Analysis of Fuel Efficiency in Cars

Stephanie Denis

17 April 2017

1 Executive Summary

Motor Trend, a US magazine about the automobile industry, would like to measure the fuel efficiency of cars by analyzing the relationship between miles per gallon (MPG) and aspects of design and performance. They are particularly interested in the following two questions:

1. Is an automatic or manual transmission better for MPG?
2. What is the difference in MPG between automatic and manual transmissions?

In order to answer these questions, we performed exploratory data analyses and applied statistical methods, such as hypothesis testing and regression analysis. When comparing the two transmission types, we found that manual cars have higher MPG and are more fuel efficient than automatic cars. However, when we control for other factors, such as cylinders, horse power and weight, the difference in MPG decreases and is no longer statistically significant.

2 Exploratory Data Analysis

We used the `mtcars` data set, which consists of 32 cars (1973–74 models) and 11 variables that cover aspects of car design and performance, and includes our outcome variable, miles per gallon (MPG). (See Appendix I, Table 1). When comparing fuel efficiency across transmission type, we find that manual cars have a higher average MPG (24.4) than automatic cars (17.1). At first glance, this suggests that manual cars are more fuel efficient. When we look at a box plot comparing the two groups, we see that the two IQRs (the two rectangles) do not overlap along the vertical axis, which implies that the two population means differ beyond just random variation. (See Appendix I, Figure 1).

3 Hypothesis Testing

Using hypothesis testing and Welch's Two-Sample T-test, we are able to infer about the differences in mean between the two groups. The data meets the conditions for conducting a hypothesis test as shown here:

$H_0 : \mu_{Automatic} = \mu_{Manual}$; $H_A : \mu_{Automatic} < \mu_{Manual}$

1. Independence − i.i.d random sample of cars representing less than 10 percent of all US cars.
2. Sample size − large enough sample with $n = 32$.
3. Error rate − probability of a Type 1 error rate is $\alpha = 0.05$.

The average difference in MPG between the two groups is -8.04. The 95 percent Student's T-confidence interval [-11.3, -3.2] is entirely below zero, confirming that both groups differ beyond random variation. Additionally, given a p-value of 0.001 (lower than our error rate of $\alpha = 0.05$), we reject H_0 in favor of H_A that the mean miles per gallon is *lower* for automatic cars. In other words, there is strong evidence that manual cars are more fuel efficient than automatic cars.

4 Simple Linear Regression

In order to quantify the difference between the two groups, we start with a simple linear regression model with `mpg` as our outcome variable and `am` as our predictor. The expected *increase* in the mean MPG for manual cars is given by the slope, 7.24, which is statistically significant at the 1% confidence level. Therefore, we reject H_0 in favor of H_A that the mean miles per gallon is *higher* for manual cars. The intercept, 17.15, is the expected mean MPG for automatic cars (the reference level). The adjusted R^2 is 0.34, meaning that only 34% of the variability is explained by this model. (See Appendix I, Table 2).

5 Multiple Linear Regression

Next, we augment our baseline model with other aspects of design and performance. We use a multivariate linear model selected by the stepwise forward and backward model selection strategy with the *AIC* as our criterion. The model selection criteria chose the model that includes `am`, `cyl`, `hp`, `wt`. It has the highest adjusted R^2, 0.84, among all estimated models in the procedure and explains 84% of the variability found in the model. Moreover, we compared both models using analysis of variance and concluded that `cyl`, `hp`, `wt` are significant factors in explaining `mpg`. (See Appendix 1, Tables 2).

The intercept, 33.71, is the expected mean MPG for automatic cars with a 4 cylinder engine when all other variables equal 0. The expected *increase* in the mean MPG for manual cars is 1.81, holding everything else constant. It is smaller than the increase in our base model, 7.24, and is no longer statistically significant at the 5% confidence level. Therefore, we fail to reject H_0 and conclude that the data do not provide convincing evidence of a difference in the mean MPG between manual and automatic cars.

