Motor Trend Data Analysis

Dinh Tuan Phan

5/3/2021

Synopsis

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they 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:

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

Install packages

```
library(ggplot2)
```

Load data

```
data(mtcars)
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Summary on data to understand the predictors and outcome

```
summary(mtcars)
```

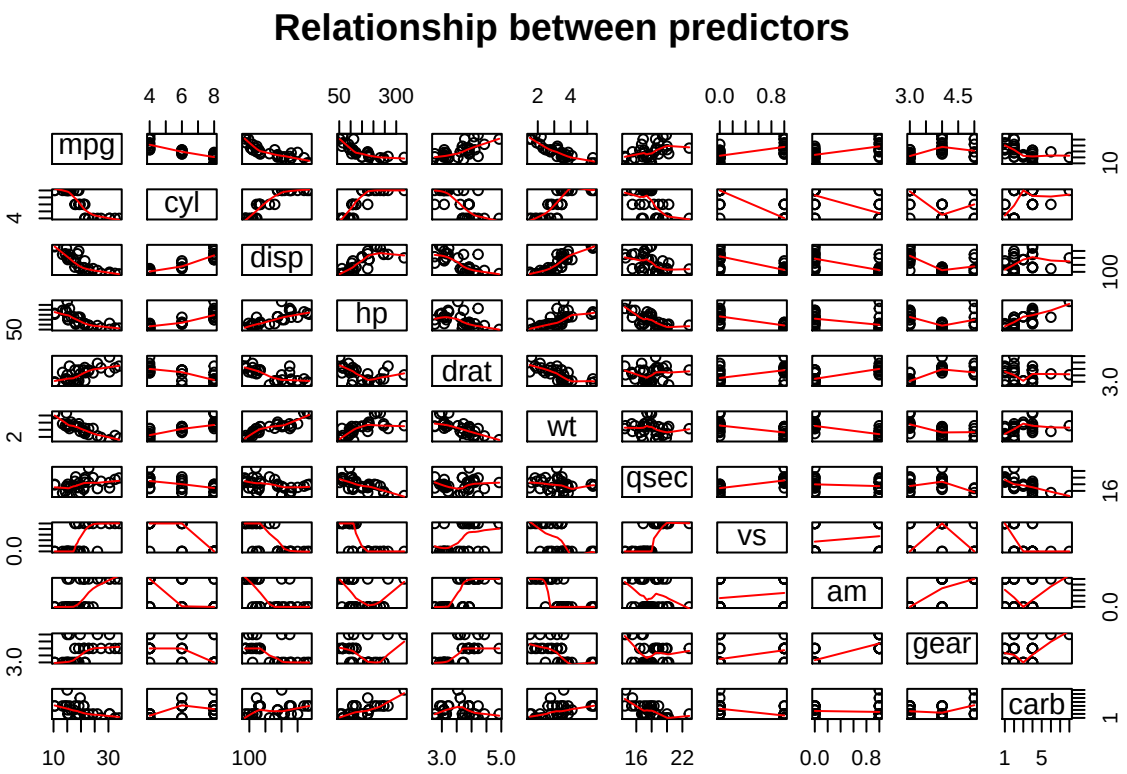
```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
```

```
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

First visualise the data

Run the pair plots to have a first glance on the correlation between input.

```
pairs(mtcars, panel=panel.smooth, main="Relationship between predictors")
```



Run statistical inference

We first run the correlation between the MPG and other predictors.

