Car Transmission Impact on Miles per Gallon: A Linear Regression Model

Francisco Ganhão

10/27/2016

Executive Summary

An analysis concerning the impact of a car's transmission type on Miles per Gallon (MPG) is presented.
Based on a exploratory analysis, **an automatic transmission has lower *MPG* compared to a manual**.
A statistical inference was also performed, **without considering other variables** and results also show that **automatic transmission outperforms manual transmissions.**
To better account other variations of the data, **a linear regression model of MPG was derived**, where an automatic transmission has a **maximum average of** 33.7 **MPG, while a manual can be** 1.8 **MPG higher**. The **automatic transmission coefficient is statistically significant, while the manual transmission coefficient is not.**
Results are therefore inconclusive due to lack of statistical significance.

Exploratory Analysis

Figure 1 (see appendix) overviews the Miles Per Gallon performance over several variables, **in general we do observe that an automatic transmission has better MPG than manual**. The performance is observed for the following variables: **Cylinders; Displacement; HP; Rear Axle Ratio; Weight; Quarter-Mile Time; Engine; Gears; Carburetors; and the transmission itself.**

Statistical Inference

Below you can find the results of the analyzed data for a **t-test that only considers the type of transmission** for the following hypotheses: H_0, where MPG is the same for both transmissions; H_A, where the MPG is lower for automatic transmissions. For a 95% interval **we can reject the null hypothesis that the consumption is the same for both transmissions**, therefore the data used for the model is **statistically significant**.

t	p-value	Lower bound	Upper Bound	Degrees of Freedom
-3.767123	0.0006868	-Inf	-3.913256	18.33225

Linear Regression Model Analysis - Nested Testing

Having studied the variables' plots, we can proceed onwards to a **linear MPG regression model that depends on the Transmission Type, as a confounder,** and at the same time **picking other variables** that make our model statistically significant. **A variance analysis/deviance of several linear methods, using *anova* and nested testing, was performed.** Variables that introduced a variance value above 5% **of significance ($P(t > F)$) were rejected**.

Model Variables	Degrees	F-value	P(t > F)	Decision
Transmission	30	16.8602788	0.0002850	Valid
Trans., Cylinder (Cyl.)	28	24.1577214	0.0000008	Valid
Trans., Cyl., HP	27	9.2140993	0.0052662	Valid
Trans., Cyl., HP, Weight (Wt.)	26	7.9490367	0.0090814	Valid
Trans.,Cyl., HP, Wt., Engine	25	1.2781784	0.2689680	Fail
Trans., Cyl., HP, Wt., Gears	24	0.1090855	0.8970957	Fail
Trans.,Cyl., HP, Wt., Carburetors	21	0.1627215	0.9734894	Fail

Model Variables	Degrees	F-value	P(t > F)	Decision
Trans., Cyl., HP, Wt., Axle Ratio	25	0.0364993	0.8500311	Fail
Trans., Cyl., HP, Wt., Displacement	25	0.1025183	0.7514890	Fail

Based on the previous results the selected model will therefore use **Transmission, Cylinders, Horse Power (HP) and Weight** as variables, where these explain $R^2 = \mathbf{0.8658799}$ of the variance, yielding the following formula: $\beta_0 + \beta_1 T_{MT} + \beta_2 Cyl_6 + \beta_3 Cyl_8 + \beta_4 HP + \beta_5 Wt$, where you can find an explanation for each coefficient in the first table below, as well as **the statistical significance and respective p-value** in the second table.

Coefficients	Values	σ	t-value	P(x>|t|)	Description
β_0	33.7083239	2.6048862	12.940421	0.0000000	Average MPG for Auto. Trans. and 4 cyl.
β_1	1.8092114	1.3963045	1.295714	0.2064597	MPG for Manual Trans.
β_2	-3.0313445	1.4072835	-2.154040	0.0406827	MPG for 6 Cyl.
β_3	-2.1636753	2.2842517	-0.947214	0.3522509	MPG for 8 Cyl.
β_4	-0.0321094	0.0136926	-2.345025	0.0269346	MPG per Horse Power
β_5	-2.4968294	0.8855878	-2.819404	0.0090814	MPG per 1000lb

St. Residual	Degrees	R^2	F-value	p-value (10^{-10})	Formula
2.41012	26	0.8658799	33.57121	1.505607	$\beta_0 + \beta_1 T_{MT} + \beta_2 Cyl_6 + \beta_3 Cyl_8 + \beta_4 HP + \beta_5 Wt$

In terms of **coefficient interpretation**: **a)** β_0 **shows that there is an average** 33.7 **MPG fuel consumption for automatic transmissions using 4 cylinders; b)** β_1 **shows that there is an average increase of** 1.8 **MPG when using manual transmission over the automatic; c)** β_2 **shows the average decrease of** -3.03 **MPG for 6 cylinders over 4; d)** β_3 **shows an average decrease of** -2.16 **MPG for 8 cylinders over 4 cylinders; e)** β_4 **shows a decrease of** -0.032 **MPG for each additional** HP**; f) finally,** β_5 **shows a decrease of** -2.49 **MPG for each additional 1000lb**.
Holding every variable constant, **using manual adds** 1.8 **MPG more in average than an automatic**; an **automatic bears a maximum average of 33.7 MPG**. The **automatic transmission constant,** β_0**, is statistically significant** as seen in the table; **the manual transmission constant** β_1 **is not - which makes our analysis inconclusive**.