The interpretation of the remaining coefficients suggests that, while holding all other variables constant, certain aspects of design and performance have high predictive powers and are negatively associated with MPG. An additional 1,000 lbs in weight will decrease mean MPG by -2.5. Similarly, an increase in cylinders from 4 to 6, lowers mean MPG by -3.03. These results make intuitive sense as heavier cars with bigger engines require more fuel to pull their weight, making them less fuel efficient.

6 Model Diagnostics

The diagnostic plots suggest that the residuals are randomly scattered around the zero line and appear to be normally distributed, as they fall close to the diagonal line. Nevertheless, there are points that appear to be influential, i.e. the Chrysler Imperial and the Toyota Corona shown in the bottom right plot. Their presence in the sample could be affecting the model's fit. (See Appendix I, Figure 2).

7 Conclusion

We compared the MPG consumption between automatic cars and manual cars to determine which transmission type had better fuel efficiency. We find that on average manual cars consume 7.24 more miles per gallon than automatic cars, pointing to more fuel efficiency. However, when we expand our model to include other variables, such as cylinders, horse power and weight, we find that the average difference in MPG for manual cars drops to 1.81 and is no longer statistically significant.

8 Appendix I: Tables and Figures

Table 1. Variable List

variable	description
`mpg`	Miles per U.S. gallon
`cyl`	Number of cylinders
`disp`	Displacement (cu.in.)
`hp`	Gross horsepower
`drat`	Rear axle ratio
`wt`	Weight (1,000 lbs)
`qsec`	1/4 mile time
`vs`	V/S (0 = v-engine, 1 = straight engine)
`am`	Transmission (0 = automatic, 1 = manual)
`gear`	Number of forward gears
`carb`	Number of carburetors“

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 2: Multivariate Analysis of Car Fuel Efficiency

	Dependent variable:	
	Miles per gallon	
	(1)	(2)
Manual	7.245^{***}	1.809
	(1.764)	(1.396)
6-Cylinder		-3.031^{**}
		(1.407)
8-Cylinder		-2.164
		(2.284)
Horse Power		-0.032^{**}
		(0.014)
Weight (1,000 lbs)		-2.497^{***}
		(0.886)
Constant	17.147^{***}	33.708^{***}
	(1.125)	(2.605)
Observations	32	32
R^2	0.360	0.866
Adjusted R^2	0.338	0.840
Note:	*p<0.1; **p<0.05; ***p<0.01	

