MAT374: Analysis of car price and interior with R

Lixin Li, Ziyan Zhu, Xinyi Huang

May 4, 2023

Contents

1 Introduction **3**

1.1 Description of the Variables 3

1.2 Research questions 4

2 Regression **5**

2.1 Collinearity 5

2.2 Linear Model 5

2.3 Ridge Regression 7

3 Classification **8**

3.1 Logistic regression 8

3.2 LDA 9

3.3 QDA 9

4 Results **10**

4.1 Regression 10

4.1.1 Collinearity 10

4.1.2 Linear model 10

4.1.3 Ridge model 10

4.2 Classification 10

4.2.1 Logistic regression 11

4.2.2 LDA 11

4.2.3 QDA 11

4.2.4 Compare error rate of the models 11

5 Conclusions **12**

Contents

List of Tables

1 Collinearity of variables table 5

2 VIF of linear model predictors table 5

3 Coefficient of linear model table 6

4 Summary of the ridge regression model 7

5 Coefficient of logistic regression model table 8

6 The confusion matrix of LDA 9

7 The confusion matrix of QDA 9

List of Figures

1 Ridge regression's log lambda vs. MSE 7

1 Introduction

We found a data set in Kaggle talking about car attributes and we will use the data of variables to do some prediction. The total amount of variables we have are 16 and the amount of variables we used are 8. Here is the website where we found the data set https://www.kaggle.com/datasets/deepcontractor/car-price-prediction-challenge.

1.1 Description of the Variables

Those are the variables we will use in our analysis:

- Levy: Tax of importing and exporting the cars, the currency we used is dollars.
- Engine volume: The Engine volume of a car, measured in liter.
- Mileage: How much car has been driven, measured in miles.
- Gear box type: The type of gear box, including automatic, manual, tiptronic, and variator.
- Airbags: The number of airbags in the car.
- Price: The price of the car, measured in dollars.
- Production year: The year that the car was produced.
- Leather interior: Whether the car has leather interior.

Below are the descriptions variables that will not appear in the analysis:

- Manufacturer: The brand of cars, including hyundai, toyota, mercedes-benz, chevrolet, lexus, ford, and others.
- Category: The body types of the cars, including sedan, jeep, hatchback, minivan universal, and others.
- Fuel type: The type of fuel the car used, including CNG, diesel, hybrid, hydrogen, LPG, petrol, and plug-in hybrid.
- Cylinders: How many cylinders the car engine has.

- Drive wheels: The wheel of car to which the engine transmits its power, the types we have are 4x4, front, rear.
- Doors: The number of doors the car has.
- Wheel: Where the driving system of the car is on, the types we have are left wheel and right-hand drive.
- Color: The color of the car, including black, white, silver, grey, blue, red and other.

1.2 Research questions

Our question for the regression analysis is whether we could predict the car price based on levy, engine volume, mileage, gearbox type, and the number of airbags. The response variable is price, and the predictors are levy, engine volume, mileage, gearbox type, and the number of airbags. In those predictor, levy, engine volume, mileage, and the number of airbags are quantitative variables, and the gearbox type is categorical variable.

The question for the classification analysis is whether we could predict the car has or not has leather interior based on price, levy, production year, engine volume, mileage and number of airbags. The response variable is the car has or not has leather interior, and predictors are price, levy, production year, engine volume, mileage and the number of airbags. All the predictors are quantitative variables here.

2 Regression

2.1 Collinearity

Table 1: Collinearity of variables table

	Price	Levy	Engine.volume	Mileage	Airbags
Price	1.0e+00	5.90e-02	1.90e-02	-9.0e-03	-2.1e-01
Levy	5.9e-02	1.00e+00	6.53e-01	1.2e-02	9.7e-02
Engine.volume	1.9e-02	6.53e-01	1.00e+00	-4.0e-03	2.1e-01
Mileage	-9.0e-03	1.20e-02	-4.00e-03	1.0e+00	-5.0e-03
Airbags	-2.1e-01	9.70e-02	2.10e-01	-5.0e-03	1.0e+00

As the table 1 shows, except the levy and engine volume has a moderate correlation coefficient, the relationship between any other two variables are very weak.

2.2 Linear Model

The linear model we have is

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 I(X_4 = Manual) + \beta_5 I(X_4 = Tiptronic) \\ &\quad + \beta_6 I(X_4 = Variator) + \beta_7 X_5 + \epsilon, \end{aligned}$$

where x_1 is levy, x_2 is engine volume, x_3 is mileage, x_4 is gear box type, x_5 is number of airbags, $\epsilon \sim N(0, \sigma^2)$, and all ϵ's are independent.

