Motor Trend Car Road Tests (Regrssion Models' Assignment)

Rishikesh Pillay

5/22/2021

Data Description: 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 (numeric) variables. You can further read about it on 'mtcars' help page. Here we need two features 1. mpg Miles/(US) gallon 2. am Transmission (0 = automatic, 1 = manual)

Objective Looking at a data set of collection of cars, we are interested in exploring the relationship between a set of variables and miles per gallon (MPG)(outcome).They are particularly interested in the following two questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmission"

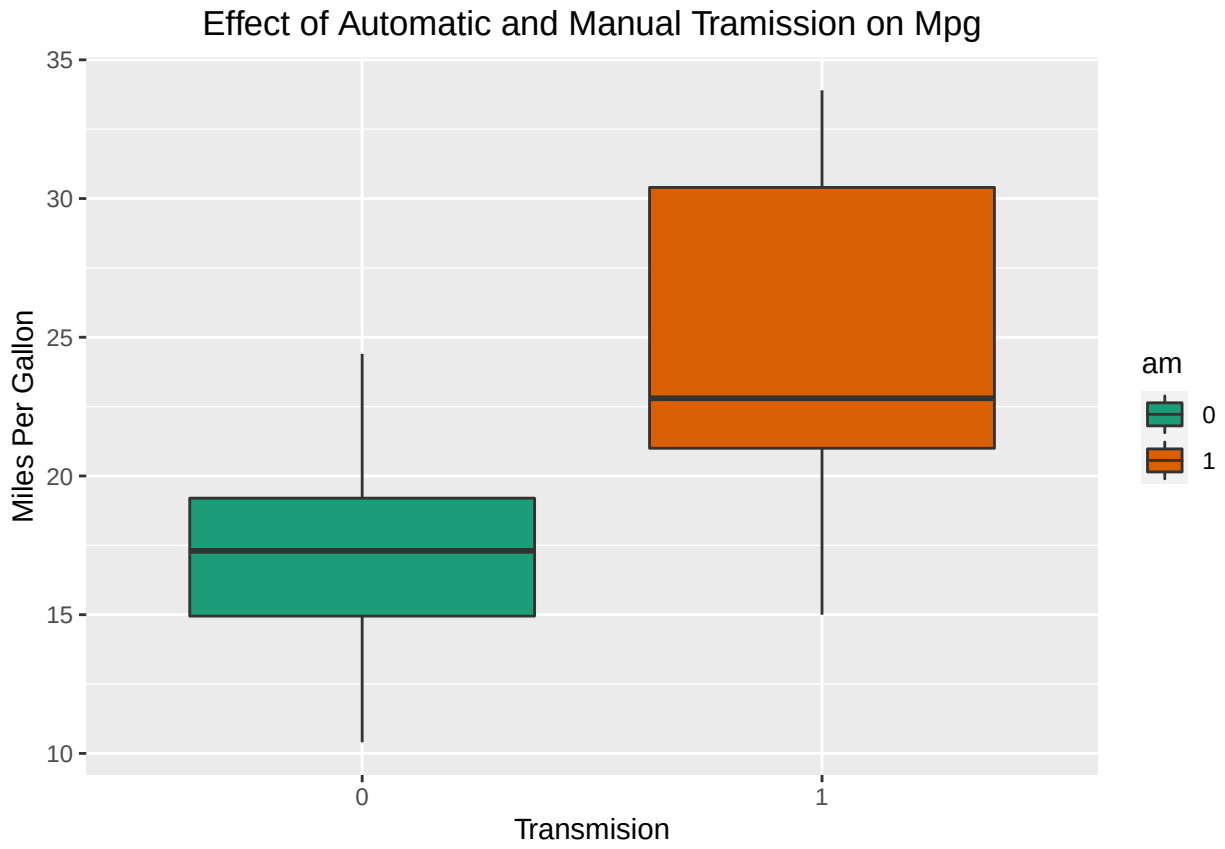
Exploratory Data Analysis

We see there are 11 variables and 32 observatio . There are 19 Automatic Transmission car and 13 Manual.

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
##       mpg        am
##  Min.   :10.40   0:19
##  1st Qu.:15.43   1:13
##  Median :19.20
##  Mean   :20.09
##  3rd Qu.:22.80
##  Max.   :33.90
```

From graph we see that annual Transmission(1) cars have large MPG compare to Automatic Transmission(0). Well now we want to find if it is statistically significant .



Regression Model

There are four assumptions associated with a linear regression model:

1. Linearity: The relationship between X and the mean of Y is linear.
2. Homoscedasticity: The variance of residual is the same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any fixed value of X, Y is normally distributed.

The below is summary is of model where output is mpg and predictor is dummy variable are of transmission (am).

```
##
## Call:
## lm(formula = mpg ~ am, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am1            7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

First notice that two sample T test tells us that there is statistically significant difference in mean of mpg(outcome) when different Transmission is used. This answer our first question ; Manual transmission is better than automatic transmission for MPG. Below we explore Residual plots.