Figure 1. MPG by Transmission Type

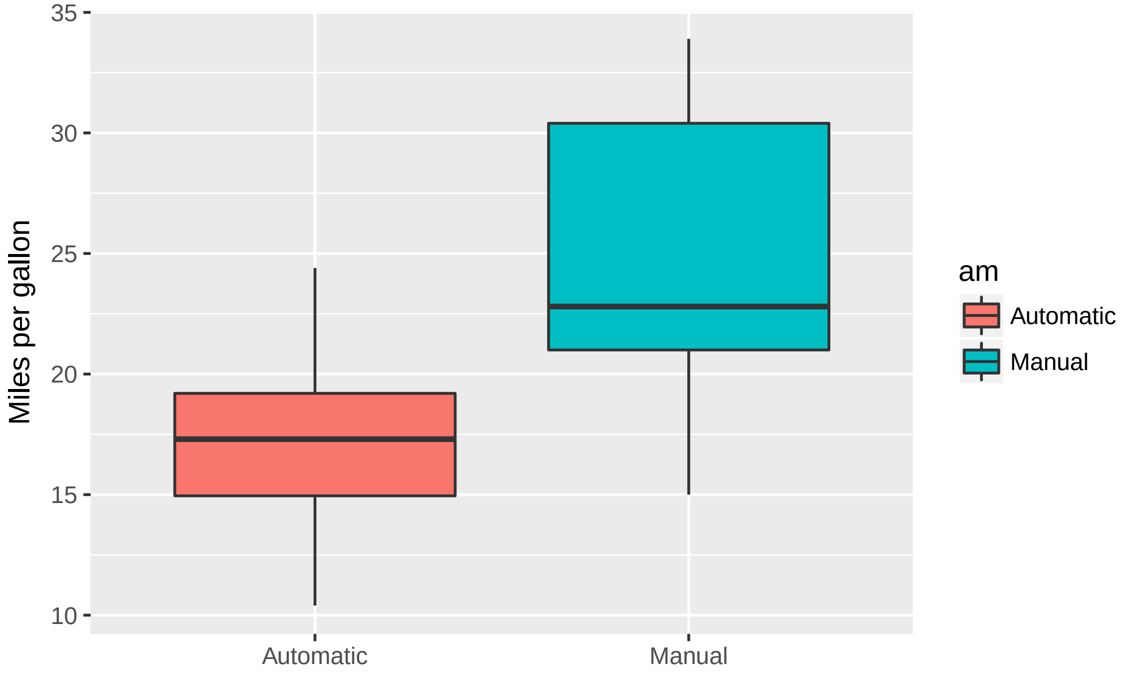
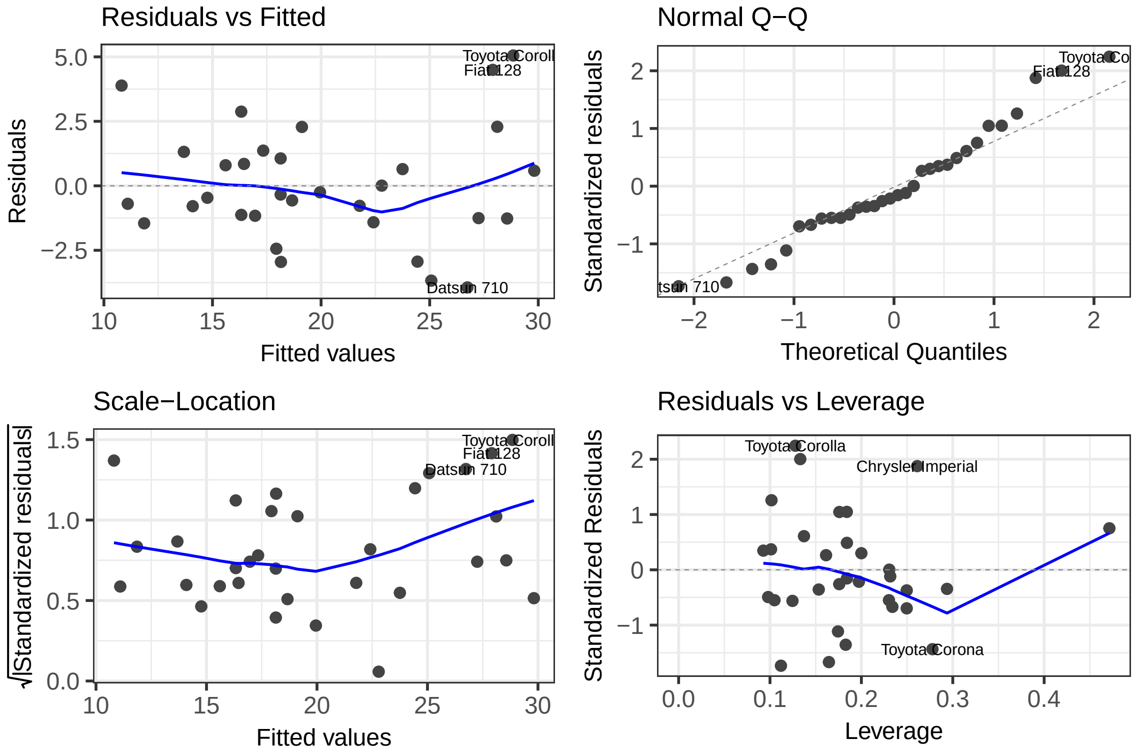


