Exploring the Relationship between Car Characteristics and Fuel Efficiency

Yohance Nicholas

3/15/2020

Executive Summary

This report was prepared by Yohance Nicholas in partial fulfilment of the Regression Models Course which comprises one of the five courses necessary for the Data Science: Statistics and Machine Specialization offered by Johns Hopikins University through Coursera. In this report, candidates wear the hat of a *Motor Trend Researcher* who seeks to explore the relationship between vehicle characteristics and fuel efficiency with the assistance of the `mtcars` data set.

This data, which was extracted from the 1974 *Motor Trend* US magazine, comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). Exploratory data analyses and regression models are utilised to mainly explore the extent to which **automatic** (am = 0) and **manual** (am = 1) transmissions features impact fuel efficiency, as measured by the the Miles/(US) gallon variable - **MPG** .

The research begins with data preparation and exploratory data analysis. Subsequently, several candidate linear regression models are estimated in order to identify the model with the best fit - i.e. the highest Adjusted R-squared value.

Research Questions

They are particularly interested in the following two questions:

1. "Is an automatic or manual transmission better for MPG?"
2. "Quantify the MPG difference between automatic and manual transmissions"

Description of the Dataset

The `mtcars` data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). The data frame contains 32 observations on 11 (numeric) variables; namely:

- **mpg** Miles/(US) gallon
- **cyl** Number of cylinders
- **disp** Displacement (cu.in.)
- **hp** Gross horsepower
- **drat** Rear axle ratio
- **wt** Weight (1000 lbs)
- **qsec** 1/4 mile time

- **vs** Engine (0 = V-shaped, 1 = straight)
- **am** Transmission (0 = automatic, 1 = manual)
- **gear** Number of forward gears
- **carb** Number of carburetors

Load Data

The data set is loaded and the necessary data transformations are done by factoring the necessary variables .

```
data("mtcars")
```

The following tables and figures provide a description of the data set

```
library(tidyverse)
library(psych)
library(knitr)
library(kableExtra)
library(broom)
landscape(kable(describe(mtcars), caption="Descriptive Statistics of Motor Trend Data Set"))
```

Table 1: Descriptive Statistics of Motor Trend Data Set

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
mpg	1	32	20.090625	6.0269481	19.200	19.6961538	5.4114900	10.400	33.900	23.500	0.6106550	-0.3727660	1.0654240
cyl	2	32	6.187500	1.7859216	6.000	6.2307692	2.9652000	4.000	8.000	4.000	-0.1746119	-1.7621198	0.3157093
disp	3	32	230.721875	123.9386938	196.300	222.5230769	140.4763500	71.100	472.000	400.900	0.3816570	-1.2072119	21.9094727
hp	4	32	146.687500	68.5628685	123.000	141.1923077	77.0952000	52.000	335.000	283.000	0.7260237	-0.1355511	12.1203173
drat	5	32	3.596563	0.5346787	3.695	3.5792308	0.7042350	2.760	4.930	2.170	0.2659039	-0.7147006	0.0945187
wt	6	32	3.217250	0.9784574	3.325	3.1526923	0.7672455	1.513	5.424	3.911	0.4231465	-0.0227108	0.1729685
qsec	7	32	17.848750	1.7869432	17.710	17.8276923	1.4158830	14.500	22.900	8.400	0.3690453	0.3351142	0.3158899
vs	8	32	0.437500	0.5040161	0.000	0.4230769	0.0000000	0.000	1.000	1.000	0.2402577	-2.0019376	0.0890983
am	9	32	0.406250	0.4989909	0.000	0.3846154	0.0000000	0.000	1.000	1.000	0.3640159	-1.9247414	0.0882100
gear	10	32	3.687500	0.7378041	4.000	3.6153846	1.4826000	3.000	5.000	2.000	0.5288545	-1.0697507	0.1304266
carb	11	32	2.812500	1.6152000	2.000	2.6538462	1.4826000	1.000	8.000	7.000	1.0508738	1.2570431	0.2855297

Exploratory Data Analysis

We begin the analysis by performing some initial explaratory data analysis to get a better idea of the existing patterns between variables in the data set. Normally in regression analysis scatter plot is a very effective tool. Below we create a nice pairwise scatter plots which offer a convenien way to investigate the relationship between all the variables in this data set.

```
library(ggplot2)
gg_boxplot <- ggplot(mtcars,
                     aes(x=factor(am),
                         y=mpg)) +
  geom_boxplot() +
  labs(title="Figure 2: Box-plot Comparison of Auto Mileage by Transmission Type",
       y="Miles/(US) gallon",
       x="Transmission (0 = automatic, 1 = manual)")
gg_boxplot
```

