Statistical Techniques on Car Price Data

University of Derby

100568146@unimail.derby.ac.uk

Abstract.
This paper contain a detailed statistical analysis on car price data. Several statistical techniques and its assumptions are evaluated to have a clear idea of the data. Several models are bult accordingly including regular Regression models and Lasso and Ridge regression.

Keywords: T-test, ANOVA, Non parametric tests, Multicollinearity, Linear Regression, Principal Component Regression, Lasso Regression, Ridge Regression.

Contents

1 Introduction

A statistical understanding of the data is crucial for any data analytical study. A clear understanding of the what the data represent can be derived only through a statistical analysis. This understanding on various factors that build up the characteristics of a particular data is necessary because these characteristics should be the building blocks for the future studies on the data.

Aim and objectives.

The aim of the project is to produce a statistical analysis on car price data. The assumptions underlying in the various statistical tests are explored and validated. The final objective is to build an optimum model that could predict the car price and to find out the significant factors that could affect a car price.

2 Dataset

The data set used in the study is Car Price Prediction dataset from Kaggle [Kaggle]. The dataset has 26 variables and 205 data points. The data is from a Chinese automobile company who aspire to enter the US market and like to understand the factors affecting the price of a data.

2.1 Data Variable Description

The dataset contains 26 variables. The predictor variable is the variable price which gives the price of the car. Each row gives characteristics of a certain car and its price. The following figure represent the dataset and the variable type that is used in this study.

```
'data.frame':   205 obs. of  26 variables:
 $ car_ID          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ symboling       : int  3 3 1 2 2 2 1 1 1 0 ...
 $ CarName         : chr  "alfa-romero giulia" "alfa-romero stelvio" "alfa-romero Quadrifoglio" "
 $ fueltype        : chr  "gas" "gas" "gas" "gas" ...
 $ aspiration      : chr  "std" "std" "std" "std" ...
 $ doornumber      : chr  "two" "two" "two" "four" ...
 $ carbody         : chr  "convertible" "convertible" "hatchback" "sedan" ...
 $ drivewheel      : chr  "rwd" "rwd" "rwd" "fwd" ...
 $ enginelocation  : chr  "front" "front" "front" "front" ...
 $ wheelbase       : num  88.6 88.6 94.5 99.8 99.4 ...
 $ carlength       : num  169 169 171 177 177 ...
 $ carwidth        : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...
 $ carheight       : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...
 $ curbweight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 3053 ...
 $ enginetype      : chr  "dohc" "dohc" "ohcv" "ohc" ...
 $ cylindernumber  : chr  "four" "four" "six" "four" ...
 $ enginesize      : int  130 130 152 109 136 136 136 136 131 131 ...
 $ fuelsystem      : chr  "mpfi" "mpfi" "mpfi" "mpfi" ...
 $ boreratio       : num  3.47 3.47 2.68 3.19 3.19 3.19 3.19 3.19 3.13 3.13 ...
 $ stroke          : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 3.4 ...
 $ compressionratio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...
 $ horsepower      : int  111 111 154 102 115 110 110 110 140 160 ...
 $ peakrpm         : int  5000 5000 5000 5500 5500 5500 5500 5500 5500 5500 ...
 $ citympg         : int  21 21 19 24 18 19 19 19 17 16 ...
 $ highwaympg      : int  27 27 26 30 22 25 25 25 20 22 ...
 $ price           : num  13495 16500 16500 13950 17450 ...
```

3 Two-Sample t-test

A two-sample t-test is used to determine if there is any statistical evidence in determining the differences between the average prices of a car whose fuel type are petrol or diesel.

T-Test and its assumptions

The few of the assumptions in t test that the underlying distribution must follow is as follows [Htoon,2020]:

1. The observations follow a continuous distribution
2. The sample observations are independent random samples.
3. The distribution of the sample observations is normally distributed.
4. And for two-sample t-tests, the variances are equal, or observations have homoscedasticity.

Figure 1 shows that there is a slight difference in the mean price of cars whose fuel type is diesel than to whose fuel type is gas.

