. The following table is the explanation of some columns.

Column	Explanation
manufacturer name	The name of car manufacturer
odometer value	Odometer state in 100000 kilometers
year produced	The year the car has been produced
engine capacity	The capacity of the engine in liters, numerical column
body type	Type of the body (hatchback, sedan, etc)
has_warranty	Does the car have warranty?
feature 0-9	Is the option like alloy wheels, conditioner, etc. is present in the car
price usd	The price of a car as listed in the catalog in USD

The data has 38531 observations and 30 columns. In the following analysis, I choose 10 manufacture brands from Japan and Korea and several variables listed above to conduct the analysis. I also conduct some transformation on some columns. For `year produced`, I use $year\_duration = log(2020 - year\_produced)$ to get the age of cars being used. For `feature 0-9`, they are 9 bool type variables. I add them together as $feature = \sum_{i=0}^{9} feature_i$ to get the sum of feature.

Exploratory Data Analysis

For continuous variables, I use quantile to separate continuous variable into 10&5 groups and calculate the mean of log(price). For category variables, I use box plot and density plot to check the distribution of log(price).

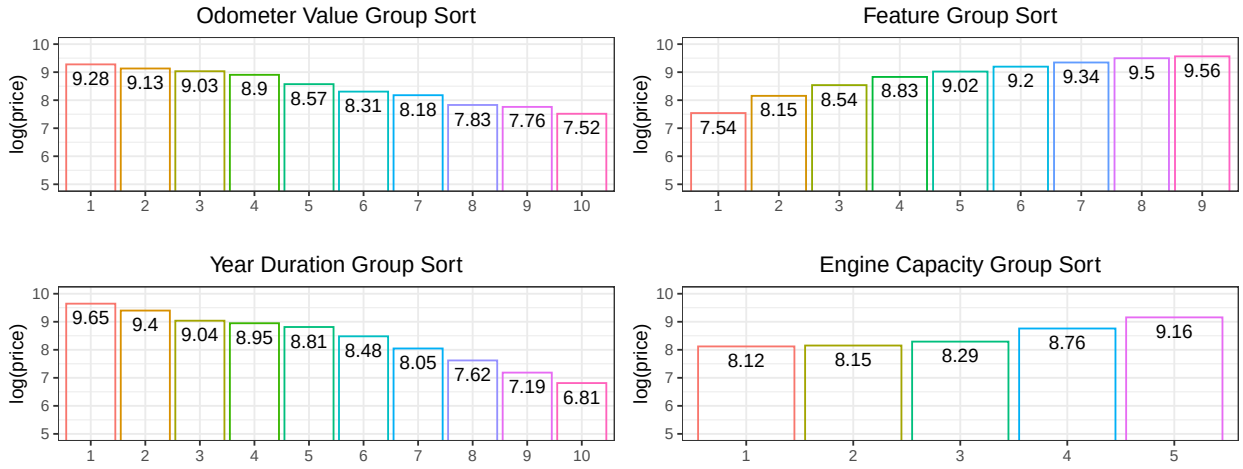

Figure 1: Group Mean of Variables