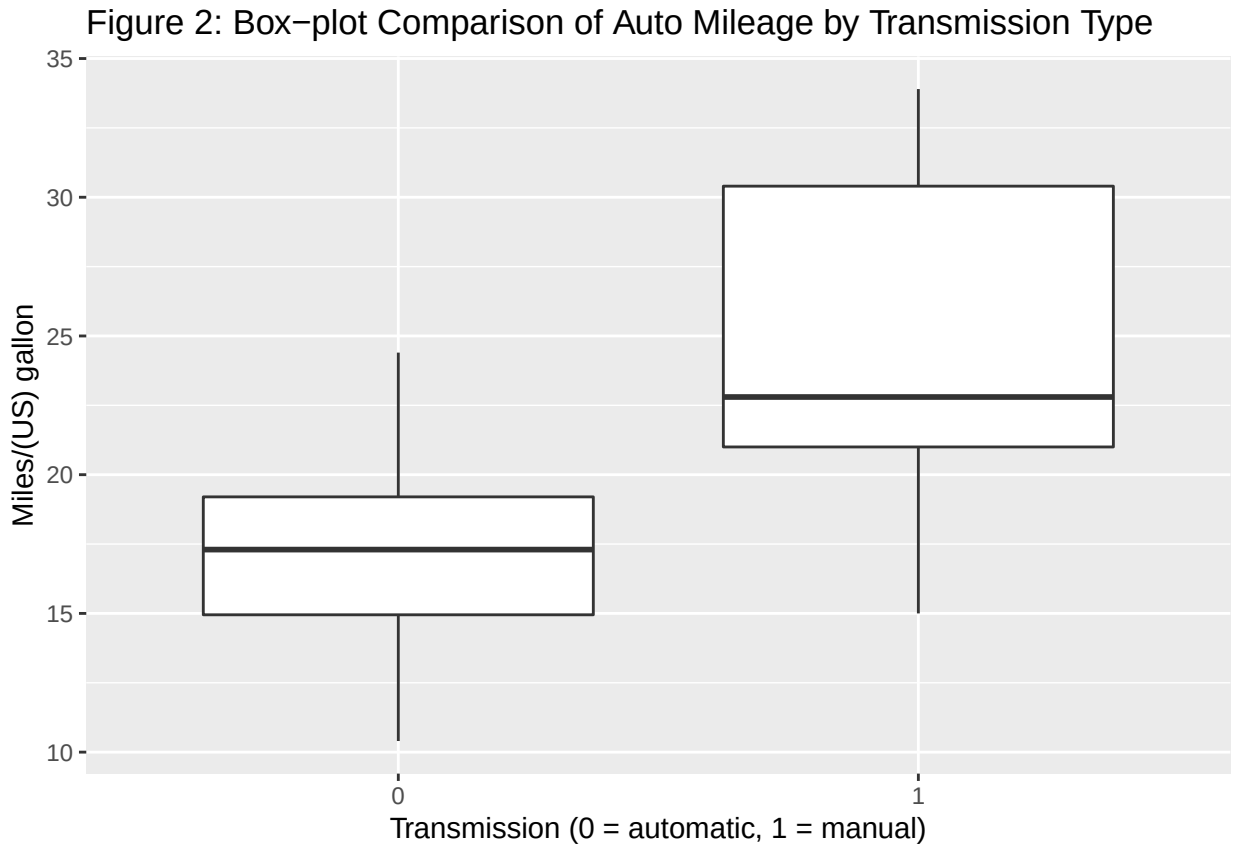


Table 2: A simple linear 'mpg' model

term	estimate	std.error	statistic	p.value
(Intercept)	17.147368	1.124602	15.247492	0.000000
am1	7.244939	1.764422	4.106127	0.000285