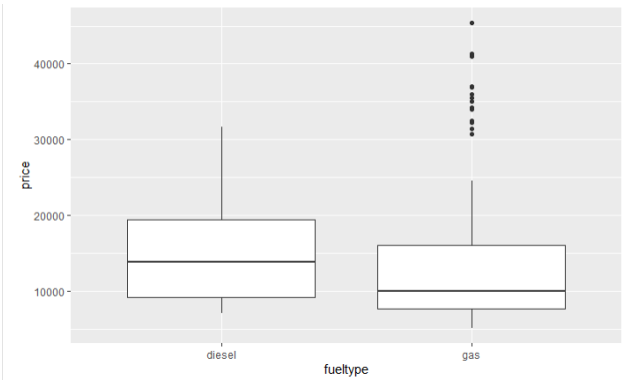

Fig-1

3.1 Wilcoxon test

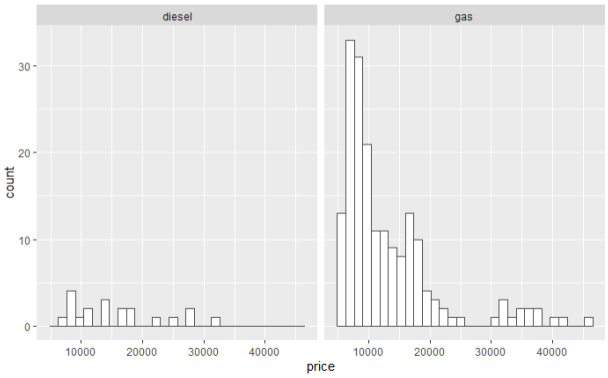

Fig-2

The distribution of the dependent variable over the variable of interest that is the fuel type suggests that there is a skewness in the data. Also, the data points are highly unequal. Hence a nonparametric test for t test is developed to check the assumptions.

The Wilcoxon rank sum test gave a p-value of 0.04617 which is slightly less than 0.05. Hence, there is statistical evidence that the means of these two groups is different.

4 Analysis of Variance

An ANOVA is used to determine the difference in car prices according to the car body type. There are five different car body type in the dataset.

carbody <chr>	variance <dbl>
convertible	125166918
hatchback	24104584
sedan	71749164
wagon	26224116
hardtop	211863184

Fig-3

Figure 3 shows the five different classes used in this test and its variance.

4.1 Assumptions

The ANOVA also follows the similar assumptions as a t-tests. The assumption of ANOVA is to have equal variance or to have no heteroscedasticity, and to have a normally distributed predictor variable. Both these conditions are first tested before using the tests.

4.2 Levene's Tests: The assessment for equality of Variances.

Homoscedasticity means having equal variance or having the same scatter. ANOVA when the sample sizes are significantly different like in this case is highly sensitive to homogeneity of variances. Levene's tests is an equality of variance tests.

The null hypothesis H0: All the groups have equal variances. And the

Alternative hypothesis H1: at least one pair of variances among these groups are unequal.

The p-value of the test is 6.299e-05, which is less than our significance level of 0.05.

Thus, we reject the null hypothesis and conclude that the variance among the five groups is not equal.

4.3 Barlett's Test for homogeneity of Variances

Another tests for checking the equality of variance is Barlett's tests. It is a statistical test that is used to determine weather or not the variances between the groups [Zach, 2020].

The null hypothesis H0: All the groups have equal variances. And the

Alternative hypothesis H1: at least one pair of variances among these groups are unequal.

p value is which is less than the alpha level of 0.05. Thus, both the tests confirms that the variances among the car body types are different.

4.4 Checking the Normality

Shapiro-Wilk test is a statistical test for checking the normality assumption that is required in both t-tests and ANOVA. The p-value is 1.849e-15 which is way less than 0.05 and hence the null hypothesis that the dataset comes from a normal distribution is rejected and the alternative hypothesis of the distribution not being a not normal distribution is accepted.

4.5 Kruskal- Wallis tests

Since two of the assumptions, homogeneity of variances and normality conditions are violated we are adapting a non-parametric test that is less sensitive to these conditions. The Kruskal Wallis tests gave a p value of 0.0001843. As the p-value is less than the significance level 0.05, we can conclude that there are significant
differences between the treatment groups.

5 Data Analysis

5.1 Regression

Regression analysis is one of the popular and oldest method to identify any relation between one or more independent variables and dependent variable. The aim is to find the optimum relationship between the independent variable that could explain most of the information withhold by the dependent variable.

The regression analysis is to estimate the f function in equation 1 that could closely describe the dependent variable [Wikipedia].

$$Y_i = f(X_{i,} \beta) + e_i \quad (1)$$

There are different potential functions that can be used as f in computing regression analysis. The researcher would need to find the potential fitting function f that is best fit for any data.

Linear regression.