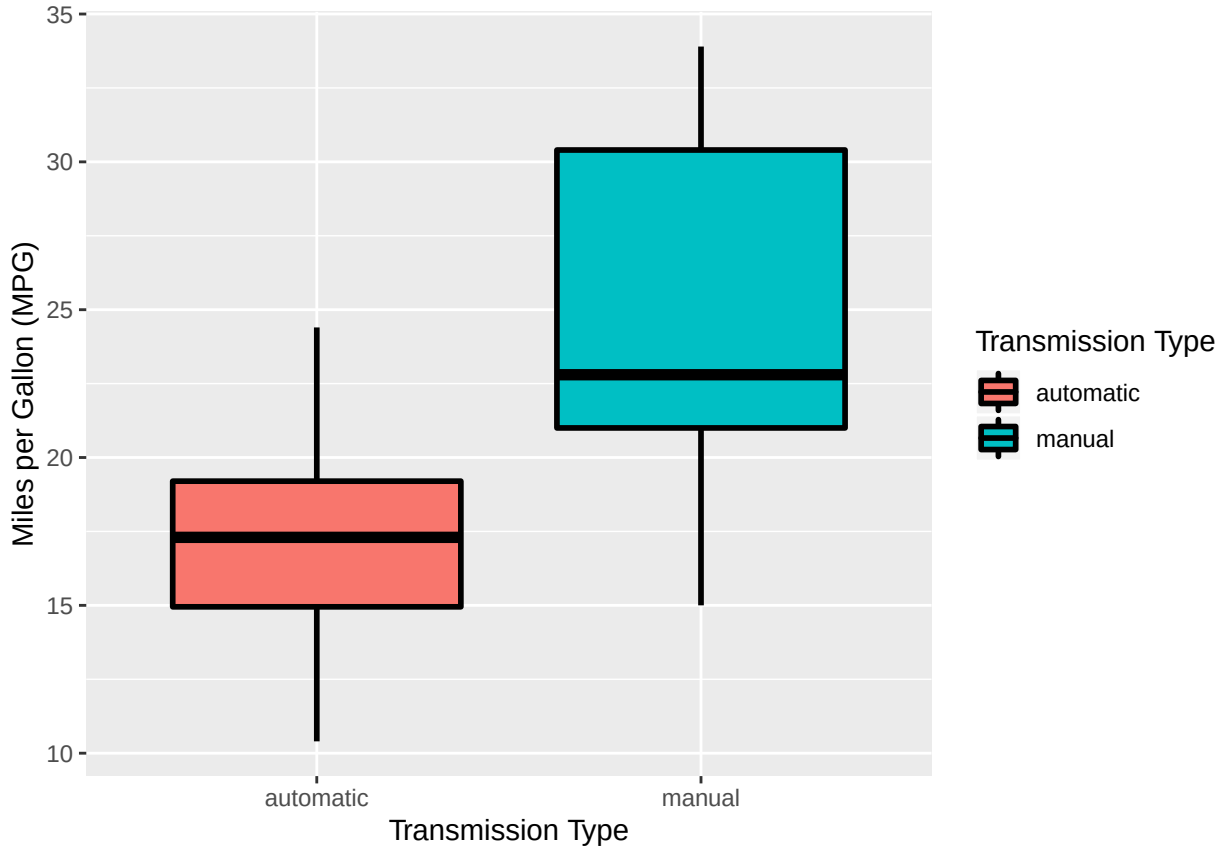
```
cor(mtcars$mpg,mtcars[,-1])
```

```
##            cyl       disp         hp      drat         wt     qsec
## [1,] -0.852162 -0.8475514 -0.7761684 0.6811719 -0.8676594 0.418684
##             vs        am      gear       carb
## [1,] 0.6640389 0.5998324 0.4802848 -0.5509251
```

We can see that the transmission type has a positive correlation (0.5998), so that answers the first question: "Is an automatic or manual transmission better for MPG". The manual is better than automatic in term of mpg. We can check by plotting the am vs mpg.

Plots

```
ggplot(mtcars, aes(y=mpg, x=factor(am, labels = c("automatic", "manual")), fill=factor(am)))+
        geom_boxplot(colour="black", size=1)+
        xlab("Transmission Type") + ylab("Miles per Gallon (MPG)") +
        scale_fill_discrete(name = "Transmission Type", labels = c("automatic", "manual"))
```



