Effect of Transmission Type on Car's Mileage

MOHAMMAD SHADAN

September 17, 2016

EXECUTIVE SUMMARY

Analyzing the data from the 1974 Motor Trend US magazine, which comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

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**	V/S	**carb**	Number of carburetors
am	Transmission Type	**gear**	Num of forward gears		

Analysis will elaborate on the below two points :
- Is an automatic (0) or manual (1) transmission better for MPG
- Quantify the MPG difference between automatic and manual transmissions

EXPLORATORY DATA ANALYSIS

Scatterplot and Boxplot for Miles Per Gallon (mpg) vs. Transmission Type (am). "+" sign represents mean

Figure 1 − Scatter Plot

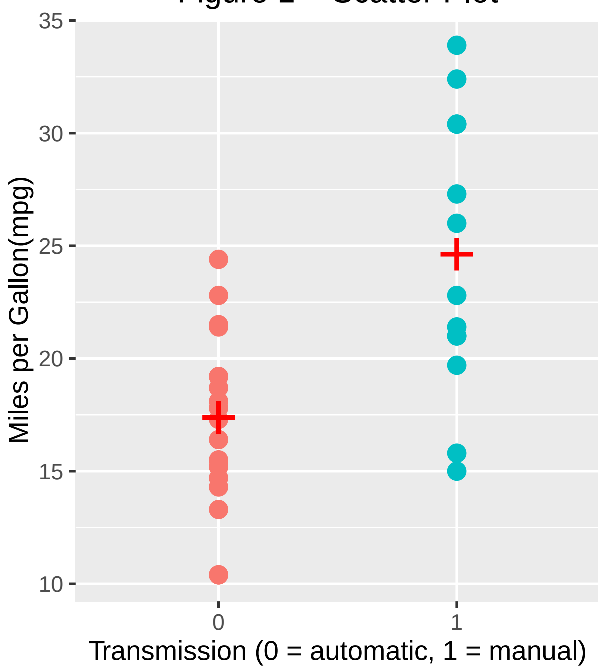
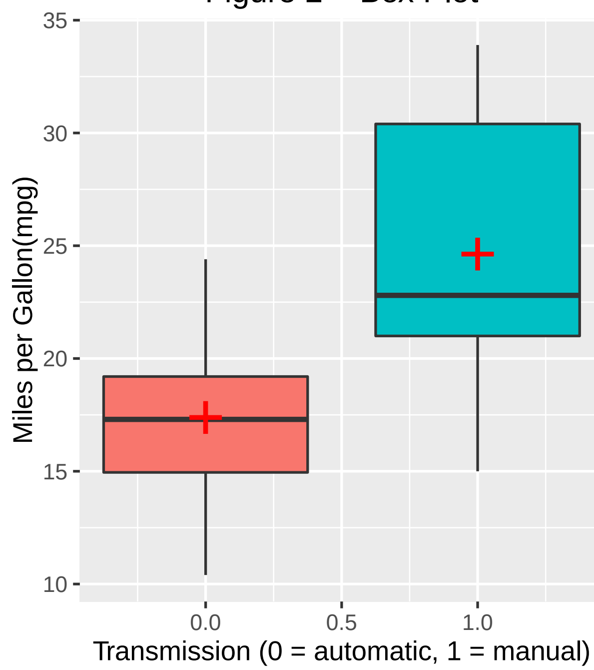


Figure 2 − Box Plot



- Average mpg for automatic and manual transmission are **17.14737** and **24.39231** respectively
- Considering other variables constant, Manual cars travel more on per gallon fuel compared to Automatic cars
- We need to analyze and select other varibles including Transmission Type which affect "mpg" to get optimized result

SELECTING BEST FIT MODEL

I have used **Backwards Elimination (p-value) Method** to find the best fit model. Steps involved are :
- Start with the full model
- Drop the variable with the highest p-value and refit a smaller model
- Repeat until all variables left in the model are significant