Residual Analysis/Diagnostic

Figure 2 (see appendix) overviews several diagnosis tests regarding the studied model's residuals. *Residuals are close to a normal distribution*, plus they do not have a systematic pattern which validates the regression model. There are a few points that have a larger leverage, but their standardized residual is smaller than *Cook's distance*. Below you can find the top three cars for the following influence measures: absolute DFFIT value; absolute residual and the leverage - **these cars can be evaluated as potentital outliers**.

Car	|DFFIT|	Car	$\|e_i\|$	Car	Leverage
Chrysler Imperial	1.1759053	Toyota Corolla	5.051260	Maserati Bora	0.4713671
Toyota Corolla	0.9377561	Fiat 128	4.494712	Lincoln Continental	0.2936819
Toyota Corona	0.9093995	Datsun 710	3.938714	Toyota Corona	0.2777872

Conclusions

Fuel consumption is **better for automatic transmission than manual if we disregard variations from other variables**. The **linear regression model** shows that **the an automatic transmission, for an average of 33.7 MPG, is statistically significant**, but an **increase of an average 1.8 MPG for manual transmissions is not**, therefore **it is inconclusive whether a manual transmission is worse than an automatic.**

Appendix

The appendix section presents the code used to produce the document's figures as well as the table from the Residuals Analysis and Diagnostics section. Most of the code could not be presented here due to the assignment's constraints.

Libraries

The code below presents libraries for plotting purposes.

```
library(ggplot2, quietly = TRUE, warn.conflicts = FALSE)
library(grid, quietly = TRUE, warn.conflicts = FALSE)
library(gridExtra, quietly = TRUE, warn.conflicts = FALSE)
```

Exploratory Analysis - Plot Construction Code

The code below presents the plot construction for the exploratory analysis for each respective variable.

```
#Cylinder
gCylinder <- ggplot(aes(x = Cylinder, y=MPG, fill=Transmission),data=pframe) +
             geom_violin() +
             labs(x="Cylinders", y = "Miles per Gallon")
#Displacement
gDisplacement <- ggplot(aes(x = Displacement, y=MPG),data=pframe) +
                 geom_point(aes(color =Transmission, shape = Transmission), size = 2) +
                 labs(x="Displacement", y = "Miles per Gallon")
#Horse Power
gHorsePower <- ggplot(aes(x = HP, y=MPG),data=pframe) +
               geom_point(aes(color =Transmission, shape = Transmission), size = 2) +
               labs(x="Horse Power", y = "Miles per Gallon")
#Rear Axle Ratio
gAxleRatio <- ggplot(aes(x = RAR, y=MPG),data=pframe) +
              geom_point(aes(color =Transmission, shape = Transmission), size = 2) +
              labs(x="Rear Axle Ratio", y = "Miles per Gallon")
#Weight
gWeight <- ggplot(aes(x = Weight, y=MPG),data=pframe) +
           geom_point(aes(color =Transmission, shape = Transmission), size = 2) +
           labs(x="Weight (1000 lbs)", y = "Miles per Gallon")
#Quarter-Mile Time
gTime <- ggplot(aes(x = Time, y=MPG),data=pframe) +
         geom_point(aes(color =Transmission, shape = Transmission), size = 2) +
         labs(x="Quarter Mile Time (sec)", y = "Miles per Gallon")
#Engine
gEngine <- ggplot(aes(x = Engine, y=MPG, fill=Transmission),data=pframe) +
           geom_violin() +
           labs(x="Engine", y = "Miles per Gallon")
#Gears
gGears <- ggplot(aes(x = Gears, y=MPG, fill=Transmission),data=pframe) +
          geom_violin() +
          labs(x="Gears", y = "Miles per Gallon")
#Carburetors
gCarburetors <- ggplot(aes(x = Carburetors, y=MPG, fill=Transmission),data=pframe) +
                geom_point(aes(color =Transmission, shape = Transmission), size = 2) +
                labs(x="Carburetors", y = "Miles per Gallon")
#Transmission
gTransmission <- ggplot(aes(x = Transmission, y=MPG, fill=Transmission),data=pframe) +
                 geom_violin() +
                 labs(x="Transmission", y = "Miles per Gallon")
```

