Course Project: Regression Models Analysis

Sabrina Kate S. Carranza

June 28, 2019

Executive summary

This report examines the relationship between a set of variables and fuel consumption measured in miles per gallon (MPG). Looking only at the transmission type, our findings suggest that cars with an automatic transmission consume more fuel than cars with a manual transmission. However, this effect can no longer be observed when including other variables in our final regression model. Based on our findings fuel consumption can be best expressed as linear combination of *Weight*, $Weight^2$ and *Quarter mile time*.

Examing the Effect of Transmission Type on Fuel Consumption

An exploratory data analysis reveals a difference in fuel consumption between cars grouped by transmission type (See Appendix). In this section we will examine if this observed difference is also statistically significant.

We assume that the data was randomly sampled from the *1974 Motor Trend* magazine. Moreover, we deal with non-paired data which means that the independence condition is also satisfied between groups. Since both samples have n < 30 observations and are not strongly skewed we can apply a two sample t-test ($H_0 : \mu_{automatic} = \mu_{manual}$. $H_A : \mu_{automatic} \neq \mu_{manual}$):

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method
-7.244939	17.14737	24.39231	-3.767123	0.0013736	18.33225	-11.28019	-3.209684	Welch Two Sample t-test

Because the p-value 0.0014 < 0.05, we reject the null hypothesis. The data provides strong evidence that fuel consumption indeed differs between transmission types. We are 95% confident that cars from the 1973/74 population with an automatic transmission drive between 3.21 and 11.28 less miles per gallon than cars with a manual transmission.

Examing the Effect of Multiple Variables on Fuel Consumption

In this section we will determine whether the difference in fuel consumption by transmission type holds, if we take additional variables and their effect on fuel consumption into account.

The first step is to perform a linear regression of *MPG* on all other variables in the *mtcars* data set leveraging the *step()* function: This ensures that we will perform a variable selection leaving only the most important variables in the model. This model we refer to as **Model 1** includes 3 significant variables:

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
am	2.935837	1.4109045	2.080819	0.0467155

When performing model diagnostics we observe a non-linear pattern in the residual plot for Model 1 (Left side of Figure 2). This is a problem because all of the conclusions that we draw from the fit are suspect. Our findings suggest that it is best to add $Weight^2$ to Model 1 to accommodate this non-linear relationship. We call this new model **Model 2**. When adding $Weight^2$ only a slight pattern can be observed in the residuals (Right side of Figure 1).

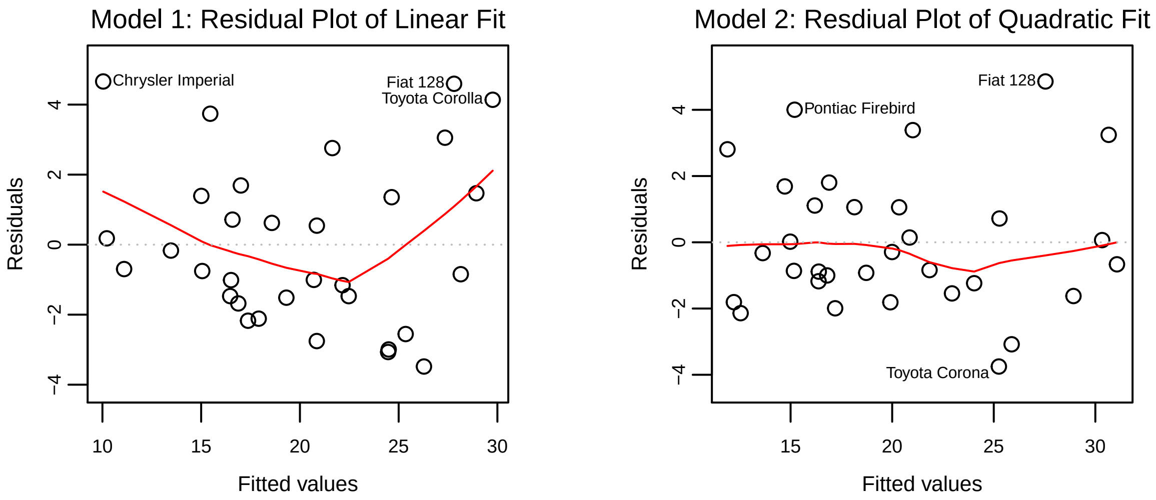

Figure 1: Plots of residuals versus fitted values for the *mtcars* data set. **Left**: A linear regression of *MPG* on *Weight*, *Quarter mile time* and *Transmission Type*. A pattern in the residuals indicates non-linearity in the data. **Right**: A linear regression of *MPG* on the same variables plus $Weight^2$. The former pattern nearly vanished.

Performing a hypothesis test comparing the two models with the `anova()` function reveals the following:

Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Model 1	28	169.2859	NA	NA	NA	NA
Model 2	27	130.8573	1	38.42859	7.929032	0.0089778

Here the F-statistic is 7.93 and the associated p-value is 0.009. This provides evidence that the model containing the predictors $Weight$ and $Weight^2$ is superior to the model that only contains the predictor $Weight$.