Table 2: VIF of linear model predictors table

	GVIF	Df	GVIF^(1/(2*Df))
Levy	1.80e+00	1e+00	1.34e+00
Engine.volume	1.88e+00	1e+00	1.37e+00
Mileage	1.01e+00	1e+00	1.00e+00
Gear.box.type	1.09e+00	3e+00	1.01e+00
Airbags	1.09e+00	1e+00	1.04e+00

From table 2, we can see that with small VIFs all less than 2, we can assume the correlations between these variables is negligible

Table 3: Coefficient of linear model table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.94e+04	4.89e+02	3.96e+01	0.00e+00
Levy	2.63e+00	4.59e-01	5.73e+00	0.00e+00
Engine.volume	5.23e+02	2.49e+02	2.10e+00	3.50e-02
Mileage	0.00e+00	0.00e+00	-8.74e-01	3.82e-01
Gear.box.typeManual	-7.25e+03	1.47e+03	-4.91e+00	0.00e+00
Gear.box.typeTiptronic	1.15e+04	6.05e+02	1.91e+01	0.00e+00
Gear.box.typeVariator	1.79e+03	1.08e+03	1.66e+00	9.80e-02
Airbags	-1.06e+03	3.72e+01	-2.85e+01	0.00e+00

Base on the result from table 3, to evaluate whether these predictors are significant factors of the response variable price, hypothesis tests are used. Thus, seven hypothesis tests are performed:

$H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$.
$H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$.
$H_0 : \beta_3 = 0$ vs. $H_a : \beta_3 \neq 0$.
$H_0 : \beta_4 = 0$ vs. $H_a : \beta_4 \neq 0$.
$H_0 : \beta_5 = 0$ vs. $H_a : \beta_5 \neq 0$.
$H_0 : \beta_6 = 0$ vs. $H_a : \beta_6 \neq 0$.

$H_0 : \beta_7 = 0$ vs. $H_a : \beta_7 \neq 0$.

Choose $\alpha = 0.05$. The p-value coefficient shows that we will reject H_0 for β_0,β_1,β_2,β_4,β_5,β_7 because their p-value is less than 0.05. We fail to reject H_0 for β_3 and β_6 because their p-value is greater than 0.05.

In order to find the test MSE of the model, we use a 6-cross validation method.

The test MSE of the linear model is calculated using 6 fold cross validation. The result is 326989750.