One of the most used regression analysis techniques is linear regression or simple (multiple) linear regression. It assumes a linear relationship between the dependent variable and independent variables and propose that it can be explained as some linear combination of the independent variables [Wikipedia].

The model proposes a linear relationship as

$$Y_i=\beta_0 +\beta_{i*}X_i + €, \text{ where } i=1,2, 3, \ldots., n \text{ (number of data points)} \quad (2)$$

The coefficients are obtained by the popular least square measures that help as to get the optimum equation minimizing the error in the data.

5.2 Multicollinearity

Multicollinearity is a phenomenon in which two or more variables in a multiple linear regression is highly correlated. Inclusion of collinear variables can inflate the regression coefficients. Figure 4 represents a correlation matrix for every numerical variable in the data. We can evidently conclude that there are variables which are highly correlated and these multicollinearity needs to be treated.

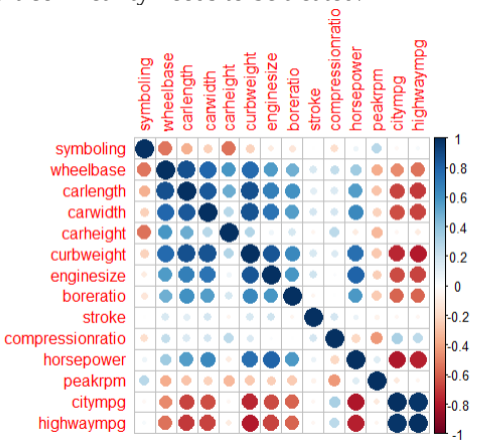

Fig-4

A ‘findCorrelation’ function from the package ‘caret’ is used to find the variables that have correlation greater than 0.80. Figure 5 is all the variables that have a correlation greater than 0.80. These variables are excluded from the original dataset before building a linear regression model.

```
[1] "curbweight" "carlength"  "highwaympg" "enginesize" "citympg"
```

Fig-5

Linear regression

A linear regression analysis to model to predict the price of the cars is executed using the dataset after removing the correlated variables. The model gave a p-value of 1.887e-05 and an adjusted R-square of 0.9566.

Checking the basic assumptions of OLS model

The following are the assumptions of a Ordinary Least Square model and these assumptions are checked. The model can be used in future predictions only if these assumptions are fulfilled

Assumptions of Linear regression

1. X and Y variables have linear relation (linear scatter diagram)
2. Errors/ residuals are normally distributed.
3. Errors are independent/ no autocorrelation between errors.
4. Constant error variance
5. Avoid multicollinearity between the predictors.

Errors/ residuals are normally distributed.

The normality of the residuals is verified using a Q-Q plot. Figure 5 is the Q-Q plot. The Q-Q plot suggests that the residuals are normally distributed.

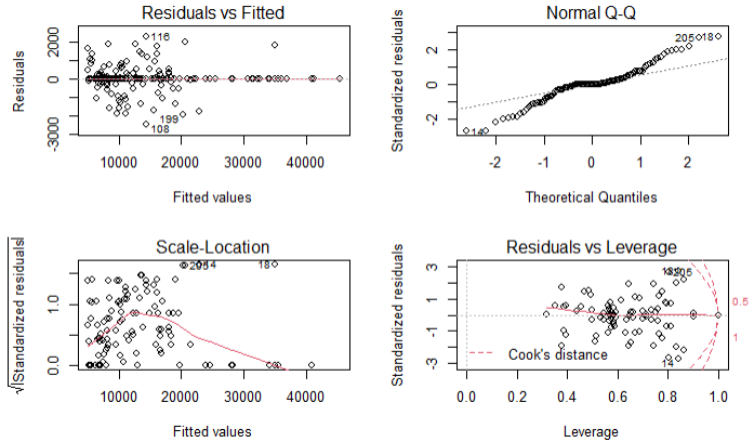

Fig-5

Errors are independent/ no autocorrelation.

A Durbin- Watson test is used to determine if the errors are independent.

The null hypothesis is that there is no autocorrelation between the errors. And since the p-value is 0.7778 is greater than 0.05, we accept the null hypothesis. Hence, it is concluded that the linear models' errors are independent.

Constant error variance

The constant error variance is checked using a NCV test. The p-value is 0.48874 which is greater than 0.05 hence we accept the null hypothesis that the errors have constant variance.

5.3 Principal Component Regression

Principal Component Analysis is a class of dimensionality reducing techniques which reduces the input features by finding the correlation between them and transforming them into a set of unrelated linear features.