Figure 2. Residual Diagnostic Plots



9 Appendix II: R Code

```
#preamble for chunks
knitr::opts_chunk$set(echo=FALSE, results="hide", message=FALSE, warning=FALSE)
#preamble for graphics
library(knitr)
opts_chunk$set(fig.path='figure/graphics-', cache.path='cache/graphics-',
                 fig.align='center', external=TRUE, echo=FALSE, warning=FALSE,
                 fig.pos='H')
a4width<- 8.3
a4height<- 11.7

#load data and packages
data(mtcars); library(plyr); library(dplyr); library(ggplot2)

#summary statistics by group
mtcars$am <- factor(mtcars$am, labels = c("Automatic","Manual"))
mn_auto <- round(mean(mtcars$mpg[mtcars$am == "Automatic"]), 1)
mn_man <- round(mean(mtcars$mpg[mtcars$am == "Manual"]), 1)

#boxplot
ggplot(aes(x = am, y = mpg), data = mtcars) +
        geom_boxplot(aes(fill = am)) +
        labs(x = "", y = "Miles per gallon") +
        ggtitle("Figure 1. MPG by Transmission Type") +
        theme(plot.title = element_text(face="bold", hjust = 0.5))

#t-test: automatic vs manual
g1 <- mtcars$mpg[mtcars$am == "Automatic"]
g2 <- mtcars$mpg[mtcars$am == "Manual"]
difference <- g1 - g2
dmn <- round(mean(difference), 2)
test <- t.test(mpg ~ am, data = mtcars,
               var.equal = FALSE, paired = FALSE, conf.level = .95)
lo_ci <- round(test$conf.int[1], 1)
hi_ci <- round(test$conf.int[2], 1)
pval <- round(test$p.value, 3)

#estimate base model
basemodel <- lm(mpg ~ am, data = mtcars)
b0 <- round(coef(basemodel)[1], 2)
b1 <- round(coef(basemodel)[2], 2)
r_sq <-  round(summary(basemodel)$r.squared,2)
adj_r <- round(summary(basemodel)$adj.r.squared,2)

#convert variables from numeric to factor
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$vs   <- factor(mtcars$vs, labels = c("V","S"))
#stepwise selection model
fullmodel <- lm(mpg ~ ., data = mtcars)
bestmodel <- step(fullmodel, direction = "both")
#re-ordering variables for regression table
bestmodel <- lm(mpg ~ am + cyl + hp + wt, data = mtcars)
b0_best <- round(coef(bestmodel)[1], 2)
b1_best <- round(coef(bestmodel)[2], 2)
```

```r
b2_best <- round(coef(bestmodel)[3], 2)
b5_best <- round(coef(bestmodel)[6], 2)
r_sq_best <- round(summary(bestmodel)$r.squared,2)
adj_r_best <- round(summary(bestmodel)$adj.r.squared,2)

#analysis of variance
anova(basemodel, bestmodel)

#regression table
library(stargazer, quietly = TRUE)
stargazer(basemodel, bestmodel, header=FALSE,
          title = "Multivariate Analysis of Car Fuel Efficiency",
          dep.var.labels = "Miles per gallon",
          covariate.labels = c("Manual", "6-Cylinder", "8-Cylinder",
                               "Horse Power", "Weight (1,000 lbs)"),
          omit.stat = c("LL","ser","f"), no.space = TRUE)

#residual plots
library(ggfortify)
autoplot(bestmodel, label.size = 2.15) + theme_bw() +
        theme(plot.title = element_text(size = 10),
              axis.title = element_text(size = 9))

#leverage points
leverage <- hatvalues(bestmodel)
tail(sort(leverage), 3)
#influencial points
influential <- dfbetas(bestmodel)
tail(sort(influential[,6]), 3)
```