Based on backwards elimination (p-value) method the best model fit is :

lm(mpg ~ am + wt + qsec, data=mtcars)

Step by Step analysis for finalizing above model using Backwards Elimination (p-value) Method and R outputs are shown in APPENDIX

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec, 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
## am            2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## 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
```

CONCLUSION

$$mpg = 9.6178 + 2.9358am - 3.9165wt + 1.2259qsec$$

Is an automatic or manual transmission better for MPG

- Manual Transmission is better than Automatic Transmission for Miles Per Gallon, but assuming all other predictors (cyl, disp, hp, drat, wt, qsec, vs, gear, carb) constant

Quantify the MPG difference between automatic and manual transmissions

- Considering weight (wt) and accelaration (qsec) speed also, Manual Cars run **2.9358** Miles Per Gallon more than the cars with Automatic Transmisson

APPENDIX

```r
#Set the Working Directory setwd("~/R/RM"), Load Required Libraries and mtcars dataset
library(ggplot2); library(gridExtra); data(mtcars); attach(mtcars);
#Figure 1 - Scatter Plot
g1 <- ggplot(mtcars, aes(x=factor(am), y=mpg)) + geom_point(aes(color=factor(am)), size=3)
g1 <- g1 + stat_summary(fun.y=mean, geom="point", shape="+", size=9, col = "red")
g1 <- g1 + theme(legend.position="none") + labs(title="Figure 1 - Scatter Plot")
g1 <- g1 + labs(x = "Transmission (0 = automatic, 1 = manual)", y = "Miles per Gallon(mpg)")
#Figure 2 - Boxplot
g2 <- ggplot(mtcars, aes(x=am, y=mpg, group=am,fill=factor(am))) + geom_boxplot()
g2 <- g2 + stat_summary(fun.y=mean, geom="point", shape="+", size=9, col = "red")
g2 <- g2 + theme(legend.position="none") + labs(title="Figure 2 - Box Plot")
g2 <- g2 + labs(x = "Transmission (0 = automatic, 1 = manual)",y = "Miles per Gallon(mpg)")
```

Step by step analysis using Backwards Elimination (p-value) to find the best fit model :