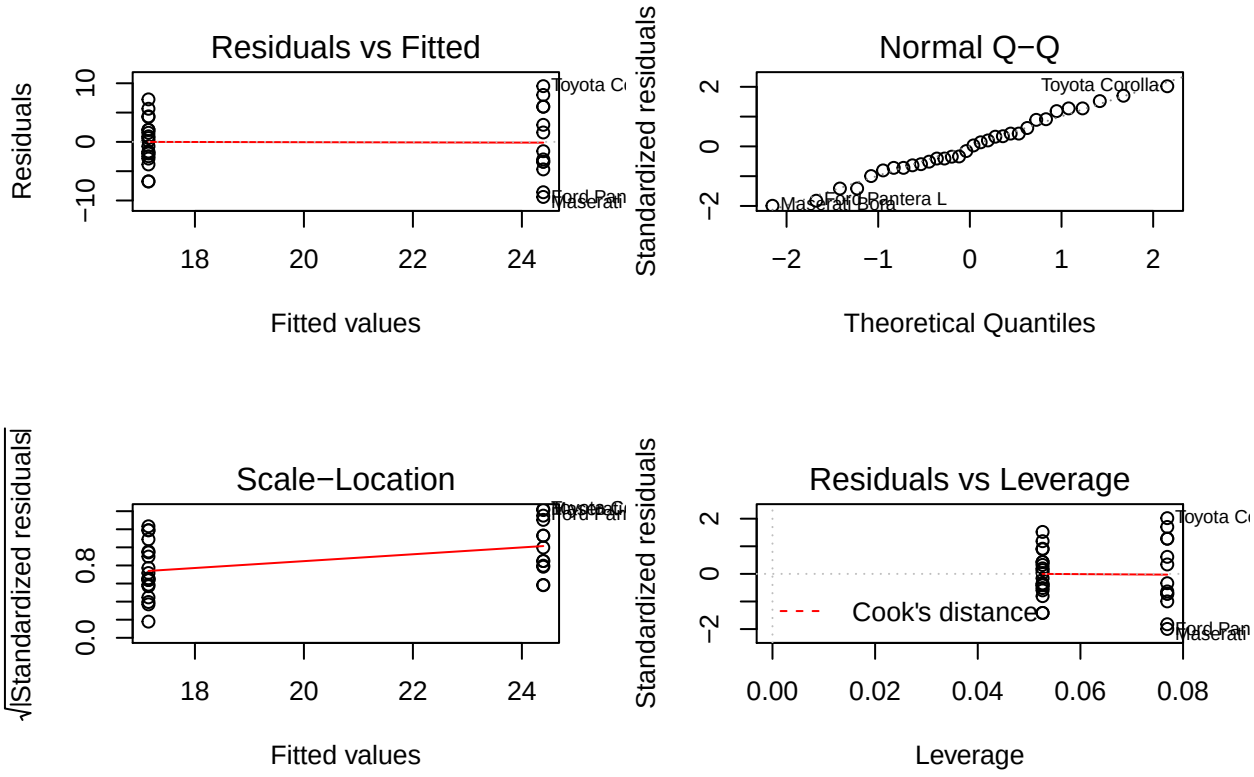
Empirical Results and Significant Findings

Estimation

Model 1: Univariate Regression Model

```
univariate_model <- lm(mpg ~ am, mtcars)
kable(tidy(univariate_model), caption="A simple linear 'mpg' model")
```

```
par(mfrow = c(2, 2))
plot(univariate_model)
```



Model 2: Multivariate Regression Model with All Available Variables

Table 3: A multiple regression model of the relationship between all available variables and 'mpg'

term	estimate	std.error	statistic	p.value
(Intercept)	23.8791324	20.0658203	1.1900402	0.2525255
cyl6	-2.6486953	3.0408904	-0.8710262	0.3974664
cyl8	-0.3361630	7.1595395	-0.0469532	0.9631700
disp	0.0355463	0.0318992	1.1143329	0.2826734
hp	-0.0705068	0.0394256	-1.7883534	0.0939316
drat	1.1828302	2.4834846	0.4762784	0.6407392
wt	-4.5297758	2.5387458	-1.7842573	0.0946186
qsec	0.3678448	0.9353957	0.3932505	0.6996672
vs1	1.9308505	2.8712578	0.6724755	0.5115079
am1	1.2121157	3.2135451	0.3771896	0.7113157
gear4	1.1143549	3.7995173	0.2932886	0.7733203
gear5	2.5283960	3.7363580	0.6767007	0.5088975
carb2	-0.9793543	2.3179745	-0.4225044	0.6786509
carb3	2.9996387	4.2935461	0.6986390	0.4954678
carb4	1.0914229	4.4496199	0.2452845	0.8095603
carb6	4.4775692	6.3840624	0.7013668	0.4938127
carb8	7.2504113	8.3605664	0.8672153	0.3994849

```
multivariate_model <- lm(mpg ~ ., mtcars)
kable(tidy(multivariate_model), caption= "A multiple regression model of the relationship between all a
```

