Cars93 dataset analysis

Emanuel Frátrik

20.7.2021

1 Introduction

In this report we will analyze fuel economy of cars in highways with three types of drive-train, namely rear, front and four wheels drive-train denoted as 4WD. Then we will compare prices of cars according to its origin. And finally we will look at relation between airbags type and origin of cars. For all three performed analyses we have used Cars93 dataset.

2 Analysis and results

2.1 Descriptive data analysis

Whole dataset contains 93 observations (rows) with 27 descriptive variables (columns). For purposes of our analysis we have chosen only five variables from dataset, namely DriveTrain, MPG.highway, Origin, Airbags and Price. Price in this case is meant as average of minimal base price of car and its maximal possible price. In table 1 we can see summary statistics of highway MPG (MPG.highway) per group of drive-train (DriveTrain) type. We need to point out that sample sizes in each group are sharply unequal. In figure 1 we can see boxplot according to three types of drive-train showing central tendency and variability in these three groups of drive-train.

Similar plot is showed in figure 3 for prices (Price) of cars according to their origin (Origin). And also descriptive statistics of prices can be seen in table 2. In case of prices we see that there is almost equal sample size in two groups according to origin of car. It seems that mean price for USA and non-USA cars is very similar but different in variance.

In table 3 one can see frequency statistics of categorical variables Origin and AirBags. There we can see that most of cars has driver only airbags but the second largest group of cars are those without airbags.

Table 1: Summary statistics for highway MPG per drive-train groups

DriveTrain	mean	variance	median	sample size
4WD	25.800	32.1778	24	10
Front	30.239	29.2754	29	67
Rear	26.312	4.8958	26	16

Table 2: Summary statistics for price of cars according to its origin

Origin	mean	variance	median	sample size
USA	18.573	61.1041	16.3	48
non-USA	20.509	127.8426	19.1	45

Table 3: Summary statistics for Airbags vs Origin data

Characteristic	N = 93
Origin	
USA	48 (52%)
non-USA	45 (48%)
AirBags	
Driver & Passenger	16 (17%)
Driver only	43 (46%)
None	34 (37%)

[1] n (%)