The final grid that arranges the exploratory analysis figure can be found below.

```
grid.arrange(gCylinder, gDisplacement, gHorsePower, gAxleRatio, gWeight, gTime,
             gEngine, gGears, gCarburetors, gTransmission, ncol=2)
```

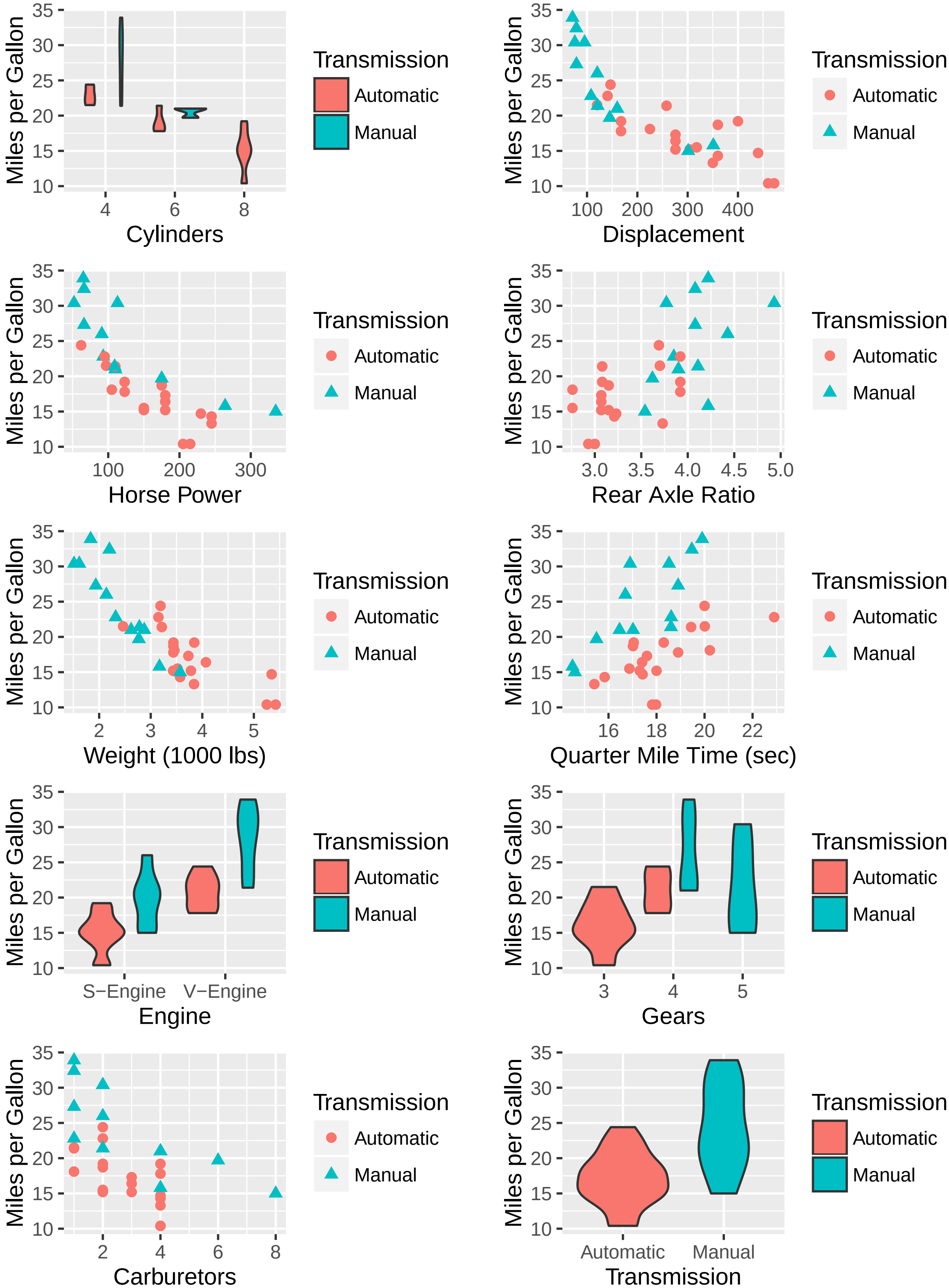

Figure 1: Miles Per Gallon Performance over Several Variables.

