Regression Models Course - Project
Relationship between a set of variables and miles per gallon (MPG)

Omer Shechter

Executive Summary

This report is done for Motor Trend, a magazine about the automobile industry. This report aims to explore the relationship between a set of variables of different type of cars and the MPG value (Mile per gallon) as an outcome. The report particularly interested in the following two questions: * Is an automatic or manual transmission better for MPG? * Quantify the MPG difference between automatic and manual transmissions? The report provides some exploratory data analysis suggest several linear regression models for predicting the outcome(MPG), compare them and analyze the residual of the selected model.

Data set

Load required libraries.

```
library(ggplot2)
library(GGally)
library(datasets)
library(kableExtra)
library(broom)
```

The 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). A data frame with 32 observations on 11 variables. Convert the categorical variables into factors (the non-continuous variables).

```
data(mtcars)
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$vs   <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am   <- factor(mtcars$am,labels=c("Automatic","Manual"))
```

Analysis

Exploratory analysis

Test the following null hypothesis: H0 –> the mean of Mile per gallon (MPG) is the same for manual and automatic transmission or what we check is the difference in means is zero.

```
kable(tidy(t.test(mtcars[mtcars$am=="Automatic",]$mpg,mtcars[mtcars$am=="Manual",]$mpg))) %>%
  kable_styling(bootstrap_options = "striped",font_size = 9 ,
                latex_options="scale_down", position = "left")
```

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

As it can be seen the result is significant p-value = 0.001374 So the Null Hypothesis need to be rejected - Hence there is a difference in the MPG data between the automatic and manual transmission types. See also figure 1. in the appendix. From figure 1. we also see that manual transmission provide larger values of mpg.

Next using a plot of the pairs (Figure 2 ggpairs) We can see : A lot of pairs exhibit high correlation Examples: disp and cyl (positive) drat and disp (negative) mpg correlated with almost all of the others coefficients.

Regression

model fit

The approach is to create a model with all parameters and mpg as an outcome. Then we perform stepwise model selection to select significant predictors for the final, selected model.

```
fitmodel <- lm(mpg ~ ., data = mtcars)
bestmodel <- step(fitmodel, direction = "both",trace=0)
tidy(summary(bestmodel))
```

```
## # A tibble: 6 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  33.7       2.60      12.9   7.73e-13
## 2 cyl6         -3.03      1.41      -2.15  4.07e- 2
## 3 cyl8         -2.16      2.28      -0.947 3.52e- 1
## 4 hp           -0.0321    0.0137    -2.35  2.69e- 2
## 5 wt           -2.50      0.886     -2.82  9.08e- 3
## 6 amManual      1.81      1.40       1.30  2.06e- 1
```

```
summary(bestmodel)$r.squared
```

```
## [1] 0.8658799
```

As it can be seen the step function select the model: mpg ~ cyl + hp + wt + am

Analysis of Variances,

Compare the model with all parameters, to the one that was chosen by the step function.

```
anova(fitmodel,bestmodel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS  Df Sum of Sq      F Pr(>F)
## 1     15 120.40
## 2     26 151.03 -11   -30.623 0.3468 0.9588
```

Looking at the Anova results, we see that removing the other coefficients as predictors have not significantly affected the model.

To verify the i.i.d Gaussian residual is not violated we look at the residual plot (figure 3.) we can see that :

Residuals vs Fitted

The dots are equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships.

Normal Q-Q

It looks like residuals are normally distributed, there is no severely deviate (Though there are some deviated points at the upper and lower edges)