Model 2 is superior to Model 1 in terms of R^2 (0.8666 vs. 0.8336) and RSE (2.2015 vs. 2.4588). However, examining the individual p-values from the predictors of Model 2 reveals that *Transmission Type (AM)* is no longer significant:

term	estimate	std.error	statistic	p.value
(Intercept)	27.7376147	8.9574469	3.0965983	0.0045282
wt	-11.2490694	2.6807530	-4.1962349	0.0002629
I(wt^2)	0.9581950	0.3402858	2.8158538	0.0089778
qsec	0.9705112	0.2739061	3.5432263	0.0014614
am	1.0215185	1.4345500	0.7120828	0.4825217

This suggests that we might drop *Transmission Type (AM)* from the quadratic model. Dropping this predictor results in **Model 3** with the respective summary information below:

term	estimate	std.error	statistic	p.value
(Intercept)	32.6418325	5.6767588	5.750083	0.0000036
wt	-12.4330965	2.0841792	-5.965464	0.0000020
I(wt^2)	1.0730270	0.2969987	3.612901	0.0011739
qsec	0.8598587	0.2235665	3.846099	0.0006339

Now, all included variables are highly significant again. Like Model 2, Model 3 does not show any pattern in the residual plot (See Appendix).

Comparing all three models we can clearly see the superiority of Model 3 in terms of *Adjusted* R^2 and *RSE*:

```
## Warning: `data_frame()` is deprecated, use `tibble()`.
## This warning is displayed once per session.
```

Model	Adj.r.squared	RSE
Model 1	0.8335561	2.458846
Model 2	0.8665743	2.201492
Model 3	0.8689233	2.182028

Model 3 would be our final choice when modeling the relationship of specific car variables and and fuel consumption. The 95% confidence intervals are as follows: (-16.702, -8.164) for *Weight*, (0.465, 1.681) for $Weight^2$, and (0.402, 1.318) for *Quarter mile time.* Moreover, Model 3 corresponds to the model derived by the *step()* function when adding $Weight^2$ as an additional variable to the original *mtcars* dataset (See Appendix).

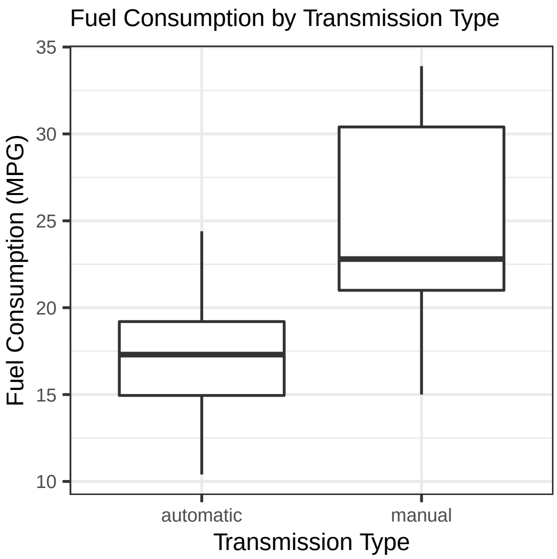

Figure 2: Boxplot of fuel consumption grouped by transmission type. A clear difference between the groups can be observed: Cars witha a manual transmission seem to have a lower fuel consumption than cars with an automatic transmission: On average they can drive more miles per gallon

Appendix

General Terms

The fuel economy of an automobile is the fuel efficiency relationship between the distance traveled and the amount of fuel consumed. Consumption can be expressed in 2 ways:

- Volume of fuel to travel a fixed distance (Gallons per Mile)
- Distance traveled per fixed fuel unit (Miles per Gallon).

In this report we will deal with fuel consumption measured in **Miles per gallon (MPG)**. At least for European readers it is important to note that MGP is interpreted the opposite way compared to Gallons per Mile:

- A car with a *higher MPG* metric *consumes less* fuel *than* a car with a *lower MPG*

Exploratory Data Analysis

- Number of observations and variables: 32 / 11
- Number of NAs: 0
- Number of cars by transmission type:

Transmission.Type	Freq
automatic	19
manual	13

Additional Plots

Additional Regression Models

Results of a regression of MPG on all variables plus $Weight^2$ with a subsequent automatic variable selection:

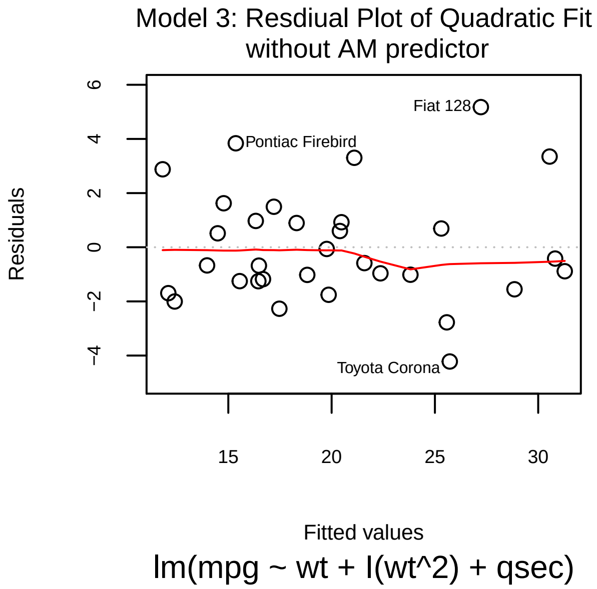

Figure 3: No pattern can be observed in the residuals

```r
lm(mpg ~ . + I(wt^2), data = mtcars) %>% step(trace = FALSE) %>% tidy %>% kable
```

term	estimate	std.error	statistic	p.value
(Intercept)	32.6418325	5.6767588	5.750083	0.0000036
wt	-12.4330965	2.0841792	-5.965464	0.0000020
qsec	0.8598587	0.2235665	3.846099	0.0006339
I(wt^2)	1.0730270	0.2969987	3.612901	0.0011739