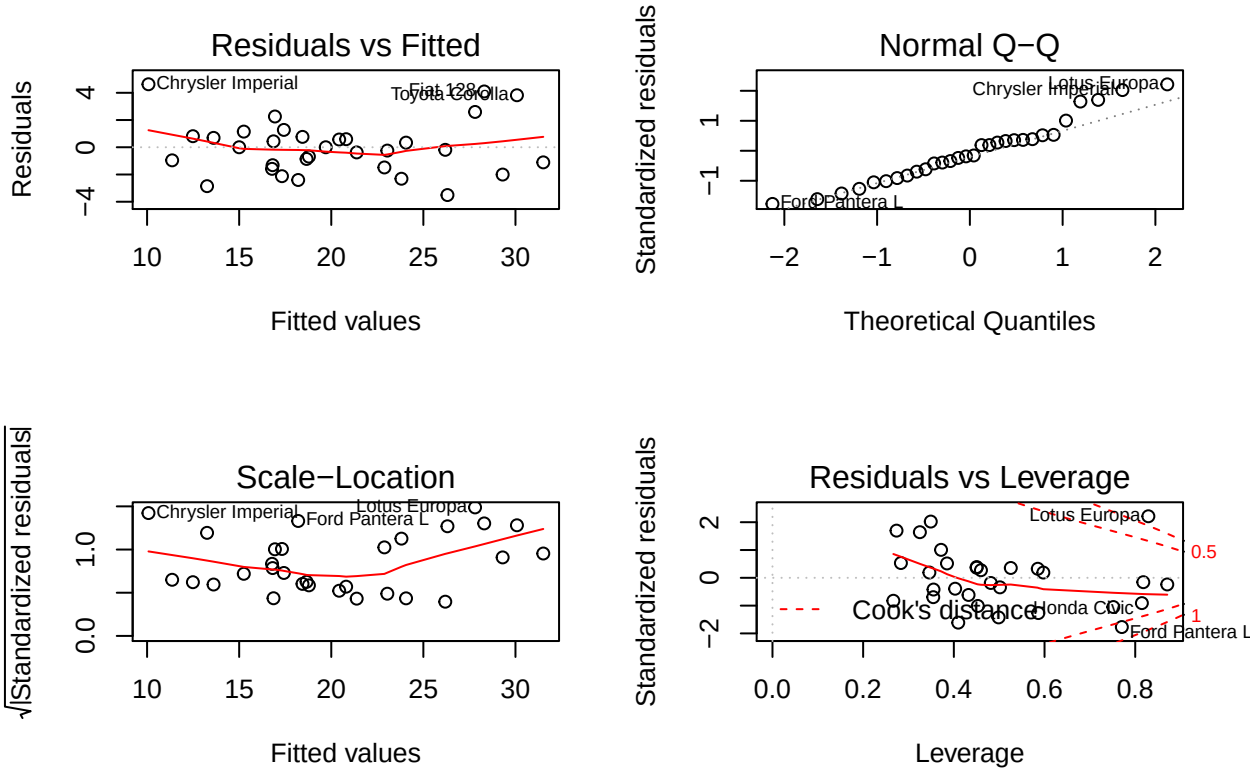
```
par(mfrow=c(2,2))
plot(multivariate_model)
```

```
## Warning: not plotting observations with leverage one:
##   30, 31
```

```
## Warning: not plotting observations with leverage one:
##   30, 31
```

Table 4: A stepwise regression model of the relationship between all available variables and 'mpg'