Scale-Location

Residuals appear randomly spread

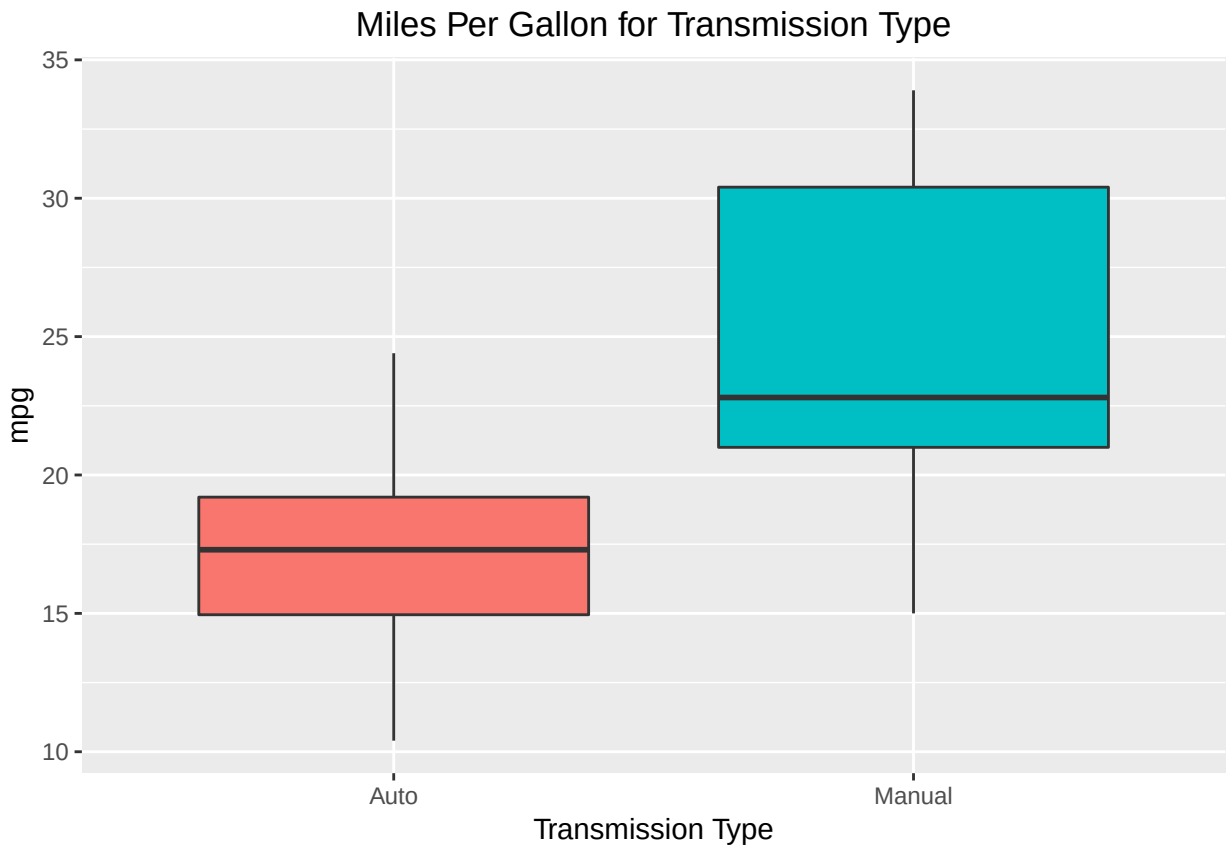
Residuals vs Leverage

Typical look when there is no influential case, as all cases are well inside of the Cook's distance lines.

Conclusions

The model explains 84.96% of the variance, so it looks like cyl, hp and wt did affect the correlation between am and mpg. Also, we see that there is a difference in the MPG between manual and automatic transmission we can see from the chosen model that manual cars have on average 1.8 miles per gallon than the automatic cars. # Appendix Figure 1. Plot the Miles per Gallon values according to manual and automatic transmissions.

