Cars93 dataset analysis

Emanuel Frátrik

19.7.2021

1 Introduction

In this report we will perform analysis to find out whether highway MPG (miles per gallon) is same for all three types of drive-train, namely rear, front and four wheels drive-train denoted as 4WD. Used data comes from Cars93 dataset.

2 Analysis and results

2.0.1 Descriptive data analysis

Whole dataset contains 93 observations (rows) with 27 descriptive variables. For purposes of our analysis we have chosen only two variables from dataset, namely DriveTrain and MPG.highway. 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.

2.0.2 Inferential analysis

As we mentioned in introduction our goal is to 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 is therefore

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

$$H_1 : otherwise$$

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 2) with significant result with p-value of 0.025 and therefore we rejected 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

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

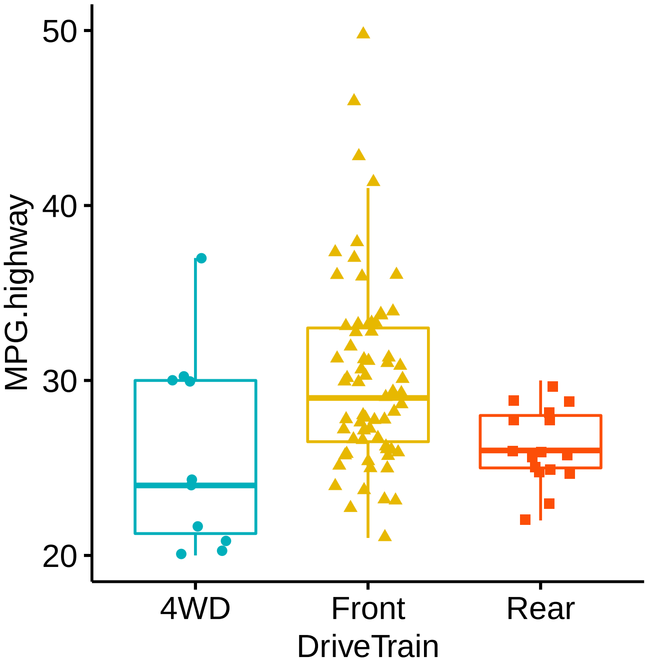

Figure 1: Boxplot of each group of drive-train

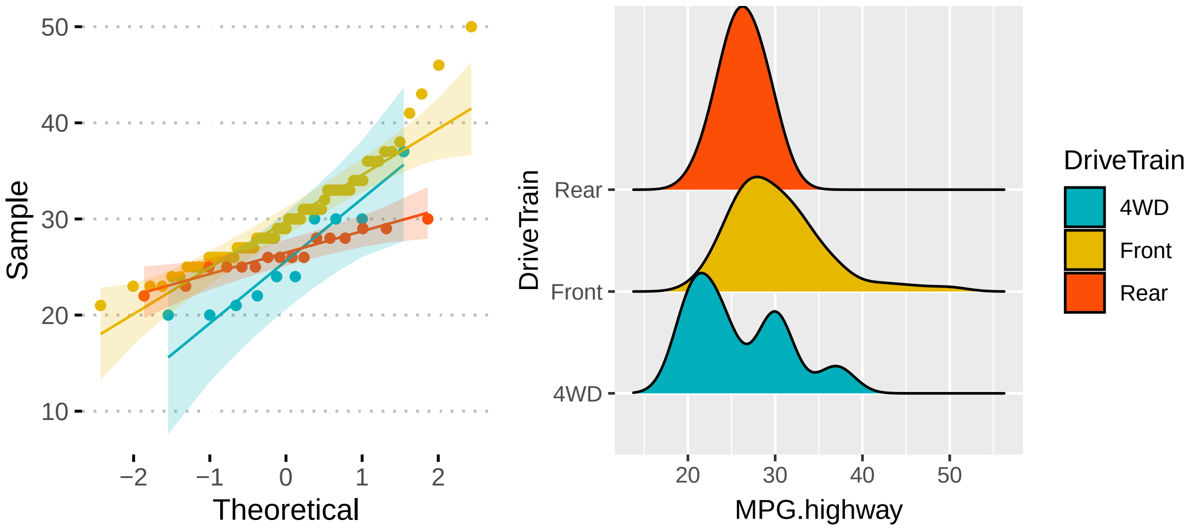

Figure 2: Normal Q-Q plot (left) and density plot (right) of highway MPG per group of drive-train

Table 2: Results of Leven's test of homogeneity of variances based on mean

Df	F statistic	p-value
2	3.829	0.0254
90		

Table 3: Results of Shapiro-Wilk's test of normality within groups

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

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 3) 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 4. 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 5 significant results occurred in pair of Rear and Front drive-train type with p-value almost equal to zero.

Table 4: Results of Welch's ANOVA

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

Table 5: Results of Games-Howell's post-hoc test

group pair	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

3 Conclusion

According to performed ANOVA we can conclude that highway MPG significantly differs among groups of cars with three types of drive-train with p-value of 0.0005. Games-Howell's test then showed significant difference in highway MPG only between groups Rear and Front with p-value almost equal to zero. Also we can conclude that cars with front drive-train have lower fuel economy in highways in comparison to cars with rear drive-train.

Rear	4WD	Front
26.31	25.82	30.24