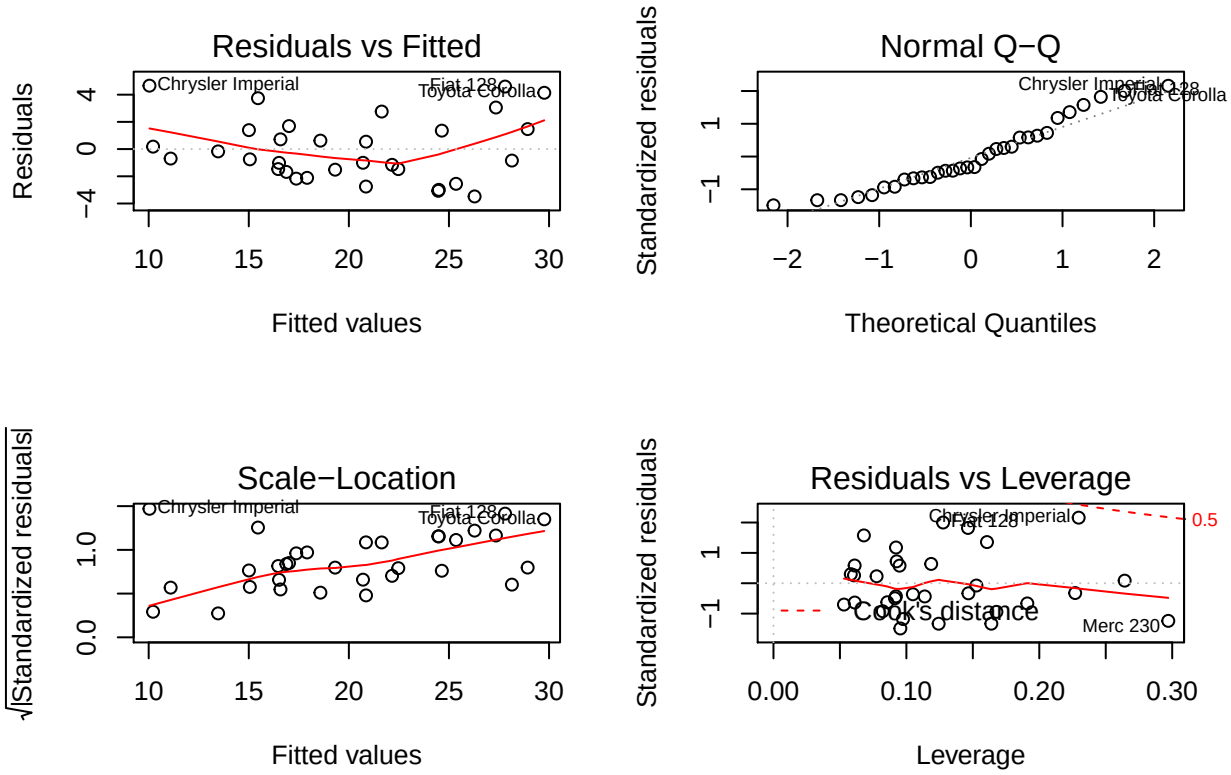
term	estimate	std.error	statistic	p.value
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
wt	-3.916504	0.7112016	-5.506882	0.0000070
qsec	1.225886	0.2886696	4.246676	0.0002162
am1	2.935837	1.4109045	2.080819	0.0467155



Model 3: Stepwise Regression Model

```
kable(tidy(stepwise_model), caption= "A stepwise regression model of the relationship between all availa
```

```
par(mfrow=c(2,2))
plot(stepwise_model)
```



Post-Estimation Diagnostics

From the above plots, one can make the following observations:

- The points in the Residuals vs. Fitted plot seem to be randomly scattered on the plot and verify the independence condition.
- The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed.
- The Scale-Location plot consists of points scattered in a constant band pattern, indicating constant variance.
- There are some distinct points of interest (outliers or leverage points) in the top right of the plots.
- We now compute some regression diagnostics of our model to find out these interesting leverage points as shown in the following section. We compute top three points in each case of influence measures.

Analysis of Variance (ANOVA): Univariate vs Multivariate vs Stepwise Model

One can employ the Analysis of Variance to compare fitted regression models.

```
anova_mtcars <- anova(univariate_model,multivariate_model, stepwise_model)
kable(tidy(anova_mtcars))
```

```
## Warning: Unknown or uninitialised column: 'term'.
```

res.df	rss	df	sumsq	statistic	p.value
30	720.8966	NA	NA	NA	NA
15	120.4027	15	600.49393	4.9873804	0.0017585
28	169.2859	-13	-48.88326	0.4684594	0.9114130

As p-value is significant, hence the null hypothesis that the variables cyl, hp and wt do not contribute to the accuracy of the model is rejected.