2.3 Ridge Regression

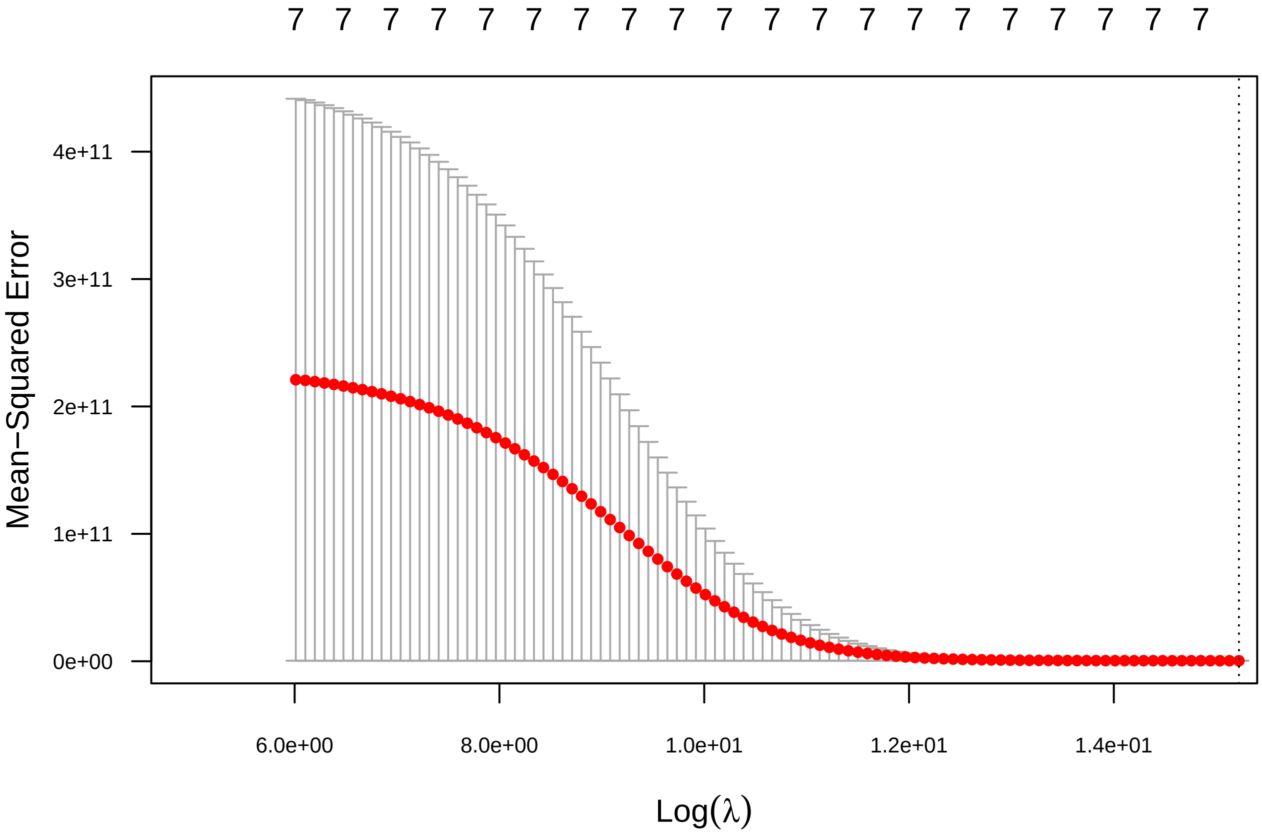

Figure 1: Ridge regression's log lambda vs. MSE

Table 4: Summary of the ridge regression model

	x
(Intercept)	1.66e+04
Levy	0.00e+00
Engine.volume	0.00e+00
Mileage	0.00e+00
Gear.box.typeManual	0.00e+00
Gear.box.typeTiptronic	0.00e+00
Gear.box.typeVariator	0.00e+00

According to Figure 1, the best λ is 4.08×10^6.

The ridge regression model is

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 I(X_4 = Manual) \\ &\quad + \beta_5 I(X_4 = Tiptronic) + \beta_6 I(X_4 = Variator) + \beta_7 X_5 + \epsilon, \end{aligned}$$

where x_1 is levy, x_2 is engine volume, x_3 is mileage, x_4 is gear box type, x_5 is number of airbags, $\epsilon \sim N(0, \sigma^2)$, and all ϵ's are independent. The MSE of ridge regression is 256510690.

3 Classification

3.1 Logistic regression

The logistic regression model is

$$p_x = P(Y = Yes|X_1 = x_1, \ldots, X_6 = x_6) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5+\beta_6 x_6}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5+\beta_6 x_6}} \tag{1}$$

Table 5: Coefficient of logistic regression model table

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.81e+02	2.01e+01	-1.40e+01	0.00e+00
Price	0.00e+00	0.00e+00	-5.96e+00	0.00e+00
Levy	-1.00e-03	0.00e+00	-8.71e+00	0.00e+00
Prod..year	1.40e-01	1.00e-02	1.40e+01	0.00e+00
Engine.volume	1.67e+00	6.40e-02	2.60e+01	0.00e+00
Mileage	0.00e+00	0.00e+00	-3.79e-01	7.04e-01
Airbags	-9.80e-02	8.00e-03	-1.26e+01	0.00e+00

Base on the equation 1, Y = leather interior(Yes, No), measures if the car has leather interior or not. X_1 = price, X_2 = levy(tax), X_3 = production year, X_4 = engine volume, X_5 = mileage in kilometers, X_6 = number of airbags.

We can see on the table 5, the error rate for this model is 0.099, which is a small value. This means logistics regression model using "binomial" looks not very bad.

The coefficients of the predictors mean that whenever X changes by 1 unit, how much log odds of leather interior will change. For example, the coefficient of engine volume is 1.673, which means that 1 unit change in engine volume will result in 1.673 change in log odds of leather interior.

3.2 LDA

Table 6: The confusion matrix of LDA

	No	Yes
No	9.00e+00	4.60e+01
Yes	8.39e+02	5.42e+03

The LDA model is

$$P(Y = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{K} f_i(x)\pi_i} \tag{2}$$

where $f_k(x)$ is normal with the same sum in each class k. From table 6, the columns of no, yes is predicted values and rows of no, yes is true values. The test error rate that we got from the linear discriminant analysis is 10.00%, which is not a very high value. This shows that this model did capture the true data precisely. However, the number of false positives is very large, which means the model will classify many cars that does not has leather inside as has leather inside. In conclusion, LDA is not a pretty good model.