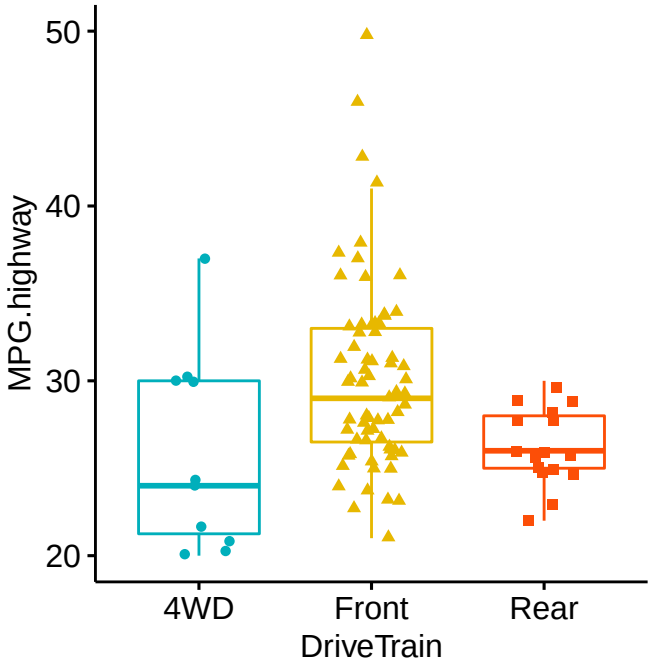

Figure 1: Boxplot of each group of drive-train

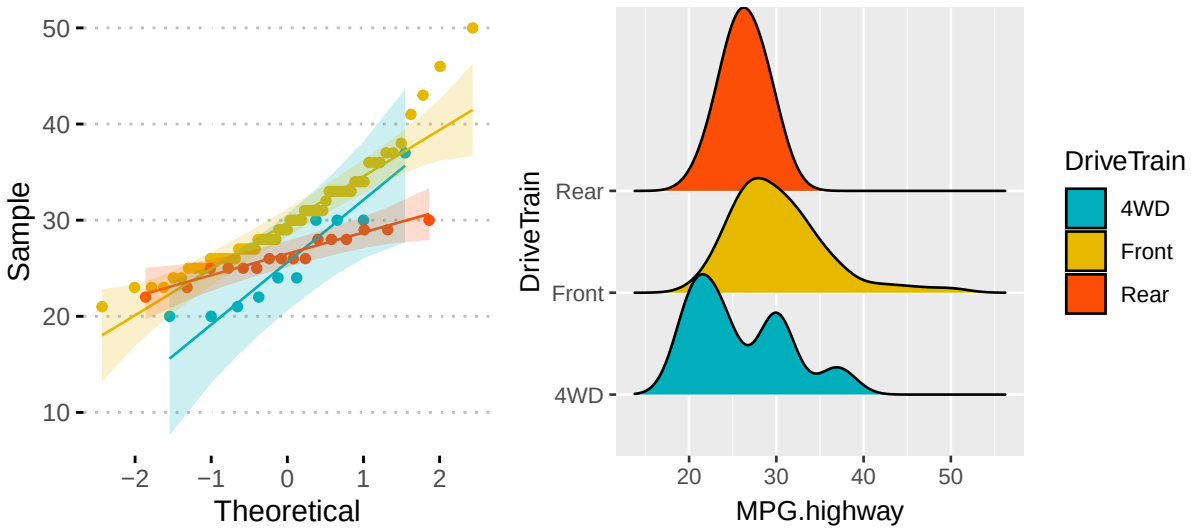

Figure 2: Normal Q-Q plot (left) and density plot (right) of highway MPG per group of drive-train