Analysis of Variance (ANOVA): All other candidate models

Conclusions

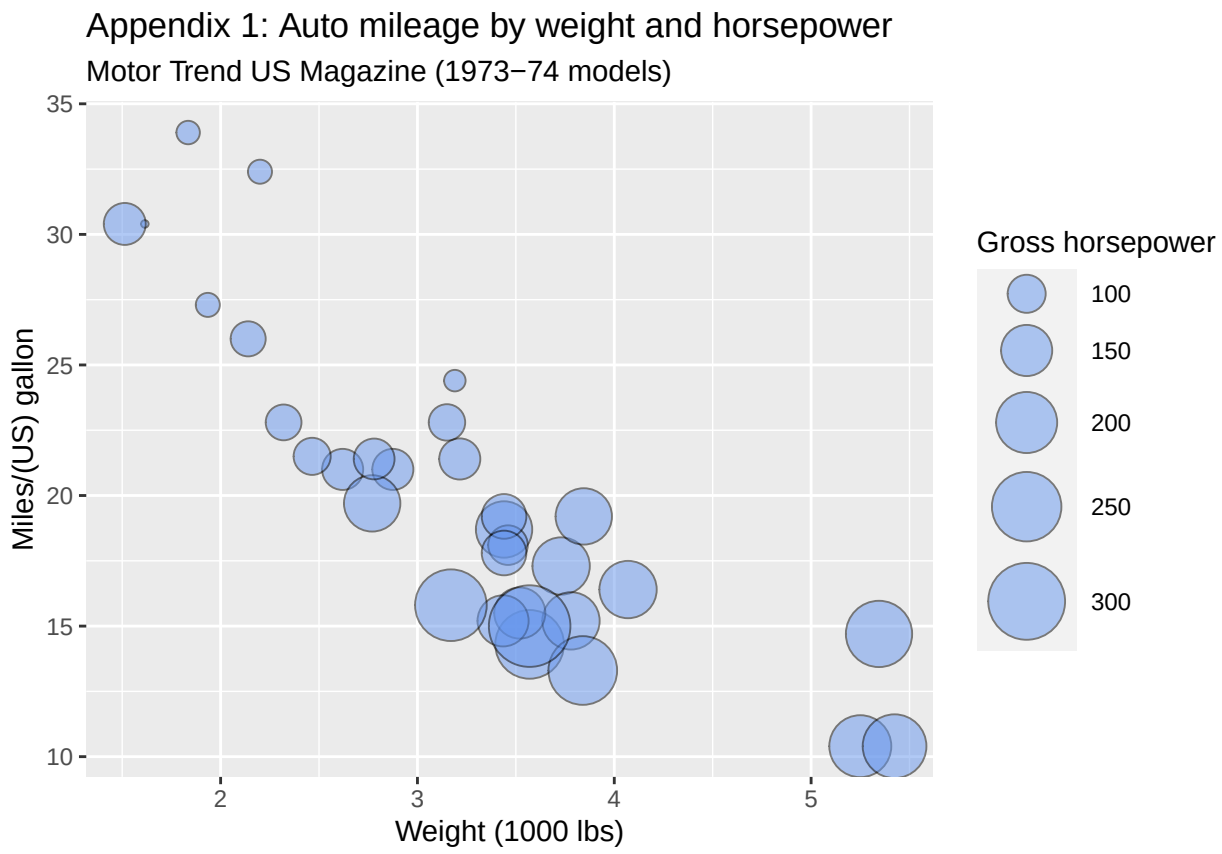
Based on the observations from our best fit model - the stepwise regression model - the following conclusions can be made:

- Cars with Manual transmission get more miles per gallon compared aganist cars with Automatic transmission. (1.8 adjusted by *hp, cyl, and wt*). *mpg* will decrease by 2.5 (adjusted by *hp, cyl, and am*) for every 1000 lb increase in wt.
- *mpg* decreases negligibly with increase of *hp.*
- If number of cylinders, *cyl* increases from 4 to 6 and 8, *mpg* will decrease by a factor of 3 and 2.2 respectively (adjusted by *hp*, *wt*, and *am*).