3.3 QDA

Table 7: The confusion matrix of QDA

	No	Yes
No	1.7e+01	8.60e+01
Yes	8.3e+02	5.38e+03

The QDA model is same as LDA shown in equation 2 where $f_k(x)$ is normal with different $\sum_k$ in each class k, $\sum_k$ means "sum over k". From the table 7, the vertical line of no, yes is predicted values and horizontal line of no, yes is true values. The test error rate from the quadratic discriminant analysis is 10.36%, which is also not a very high value. However, the number of false positives is very large, which means the model will classify many cars that does not has leather inside as has leather inside. In conclusion, QDA is also not a pretty good model.

4 Results

4.1 Regression

The regression problem is that if we can predict the price by using the quantitative variables: levy, engine volume, mileage, airbags and categorical variable gear box type. The collinearity is checked and the simple linear model, and ridge model are used in the regression problem.

4.1.1 Collinearity

We use variance inflation factor (VIF) to check the collinearity of the predictors in the model. With all the VIFs are smaller than 2, it indicates that the interaction among the predictors are very small. Thus, we do not include any interaction between predictors.

4.1.2 Linear model

The linear model uses quantitative variables: levy, engine volume, mileage, airbags and categorical variable gear box type. as predictors. The R^2 is about 0.08, which is very small. It implies that this model only explains 8 of the error are explained by the model. With p-value smaller than 2.2×10^{-16}, which is very small. The model is significant. In conclusion, use the 6-crossed validation method, there test mean square error of 326989750.

4.1.3 Ridge model

The ridge model uses quantitative variables: levy, engine volume, mileage, airbags and categorical variable gear box type. as predictors. In ridge regression, the goal is to minimize $RSS+\lambda\sum_{j=1}^{p}\beta_j^2$ where $\lambda\sum_{j=1}^{p}\beta_j^2$ is called a penalty term, which is used to penalize large β. The best λ is chosen using cross-validation with the lowest MSE. And then the test data is used to calculate the MSE which is 256510690.

4.2 Classification

For the classification problem which is whether we could predict the car has or not has Leather interior based on price, levy, production year, engine volume, mileage and number of airbags, there is no need to check the collinearity among the predictors again since the predictors are the same as regression problem. Three different models in total are fitted.

4.2.1 Logistic regression

The logistic regression model is used to predict the probability of whether a car has leather interior. The error rate of the prediction is calculated using cross-validation. The error rate of this model is 0.099.

The model with the coefficients will become

$$\begin{aligned} p_x &= P(Y = \text{Yes} \mid X_1 = x_1, \ldots, X_6 = x_6) \\ &= \frac{e^{-281.2-1.225\times10^{-5}x_1-7.838\times10^{-4}x_2+0.1397x_3+1.673x_4}}{1+e^{-281.2+-1.225\times10^{-5}x_1+-7.838\times10^{-4}x_2+0.1397x_3+1.673x_4-2.820\times10^{-9}x_5-9.806\times10^{-2}x_6}} \end{aligned}$$

From the result of the logistic regression, the Mileage has a p-value 0.704 which is greater than $\alpha = 0.05$, so we can use backward selection to drop the Mileage.

4.2.2 LDA

The second model is LDA. The prediction will assign car to the class with highest $\delta_k(x) = x^T \sum^{-1} \mu_k - \frac{1}{2}\mu_k^T \sum^{-1} \mu_k + \log\pi_k$. And then the error rate is calculated again using cross-validation. The error rate of this model is 0.1000, which is not very large. Compare with logistic regression the error rate is large, and similar as error rate of QDA.

4.2.3 QDA

The third model is QDA. After the model is fitted, the prediction will assign car to the class with highest $\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \sum_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\sum_k| + \log\pi_k$. And then the error rate is calculated again using cross-validation. The error rate of this model is 0.1036 which is not very large. Compare with logistic regression the error rate is large, and similar as error rate of LDA.

4.2.4 Compare error rate of the models

Finally, all the models are compared using error rate. Since the goal is to predict leather interior is used in a car as correct as possible, the model with the lowest error rate is chosen, which is the first logistic model that include all the predictors.

5 Conclusions

In conclusion, if people want to choose car based on price, they should look at a car's levy, engine volume, if it has a manual or tiptronic gearbox and the number of airbags.

On the other hand, if people want to see whether a car has leather interior or not, they should look at a car's price, production year, engine volume, mileage and number of airbags. This model can help people better decide whether car has leather interior or not.