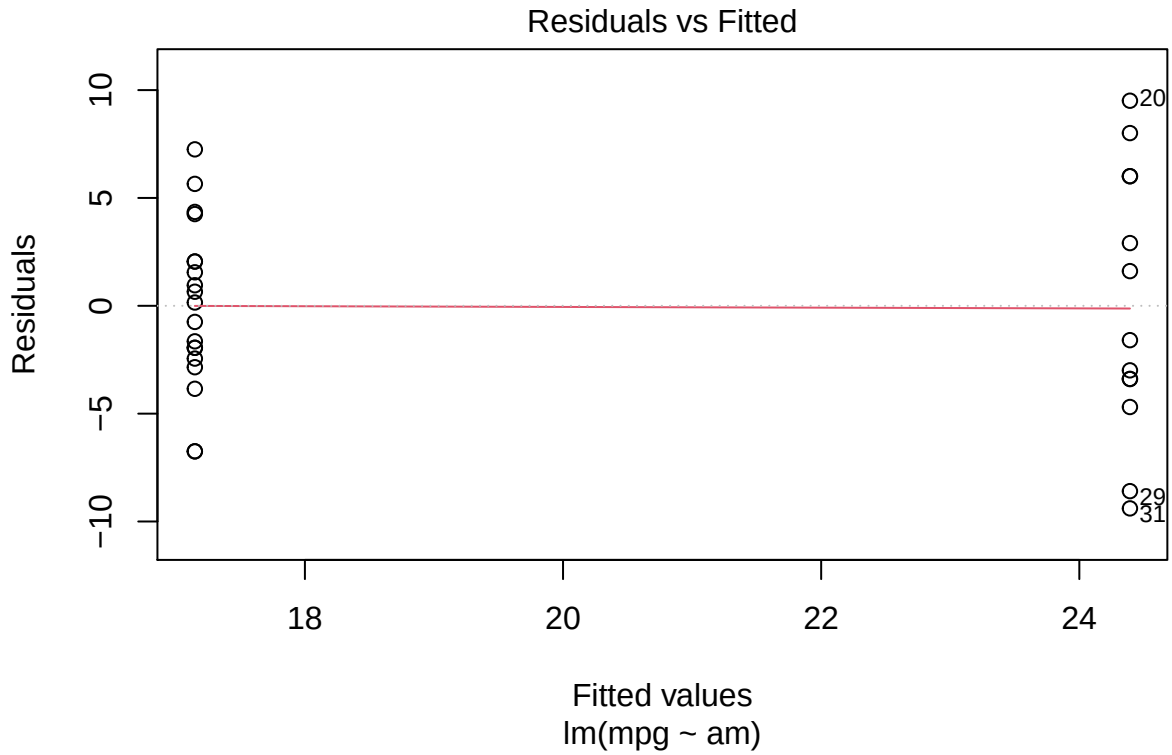
The proportion of the variance for a dependent variable(mpg) that's explained by an independent variable(am) is given by R^2 = 0.3385

Coefficient Interpretation

From above model fit

1. Estimate 17.147 is coefficient of automatic transmission predictor(as it is first level '0') that means 17.147 Intercept is the mean of the mpg when am is "0"
2. 17.147 + 7.244939 = 24.39231 is the mean of mpg when "am" is "1" or Manual transmission I hope this answers our second question "difference in between"Quantify the MPG difference between automatic and manual transmission".

Residual Diagnostics



From residual plot we see patterns which indicate poor model fitting (Underfittied) because we use only one predictor for modeling.

Better model

Assuming that the model is linear with additive iid errors (with finite variance); Two important point to consider is overfiting result in Variance Inflation and Underfiting introduce bias. Also Variance increase dramatically when predictors are highly correlated with each other.

Nested Model

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ am
## Model 3: mpg ~ cyl
## Model 4: mpg ~ cyl + wt + disp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     21 147.49
## 2     30 720.90 -9   -573.40 9.0711 1.779e-05 ***
## 3     30 308.33  0    412.56
## 4     28 188.49  2    119.84 8.5314  0.001941 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix

```
library(ggplot2)
library(dplyr)
data ("mtcars")
cars <- mtcars %>%
        select(mpg,am) %>%
        mutate(am = as.factor(am))
str(mtcars)
summary(cars)


#boxplot
g <- ggplot(cars, aes(am, mpg))
g <- g + geom_boxplot(aes(fill = am))
g <- g +  scale_fill_brewer(palette="Dark2")
g <- g+ xlab('Transmision') + ylab("Miles Per Gallon") + labs(title = "Effect of Automatic and Manual Tr
g <- g + theme(plot.title = element_text(hjust = 0.5))
g

#Regression Model
fit <- lm(mpg ~ am, cars)  # cars have factor "am" variable
summary(fit)
```

```r
#residual plot
plot(fit, which = 1)


#Nested model
library(car)
fit1 <- lm(mpg ~., mtcars)
fit2 <- update(fit1, mpg ~ cyl)
fit3 <- update(fit1, mpg ~ cyl+ wt + disp)
fit <- lm(mpg ~ am, cars)
anova( fit1, fit, fit2, fit3)
```

Note :

1. I have to exceed the 2 pages limit for showing multiple models (apologies)
2. Conclusion is under Regression model and coefficient interpretation
3. And Limitation of model is that residual plot show pattern as our model is Bias