Appendices

Appendix 1: Auto mileage by weight and horsepower

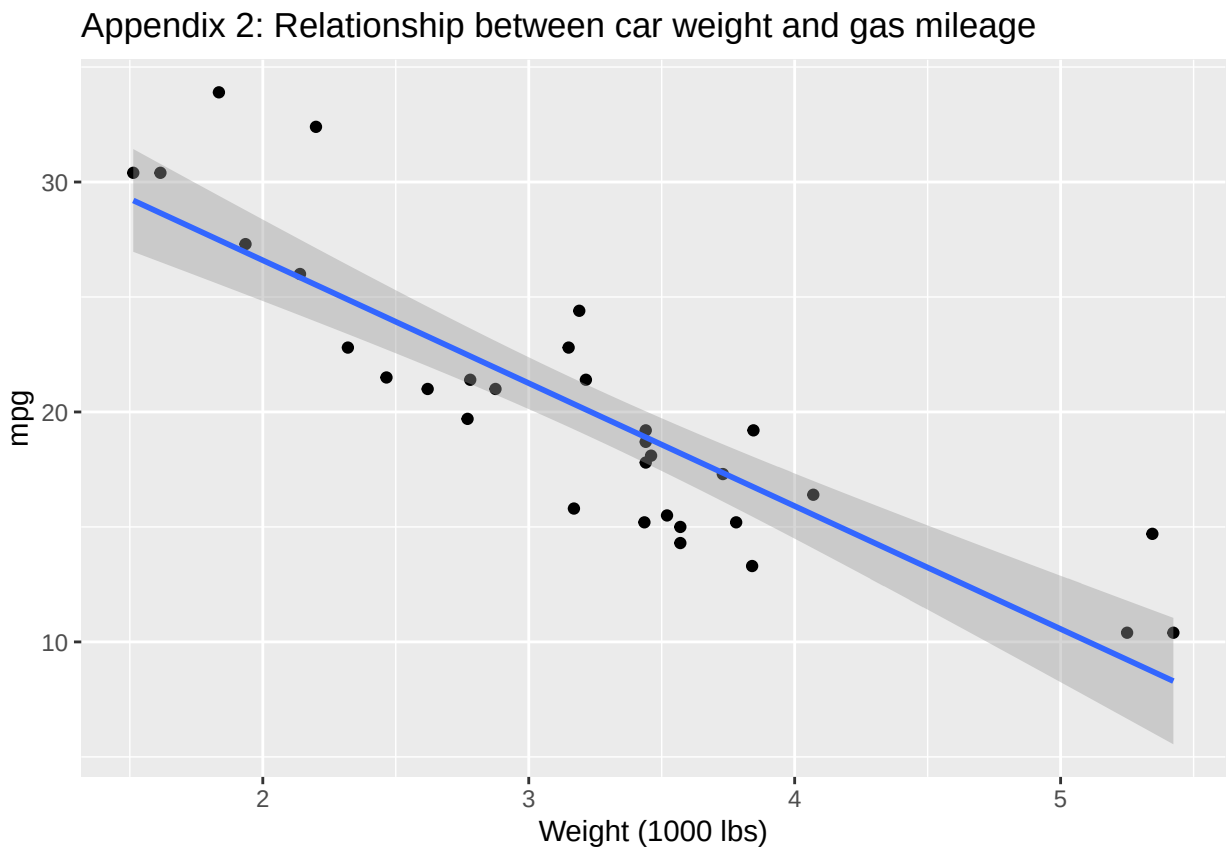
```
ggplot(mtcars,
       aes(x = wt,
           y = mpg,
           size = hp)) +
geom_point(alpha = .5,
           fill="cornflowerblue",
           color="black",
           shape=21) +
scale_size_continuous(range = c(1, 14)) +
  labs(title = "Appendix 1: Auto mileage by weight and horsepower",
       subtitle = "Motor Trend US Magazine (1973-74 models)",
       x = "Weight (1000 lbs)",
       y = "Miles/(US) gallon",
       size = "Gross horsepower")
```



Appendix 2: Relationship between car weight and gas mileage

```
mtcars %>%
  ggplot(aes(wt, mpg)) +
  geom_point() +
  stat_smooth(method = "lm") +
  ggtitle("Appendix 2: Relationship between car weight and gas mileage") +
  xlab("Weight (1000 lbs)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Appendix 3: Scatter Plot of MPG vs. Weight by Transmission

```
ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am, height=3, width=3)) + geom_point() +
scale_colour_discrete(labels=c("Automatic", "Manual")) +
xlab("weight") + ggtitle("Appendix 3: Scatter Plot of MPG vs. Weight by Transmission")
```

Appendix 3: Scatter Plot of MPG vs. Weight by Transmission