Residual Analysis/Diagnostic - Plot and Table Code

Below you can find the plot code for the residual analysis and respective diagnostic.

```
par(mfrow=c(2,2))
#Plots the residual analysis diagnostics for the model lmTxCylHpWt
#Transmission + Cyl + HP + Wt
plot(lmTxCylHpWt,which=1)#Residuals vs Fitted
plot(lmTxCylHpWt,which=2)#Normal Q-Q
plot(lmTxCylHpWt,which=3)#Scale-Location
plot(lmTxCylHpWt,which=5)#Residuals vs Leverage
```

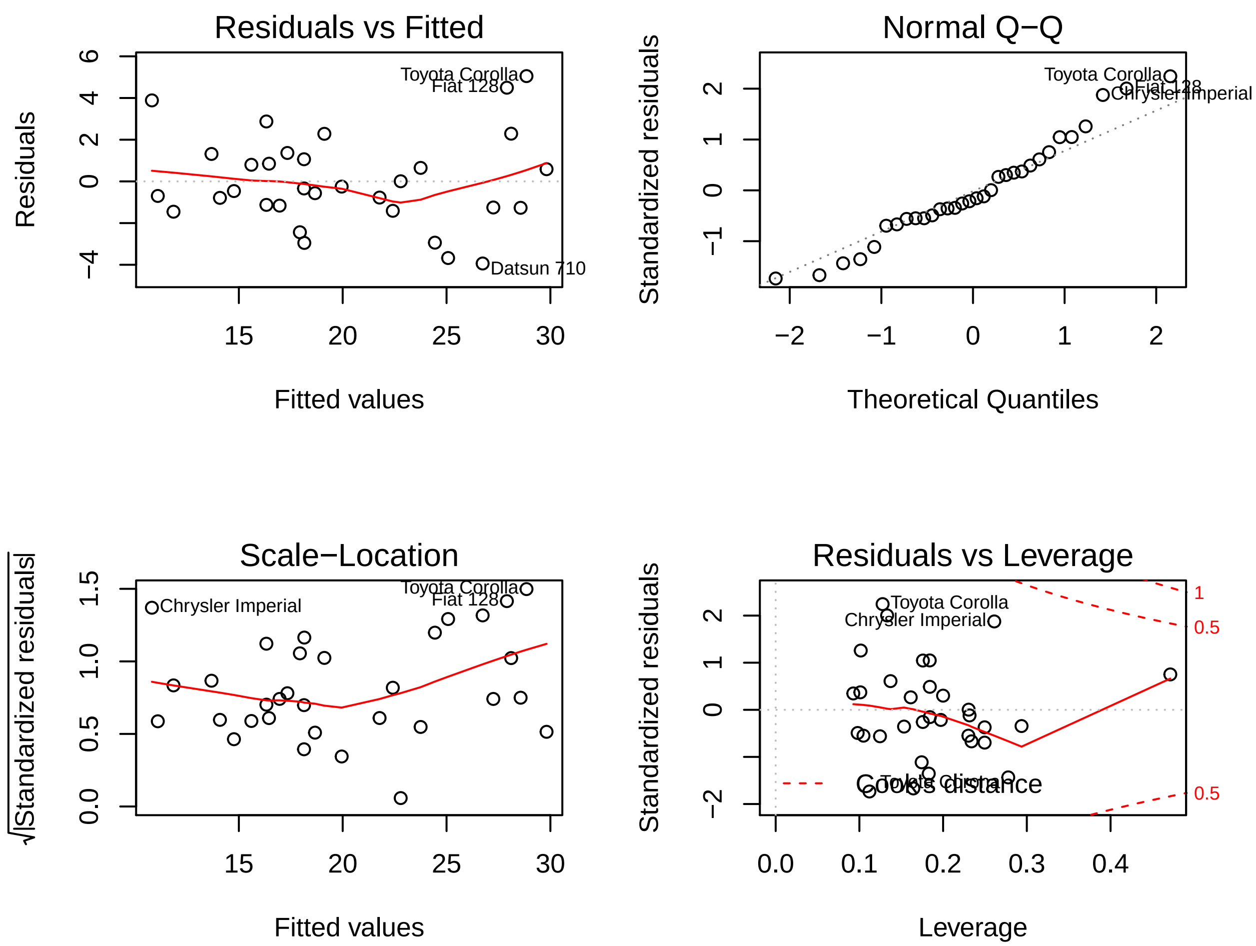

Figure 2: Residual Analysis Plots.

Below you can find the code to output the table concerning the Residual Analysis/Diagnostic section.

```
dffit <- sort(abs(dffits(lmTxCylHpWt)), decreasing = TRUE)[1:3]
carfit <- names(dffit)
residuals <- sort(abs(resid(lmTxCylHpWt)), decreasing = TRUE)[1:3]
carresid <- names(residuals)
hats <- sort(hatvalues(lmTxCylHpWt), decreasing = TRUE)[1:3]
carhat <- names(hats)
columns <- c("Car","|DFFIT|","Car","$|e_i|$","Car","Leverage")
rtable <- data.frame(carfit,as.vector(dffit),carresid,as.vector(residuals),carhat,as.vector(hats))
names(rtable) <- columns
kable(rtable)
```