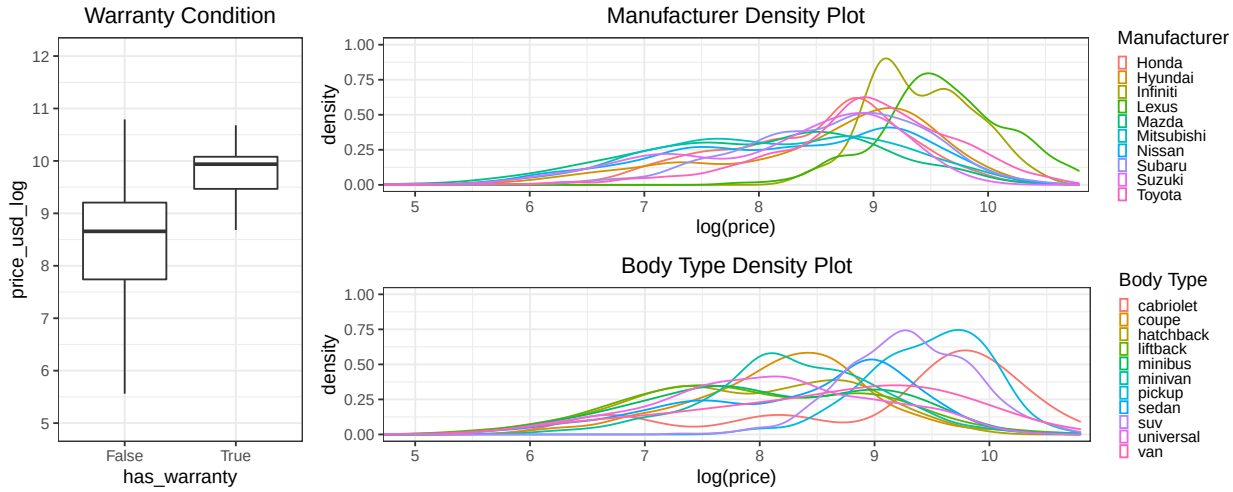

Figure 2: Gategory Variables

From Figure 1, we can observe that each variable shows clear trend with mean of log(price) when divided through quantile. For odometer value, the larger the odometer value is, the lower the car price is. For features, the more features included, the higher the car price is. For year duration, the longer the car is, the lower the price is. For engine capacity, the larger the engine capacity is, the higher the car price is.
Figure 2 consists of three plots, the first one is the log(price) boxplot of warranty condition, we can see that

cars with warranty are more likely to have higher price. The rest two plots are the price density plot of manufacturer name and body type respectively. It is rational that different brands and body types have different density distribution.

Model Fitting

Previous exploratory data analysis section shows vivid trends and difference of car price. More importantly, these characteristics corresponds to most people's common sense. In this section, I elaborate the model fitting process. First, it is important to check if potential variables have polynomial effect on log(price). Therefore, I apply marginal model plots without polynomial effect on odometer value, car feature, year duration and engine capacity. The plots are shown below:

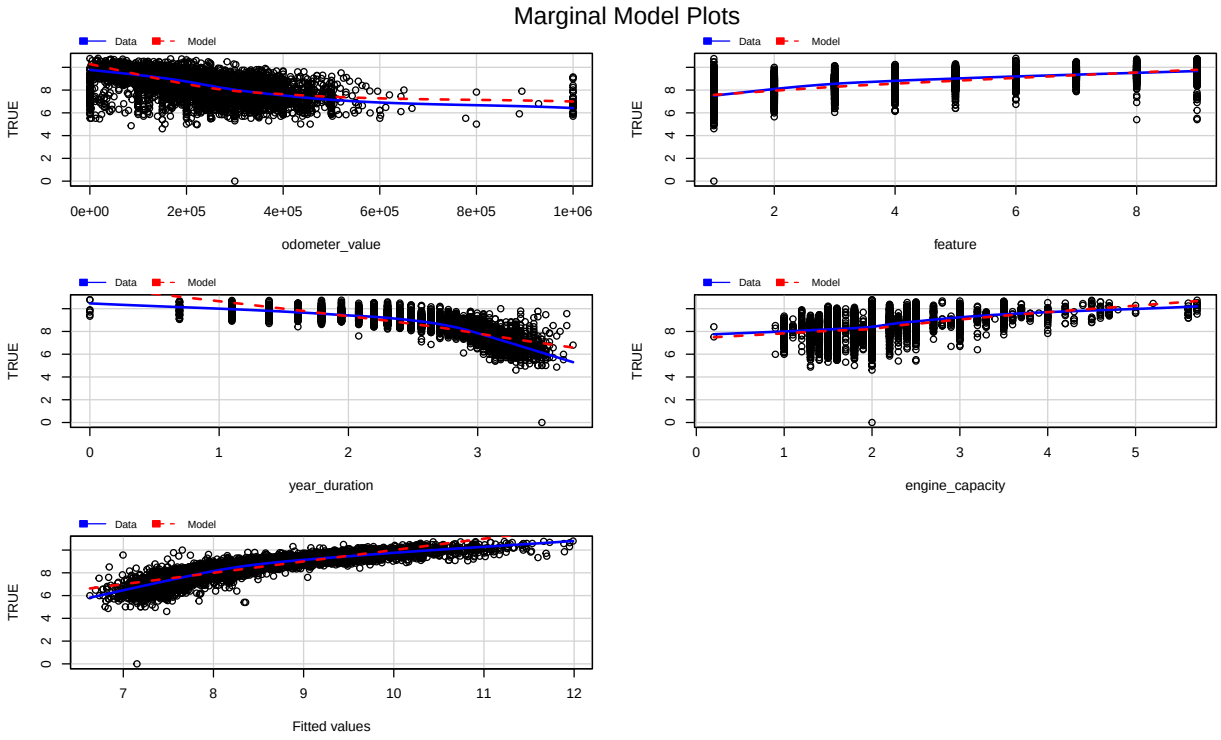

Figure 3: Marginal Model Plots

Figure 3 shows that the marginal model line of year duration deviates from the data line at both tails and it may be the main reason results in the deviance of fitted values and true value. Other marginal lines generally fit well. Note that even though the car feature is integral but not continuous, I regard it as continuous value instead for simplicity.

Then I fit the multilevel model. For year duration, I use polynomial with degree of 3. Given that manufacturer and body type have influence on car price, I use the interaction of these two variables to construct the multilevel term. Here is the model fitting result:

$$\begin{aligned} log(price) =& 7.981 + 0.027 feature - 60.510 year\_duration - 24.011 year\_duration^2 - 7.406 year\_duration^3 \\ & - 0.034 odometer\_value + 0.254 engine\_capacity + 0.085 warranty + n_j + \epsilon \end{aligned}$$