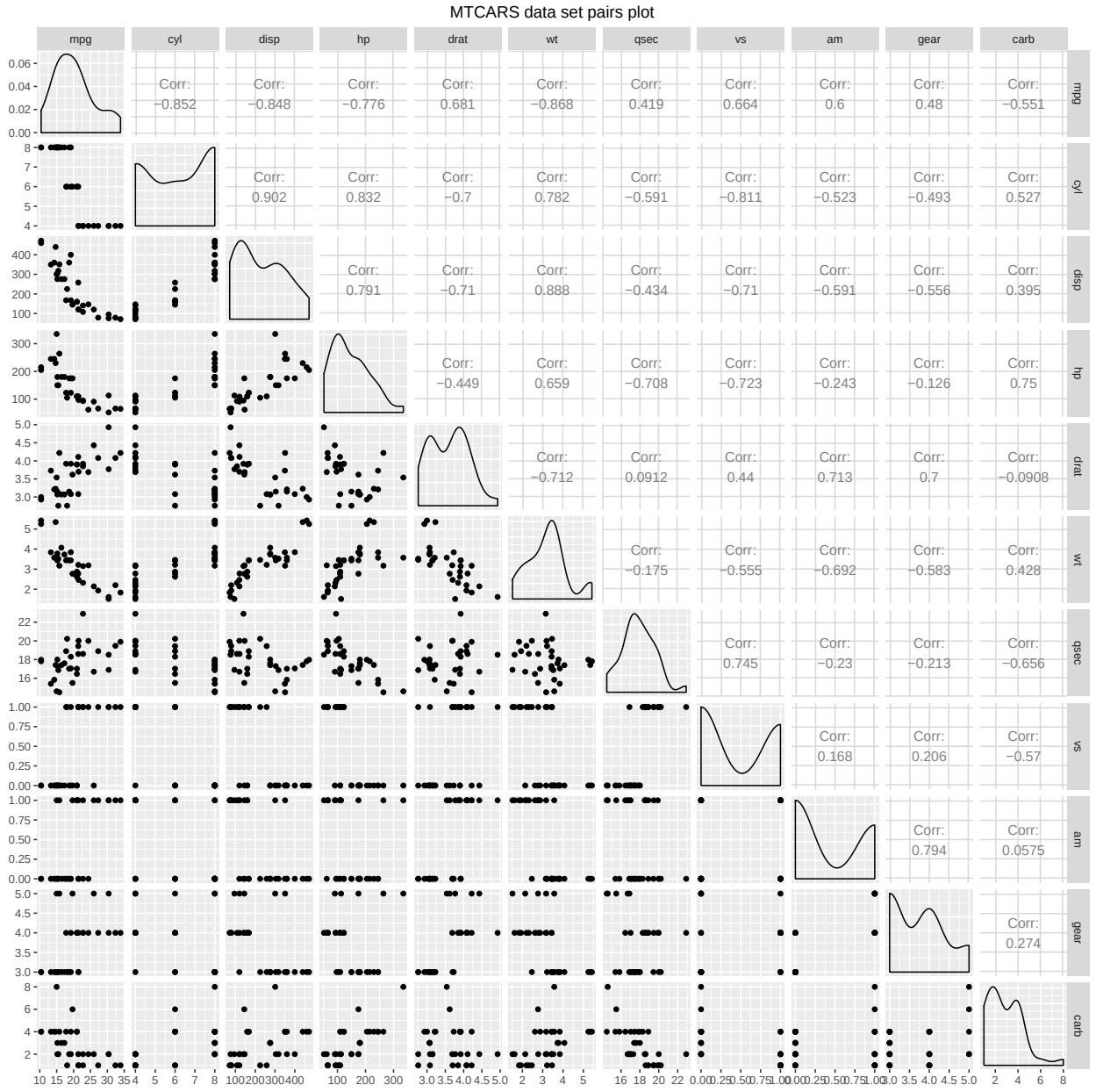
```
theme_update(plot.title = element_text(hjust = 0.5))
ggplot(data=mtcars,aes(y=mpg,x = factor(am)))+geom_boxplot(aes(fill=factor(am)))+
  scale_x_discrete(labels=c("Auto","Manual")) +
  xlab("Transmission Type")+ggtitle("Miles Per Gallon for Transmission Type")+
  theme(legend.position='none')
```




```
knitr::asis_output("\\pagebreak")
```

Figure 2. Plot pairs of the mtcars dataset.

```
data("mtcars")
ggpairs(mtcars)+ggtitle("MTCARS data set pairs plot")
```




```
knitr::asis_output("\\pagebreak")
```

Figure 3. plot Residual panel of the final model.

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