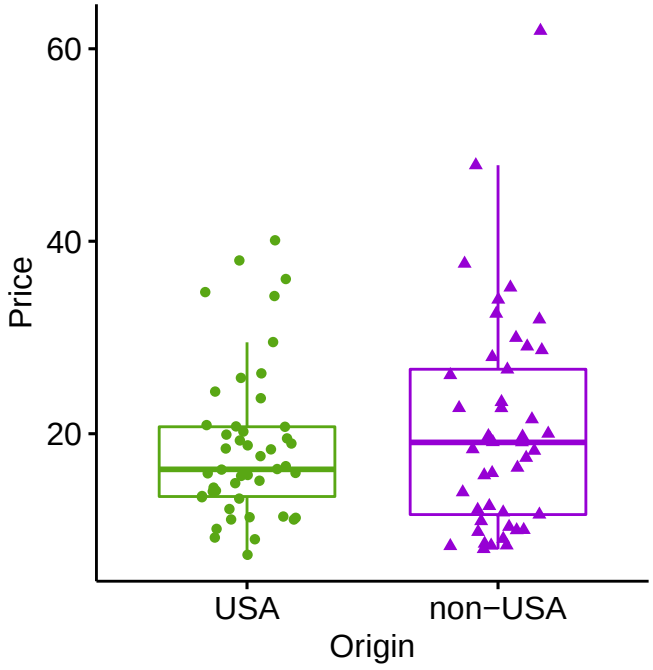

Figure 3: Boxplot of prices of cars according to its origin

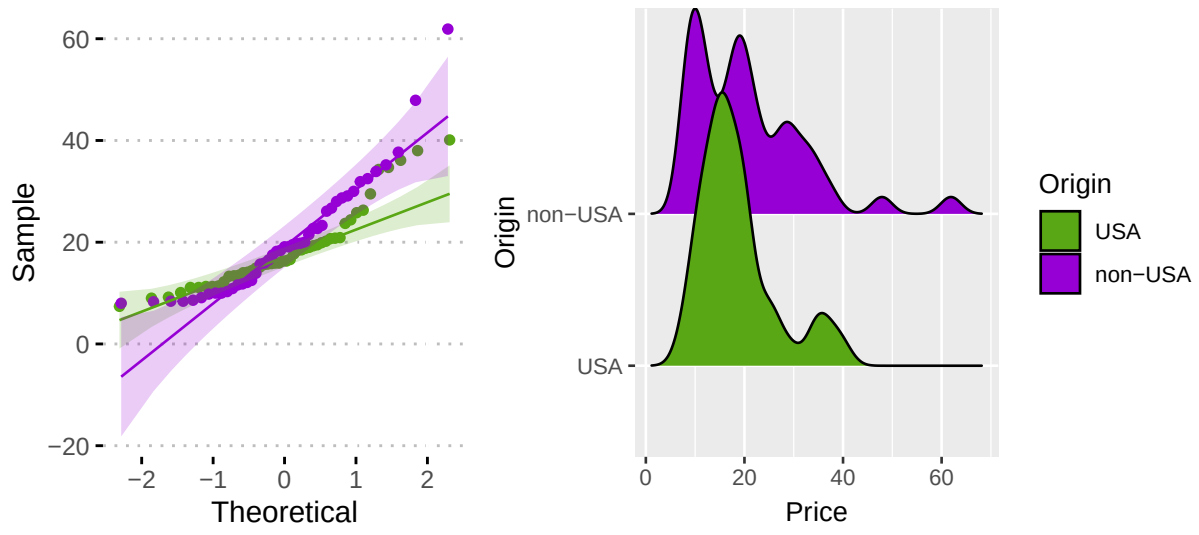

Figure 4: Normal Q-Q plot (left) and density plot (right) of prices of cars according to its origin

2.2 Inferential analysis

2.2.1 Analysis of fuel economy

As we mentioned in introduction our goal is to analyze fuel economy of cars in highway according to used drive-train. We want therefore perform Fisher's one-way ANOVA to find out whether highway MPG is equal according to type of drive-train used in vehicles. Our null hypothesis looks like

$$H_0 : \mu_{front} = \mu_{rear} = \mu_{4WD}$$

$$H_1 : otherwise$$

First of all we performed tests to check whether the assumptions of ANOVA were met. We assume that independence of observations was ensured by experiment setting. Homogeneity of variances between groups can be checked visually in boxplot in figure 1. According to this boxplots it seems that observations may have different variance among groups. We also tested this assumption using Leven's test (see table 4) with significant result with p-value of 0.025 and therefore we rejected null hypothesis about homogeneity of variances between groups. Normality of observations in each group can be also checked by looking at q-q plot or density plot. Both plots are showed in figure 2. According to this plots it seems that assumption about normality of data in groups is deviated a bit. Although Shapiro-Wilk's normality test gives significant results with p-value of 0.0002 (see table 5) for group "Front" we still did not consider this as severe violation of normality assumption and therefore we proceeded with Welch's ANOVA instead of Fisher's ANOVA which does not assume same variances among groups. Summarized results of Welch's ANOVA can be seen in table 6. As Welch's test gives significant results we reject null hypothesis about equal means and continue with Games-Howell's test to find out which groups are different from each other significantly. Games-Howell's test was used instead of Tukey's HSD method because it does not assume same variances and sample sizes between groups. As we can see in table 7 significant results occurred in pair of Rear and Front drive-train type with p-value almost equal to zero. Finally we performed also non-parametric Kruskal-Wallis test which also turned out to be significant with p-value of 0.0015 (see table 8). Dunn's test was chosen as optimal for post-hoc test which resulted in slightly different results in comparison to Games-Howell's test. Dunn's test

Table 4: Results of Leven's test of homogeneity of variances based on mean for highway MPG

Df	F statistic	p-value
2	3.829	0.0254
90		

Table 5: Results of Shapiro-Wilk's test of normality within groups of drive-train

DriveTrain	statistic	p-value
4WD	0.879	0.1269
Front	0.914	0.0002
Rear	0.948	0.4643

showed significant difference in pair of 4WD and Front drive-train type with p-value of 0.0122 in addition to Games-Howell§s test.(see table 9).

Table 6: Results of Welch's ANOVA

Df	F statistic	p-value
2	10.725	5e-04
90		

Table 7: Results of Games-Howell's post-hoc test

comparison	mean diff.	conf.int.lower	conf.int.upper	t statistic	Df	p-value
Front-4WD	4.439	-0.6864	9.5640	2.3219	11.581	0.092310
Rear-4WD	0.512	-4.5761	5.6011	0.2730	10.735	0.959895
Rear-Front	-3.926	-5.9974	-1.8553	4.5552	60.424	0.000076

Table 8: Results of non-parametric Kruskal-Walis test

Df	F statistic	p-value
2	12.979	0.0015
90		

Table 9: Results of Dunn's post-hoc test

comparison	Z statistic	p-value
4WD - Front	-2.506	0.0122
4WD - Rear	-0.070	0.9444
Front - Rear	2.952	0.0032