A Principal component regression model is built on the data set because of the multicollinearity present in the data. The number of components is decided from the figure 6, where it is evident that after a certain number of components the RMSE value is almost a constant. The number of components ais this taken as eight.

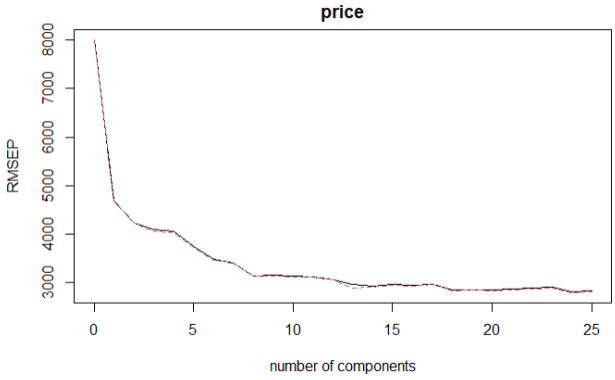

Fig-6

The figure 7 also justifies the decision to be eight. The figure below gives the R-square value along the number of components. And after a certain point (in this case eight) it is evident that the R-square is almost constant.

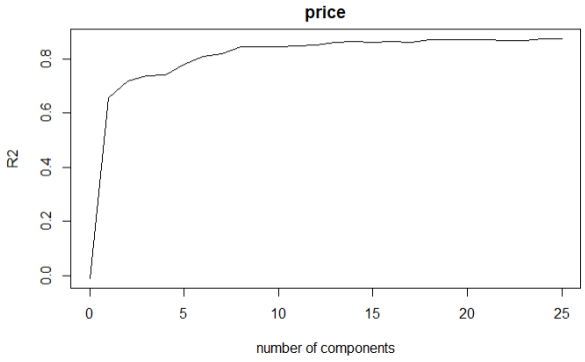

Fig-7

The model is built using eight components and have got an RMSE value of 3200.

5.4 Ridge Regression

Ridge regression is a technique that can be used when the data suffer from multicollinearity. Multicollinear variables inflate the coefficients and thus the true value can never be estimated. The biggest benefit of using a ridge regression is that it reduces the test mean square error [Zach,2020].

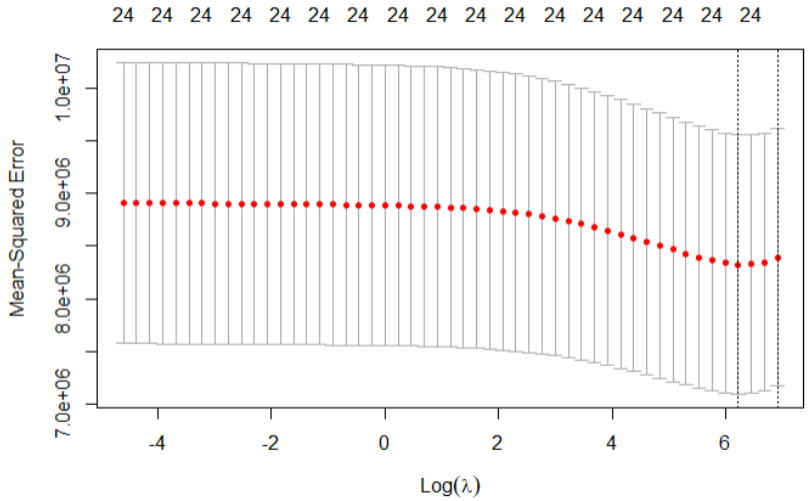

Fig-8

A 'glmnet' package is used to create a ridge regression model. λ is calculated by giving different values for lambda in a range and the test MSE of each model is calculated and the lambda that gives the lowest MSE is chosen [Zach,2020]. Figure 8 shows the MSE value for different lambda. The lambda that gives the lowest MSE is found to be 630.95. The model build using this lambda has an r-square of 0.9723. That implies 97% of the variation was able to be explained by the ridge regression.

5.5 Lasso Regression

Lasso regression is another regression technique which can be used when there is multicollinearity in the data. The idea behind a Lasso regression or the least absolute shrinkage and selection operator is to utilizes the shrinkage [datascievo, 2021]. Lasso and ridge regression uses estimation method unlike regression which uses inference.