Run regression model

We run linear model between all predictors vs outcome (mpg) and we examine parameters in the model.

```
fullModel <-lm(mpg~.,mtcars)
summary(fullModel)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
```

```
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

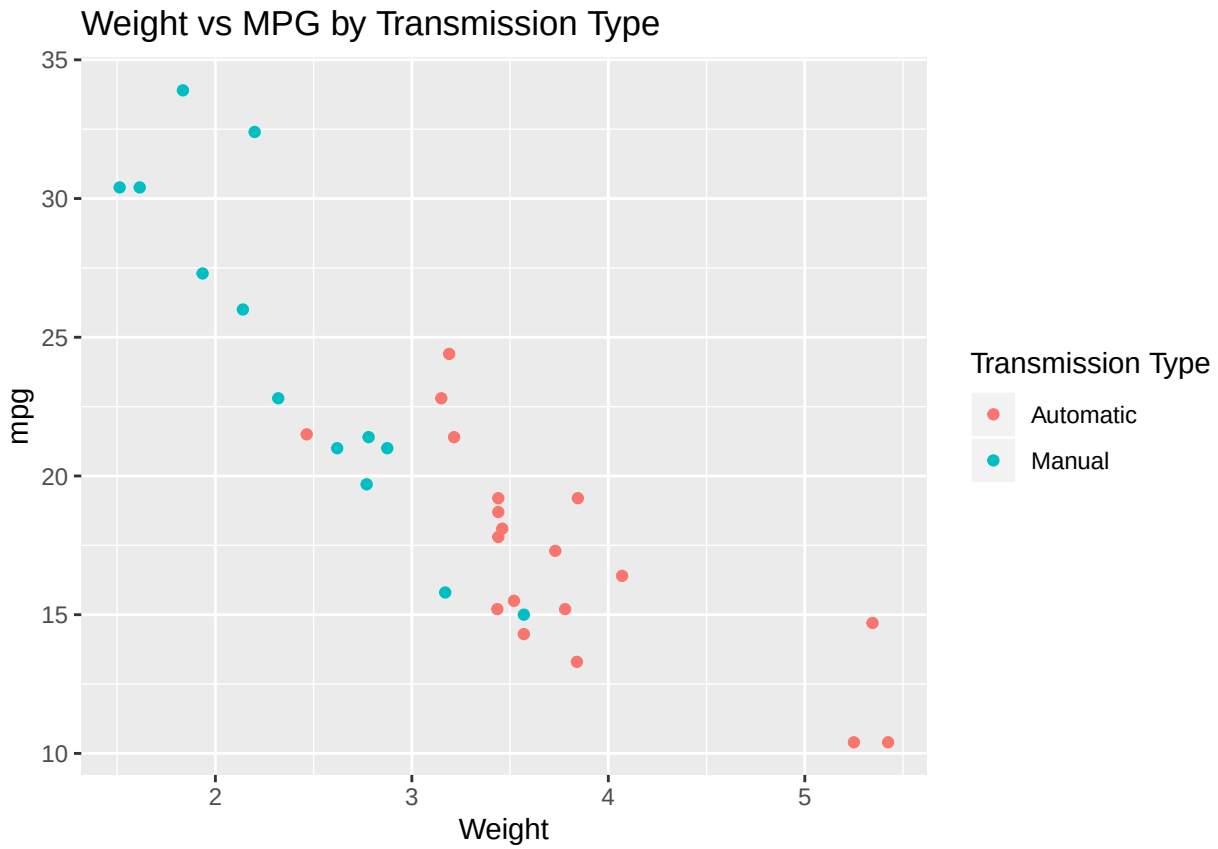
From the results, we can see that only wt is significant to the model because p-value < 0.1 (theoretically, p-values should be smaller than 0.05). We run the model again with only 1 predictor wt.

```
lm1 <-lm(mpg~wt,mtcars)
summary(lm1)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

This is a good model with very low p-value (0.001). It makes sense because the car efficiency (i.e., measured in mpg) heavily depending on its weight. We can further visualize by plotting.

```
mtcars$tran=ifelse(mtcars$am == 0,"automatic","manual")
ggplot(mtcars, aes(x=wt, y=mpg, group=tran, color=tran, height=3, width=3)) + geom_point() +
scale_colour_discrete(name = "Transmission Type",labels=c("Automatic", "Manual")) +
xlab("Weight") + ggtitle("Weight vs MPG by Transmission Type")
```



Next we can try to run the model with one more predictor (i.e., the Transmission am) because predictor am has the lowest p-value among the remaining predictors.

```
lm2 <-lm(mpg~wt+am,mtcars)
summary(lm2)
```

```
##
## Call:
## lm(formula = mpg ~ wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.32155    3.05464  12.218 5.84e-13 ***
## wt          -5.35281    0.78824  -6.791 1.87e-07 ***
## am          -0.02362    1.54565  -0.015    0.988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 29 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7358
## F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

The predictor wt is significant (p-value less than 0.001) but am is not significant (p-value is to high 0.988) in

the model.

Next, we might want to run a linear model with only predictor am to exam it independently.

```
lm3 <-lm(mpg~am,mtcars)
summary(lm3)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## 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 ***
## am             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
```

This model seems to answer the question. The car efficiency (i.e., measured in mpg) depends on the transmission type.

The answer for the second question: "Quantify the MPG difference between automatic and manual transmissions" is: the cars with manual transmission is 7.245 mpg better than the automatic cars.

Best model

```
bestmodel = step(lm(data = mtcars, mpg ~ .), trace=0)
summary(bestmodel)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
```

```
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Residual Analysis

We observe that the residuals seems randomly scattered around zero and that quantiles for the residuals fall somewhat close to the theoretical normal quantiles in the Q-Q plot.

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