$$n_j \sim N(0, \sigma_a^2)$$

where n_j is the random effect of manufacturer name:body type

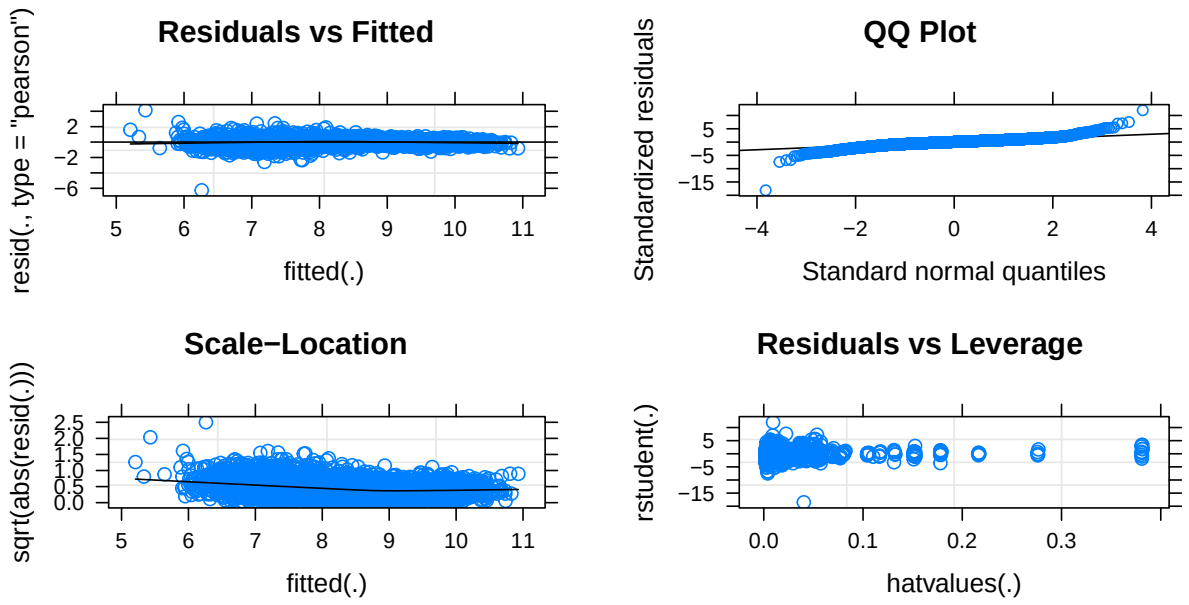

Figure 4: Residual Plot

Finally, I conduct residual analysis. The Residuals vs Fitted plot shows that the mean of residuals lies near zero, indicating that the residuals don't have obvious correlation with predictors. In QQ plot, many residual points aren't on the line, which shows that the residuals don't follow normal distribution. In Scale Location plot, the standard error of residuals are rather high when the fitted value is small, but the standard error gradually reduce when the fitted value increase. Such condition indicates potential heteroskedasticity in residuals. In Residuals vs Leverage plot, it is hard to identify the outlier. I also use `cook.distance` to check if the distance is larger than 0.5. The result shows that there are 4 influential observations.

Result

Under the fitted multilevel model, for every one unit increase in feature, the average car price is expected to increase by 2.7% when other variables remain unchanged. For every 100000 kilometers increase in odometer value, the average car price is expected to reduce by 3.4% when other variables remain unchanged. For every one unit increase in engine capacity, the average car price is expected to increase by 28.9% when other variables remain unchanged. For cars with warranty, the average car price is expected to increase by 8.5% compared with those without warranty when other variables remain unchanged. For year duration, it is hard to directly interpret in that this term include polynomial degree. Here I use derivative instead to illustrate it. When we control other variables, We have:

$$
\begin{aligned}
price\_increment =& \frac{\mathrm{d}exp(-60.510log(y) - 24.011log^2(y) - 7.406log^3(y))}{\mathrm{d}y} \\
=& exp(-60.510log(y) - 24.011log^2(y) - 7.406long^3(y)) * (-60.510\frac{1}{y} - 24.011\frac{2log(y)}{y} - 7.406\frac{3log^2(y)}{y}) \\
=& exp(-60.510log(y) - 24.011log^2(y) - 7.406long^3(y)) * (-60.510\frac{1}{y} - 48.022\frac{log(y)}{y} - 22.218\frac{log^2(y)}{y})
\end{aligned}
$$

Therefore, for every one unit year increase, the average price is expected to change by $exp(-60.510log(y) - 24.011log^2(y) - 7.406long^3(y)) * (-60.510\frac{1}{y} - 48.022\frac{log(y)}{y} - 22.218\frac{log^2y}{y})$.