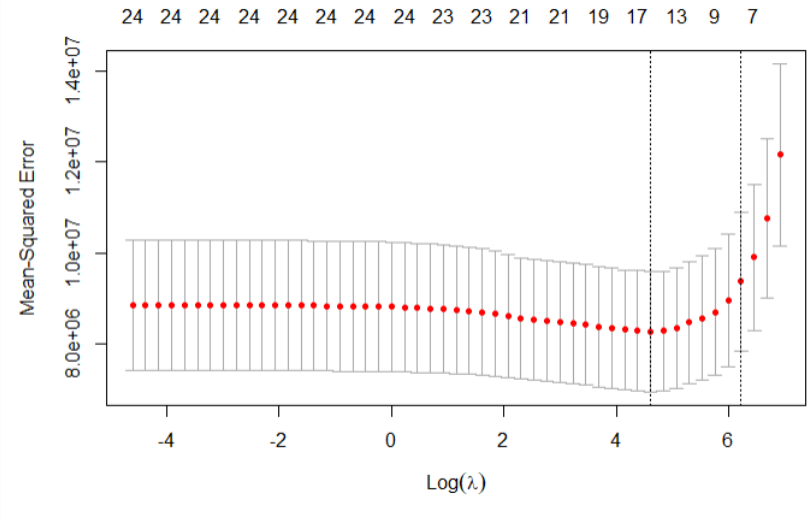

Fig-9

Similarly like in Ridge regression 'glmnet' is used to model a lasso regression but with alpha as 1. The optimum lambda is obtained by minimizing the mean-square error and is found to be 100. Figure 9 shows the MSE value over different lambda for lasso regression. The model built upon has an adjusted R-square of 0.9730818.

6 Conclusion

A detailed data analysis was conducted on the Car price data. The data was found to have multicollinearity between the variables and so the data was treated for that. Models were built either after treating the multicollinearity or my using techniques which incorporate multicollinearity.

A linear regression is built after dropping all the variables which have a high collinearity and the adjusted r-square for the model was 0.9566.

A Principal component regression is used to incorporate the multicollinear variables and the model got a RMSE value of 3200.

A Ridge regression for multicollinear data set is built and the model gave and adjusted r-square 0.9723.

And finally, a Lasso regression is built, and the model r-square is 0.9730818.

7 References

1. Wikipedia contributors, "Regression analysis," Wikipedia, The Free Encyclopedia, Available from : Regression analysis - Wikipedia (accessed January 3, 2022).
2. Kaggle, (2020), 'Car Price Prediction', Available at: Car Price Prediction Multiple Linear Regression | Kaggle, (Accessed: 22 December 2021).
3. Frigaard, Martin., (2019),' Diagnosing the accuracy of your linear regression in R', Available at: Diagnosing the accuracy of your linear regression in R - Storybench, (Accessed: 14-Jan-2022).
4. UC Business Analytics R Programming Guide ,'t-test: Comparing Group Means', Available at t-test: Comparing Group Means · UC Business Analytics R Programming Guide (uc-r.github.io) (Accessed: 4-Jan-2022).
5. Htoon, Kyaw.,' Levene's Test: The Assessment for Equality of Variances',Available at: Levene's Test: The Assessment for Equality of Variances | by Kyaw Saw Htoon | Medium (Accessed : 12 January 2022).
6. Zach,(2019), 'How to Conduct Levene's Test for Equality of Variances in R', Available at: How to Check ANOVA Assumptions - Statology (Accessed 5 December 2021).
7. Zach,(2019),' How to Check ANOVA Assumptions', Available at: How to Conduct Levene's Test for Equality of Variances in R - Statology (Accessed 5 December 2021).
8. Taylor,Jeremy., (2020) 'Statistical Soup: ANOVA, ANCOVA, MANOVA, & MANCOVAAvailable at: Statistical Soup: ANOVA, ANCOVA, MANOVA, & MANCOVA — Stats Make Me Cry Consulting (Accessed: 31 December 2021).
9. Great Learning Team, (2020), 'What is Ridge Regression?', Available at: Ridge Regression Definition & Examples | What is Ridge Regression? (mygreatlearning.com) (Accessed: 17 January 2022)
10. Zach., (2019) 'Introduction to Ridge Regression, Available at: Introduction to Ridge Regression - Statology (Accessed :10 Jan 2022).
11. Zach., (2019) 'Introduction to Lasso Regression, Available at: Introduction to Lasso Regression - Statology (Accessed :15 Jan 2022).
12. Zack,(2019),'A Complete Guide to Stepwise Regression in R',Available at :https://www.statology.org/stepwise-regression-r/, Accessed on :4-Jan-2022