2.2.2 Analysis of price of cars

In this section we have analyzed prices of cars according to their origin (USA, non-USA). Optimal test to be chosen to compare prices would be two samples t-test or Welch's test. We have therefore firstly analyzed variances of prices. From boxplot in figure 3 it seems that variances of both groups may be different. Leven's test was used to test hypothesis about homogeneity of variances of prices with significant result with p-value

of 0.0425 (see table 10). After this finding we proceeded with analysis of normality visually in q-q and density plots (see figure 4) and then we performed Shapiro-Wilk's test. According to this test the hypothesis about normality of samples was rejected with p-values equal almost to zero (see table 11). As we cannot use t-test we have chosen non-parametric Kolmogorov-Smirnov two samples test and asymptotic test of means to test following hypotheses with significance level of 0.05

$$H_0 : distribution_{USA} = distribution_{non-USA}$$

and

$$H_0 : \mu_{USA} = \mu_{non-USA}$$

$$H_1 : otherwise$$

Both tests given non-significant results with p-values of 0.517 for Kolmogorov-Smirnov test and 0.34 for asymptotic test also with wide confidence interval (see tables 12 and 13).

Table 10: Results of Leven's test of homogeneity of variances based on mean for prices according to origin of cars

Df	F statistic	p-value
1	4.234	0.0425
91		

Table 11: Results of Shapiro-Wilk's test of normality of prices according to origin of cars

Origin	statistic	p-value
USA	0.884	2e-04
non-USA	0.878	2e-04

Table 12: Results of two samples Kolmogorov-Smirnov's test for prices

statistic	p-value
0.169	0.517

Table 13: Results of asymptotic two sample test of means for prices

statistic	conf. int. lower	conf. int. upper	p-value
-0.954	-5.911	2.0394	0.34

2.2.3 Airbags analysis

Finally we briefly analyzed contigency table for airbags vs origin which can be found in table 14. This table contains two groups according to origin of cars and three categories of airbags types. We wanted to know whether there is same distributions of airbags according to origin of cars. In other words we have analyzed homogeneity of contigency table. We have assumed that all observations are independent which was ensured by experiment setting and therefore we can test independence of groups vs categories which is equivalent to homogeneity of cross table. To perform test of independence we have used Chi squared test which turned out to be non-significant with p-value of 0.7864 (see table 15).

Table 14: Cross-table for AirBags vs Origin

Characteristic	Origin		Total
	USA	non-USA	
AirBags			
Driver & Passenger	9 (9.7%)	7 (7.5%)	16 (17%)
Driver only	23 (25%)	20 (22%)	43 (46%)
None	16 (17%)	18 (19%)	34 (37%)
Total	48 (52%)	45 (48%)	93 (100%)

Table 15: Results of Chi squared test of homogeneity of distribution of airbags according to origin of cars

Df	chi2 statistic	p-value
2	0.481	0.7864

3 Conclusion

According to performed Welch's ANOVA we can conclude that highway MPG significantly differs among groups of cars with three types of drive-train with p-value of 0.001. Games-Howell's test then showed significant difference in highway MPG only between groups Rear and Front with p-value almost equal to zero. Similar finding was given by non-parametric Kruskal-Wallis test with p-value of 0.002. Dunn's test then revealed that there is also significant difference between groups Front and 4WD with p-value 0.012 in addition to Games-Howell's test. As the normality requirement was violated we consider Kruskal-Wallis and Dunn's test more trustworthy. Finally we can conclude that cars with front drive-train have higher mean MPG or in other words lower fuel economy in highways in comparison to cars with rear and 4WD drive-train and therefore we can say that cars with front drive-train are more economical in highways.

4WD	Rear	Front
25.82	26.31	30.24

Analysis of prices of cars according to their origin was performed by test of means of prices with asymptotic test and test of distributions with Kolmogorov-Smirnov test. Both tests turned out to be non-significant with p-values of 0.340 for asymptotic test of means and 0.517 for Kolmogorov-Smirnov test. As we failed to reject both of this hypotheses we can only conclude that our data does contain no evidence to say whether prices of cars are same or not according to their origin. Finally, analysis of homogeneity of distribution of airbags according to origin of cars was performed using chi squared test which has given also non-significant results with p-value of 0.79 and we therefore fail to reject hypothesis about homogeneity of distribution of airbags on significance level of 0.05.