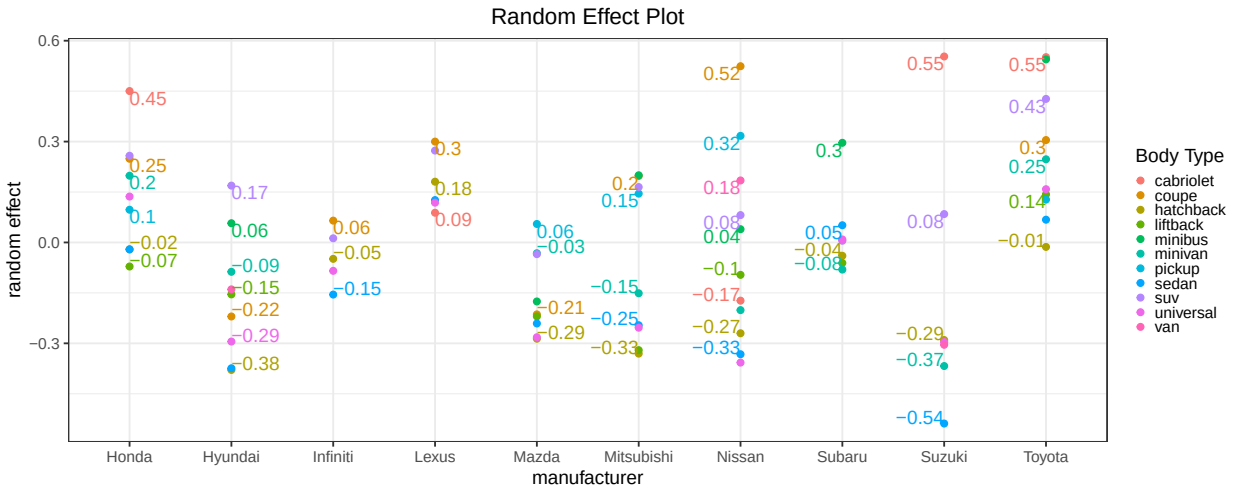

Figure 5: Random Effect Plot

The random effect plot shows that the car price varies from manufacturer and body type. Take Toyota as an example, most of their cars have positive random effect, but body types like cabriolet, suv and minibus enjoy a rather high random effect compared with other types. We can also compare the same body type with different manufacturers. For example, Toyota's suv has the highest random effect among all manufacturers' suv.

Discussion

In this report, I conduct a multilevel linear regression based on used on car dataset downloaded in Kaggle. I select feature, year duration, odometer value, engine capacity and warranty as predictor and apply manufacture name and body type as random effect. The interpretation of coefficients mostly corresponds to people's common sense. This report gives more precise relationship between selected variables and car price.

However, there are still some limitation in this report. First, this report only focuses on 10 manufacturer brands. Chances are good that brands excluded from this report may have different explanation on predictors and random effect. Second, the predictors selected in this report may be more or less correlated. This report don't further analyze the correlation and deal with them. Third, the residuals in the fitted model don't follow normal distribution and have heteroskedasticity. Therefore, the residuals don't completely confirm to the assumptions based on linear regression. Fourth, the overall used car market price may change over time and the market price may be affected by macro index as well. This report don't take these factors into consideration. Finally, linear regression is usually suitable for interpretation but not accurate in prediction. Many other methods like machine learning and deep learning have better predictive power.

For the next step, I plan to include more brands into account and conduct more feature engineering in the basis of current exploratory data analysis, then I plan to apply machine learning models like Random Forest, XGboost and LightGBM to build the model for more accurate prediction. Also, I consider combining models through stacking to further improve the predictive power.

Appendix

Random Effect Table:

manufacturer	cabriolet	coupe	hatchback	liftback	minibus
Honda	0.45	0.25	-0.02	-0.07	NA
Lexus	0.09	0.30	0.18	NA	NA
Nissan	-0.17	0.52	-0.27	-0.10	0.04
Suzuki	0.55	NA	-0.29	NA	NA
Toyota	0.55	0.30	-0.01	0.14	0.54
Hyundai	NA	-0.22	-0.38	-0.15	0.06
Infiniti	NA	0.06	-0.05	NA	NA
Mazda	NA	-0.21	-0.29	-0.22	-0.18
Mitsubishi	NA	0.20	-0.33	-0.32	0.20
Subaru	NA	NA	-0.04	-0.06	0.30

manufacturer	minivan	pickup	sedan	suv	universal
Honda	0.20	0.10	-0.02	0.26	0.14
Lexus	NA	NA	0.13	0.27	0.12
Nissan	-0.20	0.32	-0.33	0.08	-0.36
Suzuki	-0.37	NA	-0.54	0.08	-0.30
Toyota	0.25	0.13	0.07	0.43	0.16
Hyundai	-0.09	NA	-0.37	0.17	-0.29
Infiniti	NA	NA	-0.15	0.01	-0.08
Mazda	-0.03	0.06	-0.24	-0.03	-0.28
Mitsubishi	-0.15	0.15	-0.25	0.16	-0.25
Subaru	-0.08	NA	0.05	0.01	0.01

Model Checking Details:

	Estimate	Std. Error	t value
(Intercept)	7.981	0.038	208.777
feature	0.027	0.002	11.498
poly(year_duration, 3)1	-60.510	0.554	-109.209
poly(year_duration, 3)2	-24.011	0.377	-63.642
poly(year_duration, 3)3	-7.406	0.355	-20.835
odometer_value_th	-0.034	0.004	-7.662
engine_capacity	0.254	0.009	28.928
has_warrantyTrue	0.085	0.081	1.039