```r
#Step 1 : Start with the full model
pfit1 <- lm(mpg ~ ., data=mtcars); coef(summary(pfit1))
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

```r
#Step 2 : cyl has the largest pvalue(0.91608738), so drop cyl and refit the model
pfit2 <- lm(mpg ~ am + disp + hp + drat + wt + qsec + vs + gear + carb, data=mtcars)
coef(summary(pfit2))
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 10.96007405 13.53030251  0.8100391 0.42659327
## am           2.57742789  1.94034563  1.3283344 0.19768373
## disp         0.01282839  0.01682215  0.7625891 0.45380797
## hp          -0.02190885  0.02091131 -1.0477031 0.30615002
## drat         0.83519652  1.53625251  0.5436584 0.59214373
## wt          -3.69250814  1.83953550 -2.0073046 0.05715727
## qsec         0.84244138  0.68678068  1.2266527 0.23291993
## vs           0.38974986  1.94800204  0.2000767 0.84325850
## gear         0.71155439  1.36561933  0.5210489 0.60753821
## carb        -0.21958316  0.78855537 -0.2784626 0.78325783
```

```r
#Step 3 : vs has the largest pvalue (0.84325850), so drop vs and refit the model
pfit3 <- lm(mpg ~ am + disp + hp + drat + wt + qsec + gear + carb, data=mtcars); coef(summary(pfit3))
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept)  9.76827789 11.89230469  0.8213949 0.41985460
## am           2.52390094  1.88128007  1.3415870 0.19282690
## disp         0.01214441  0.01612373  0.7532010 0.45897019
## hp          -0.02095020  0.01992567 -1.0514175 0.30398892
## drat         0.87509822  1.49112525  0.5868710 0.56300717
## wt          -3.71151106  1.79833544 -2.0638592 0.05049085
## qsec         0.91082822  0.58311935  1.5619928 0.13194532
## gear         0.75984464  1.31577205  0.5774896 0.56921947
## carb        -0.24796312  0.75933250 -0.3265541 0.74695821
```

```r
#Step 4 : carb has the largest pvalue (0.74695821), so drop carb and refit the model
pfit4 <- lm(mpg ~ am + disp + hp + drat + wt + qsec + gear, data=mtcars); coef(summary(pfit4))
```

```
##                Estimate  Std. Error    t value    Pr(>|t|)
## (Intercept)  9.19762837 11.54220381  0.7968693 0.433339841
## am           2.58979984  1.83528342  1.4111171 0.171042438
## disp         0.01551976  0.01214235  1.2781513 0.213420001
## hp          -0.02470716  0.01596302 -1.5477746 0.134763097
## drat         0.81022794  1.45006779  0.5587518 0.581507634
## wt          -4.13065054  1.23592980 -3.3421401 0.002717119
## qsec         1.00978651  0.48883274  2.0657097 0.049814778
## gear         0.60644020  1.20596266  0.5028681 0.619640616
```

```r
#Step 5 : gear has the largest pvalue (0.619640616), so drop gear and refit the model
pfit5 <- lm(mpg ~ am + disp + hp + drat + wt + qsec, data=mtcars); coef(summary(pfit5))
```

```
##                Estimate  Std. Error    t value    Pr(>|t|)
## (Intercept) 10.71061639 10.97539399  0.9758753 0.338475309
## am           2.98468801  1.63382423  1.8268110 0.079692318
## disp         0.01310313  0.01098299  1.1930387 0.244054196
## hp          -0.02179818  0.01465399 -1.4875257 0.149381426
## drat         1.02065283  1.36747598  0.7463772 0.462401185
## wt          -4.04454214  1.20558182 -3.3548467 0.002536163
## qsec         0.99072948  0.48002393  2.0639168 0.049550895
```

```r
#Step 6 : drat has the largest pvalue (0.462401185), so drop drat and refit the model
pfit6 <- lm(mpg ~ am + disp + hp + wt + qsec, data=mtcars); coef(summary(pfit6))
```

```
##                Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 14.36190396 9.74079485  1.474408 0.152378367
## am           3.47045340 1.48578009  2.335779 0.027487809
## disp         0.01123765 0.01060333  1.059823 0.298972150
## hp          -0.02117055 0.01450469 -1.459565 0.156387279
## wt          -4.08433206 1.19409972 -3.420428 0.002075008
## qsec         1.00689683 0.47543287  2.117853 0.043907652
```

```r
#Step 7 : disp  has the largest pvalue (0.298972150), so drop disp and refit the model
pfit7 <- lm(mpg ~ am + hp + wt + qsec, data=mtcars); coef(summary(pfit7))
```

```
##                Estimate Std. Error   t value    Pr(>|t|)
```

```
## (Intercept) 17.44019110  9.3188688  1.871492 0.072149342
## am           2.92550394  1.3971471  2.093913 0.045790788
## hp          -0.01764654  0.0141506 -1.247052 0.223087932
## wt          -3.23809682  0.8898986 -3.638726 0.001141407
## qsec         0.81060254  0.4388703  1.847021 0.075731202
```

```
#Step 8: hp  has the largest pvalue (0.223087932), so drop hp and refit the model
pfit8 <- lm(mpg ~ am + wt + qsec, data=mtcars); coef(summary(pfit8))
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## am           2.935837  1.4109045  2.080819 4.671551e-02
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
```

Since all the p values are significantly low, below is our best fit model and should not remove more predictors

lm(mpg ~ am + wt + qsec, data=mtcars)

RESIDUAL PLOT

```
par(mfrow=c(2,2)); plot(pfit8); par(mfrow=c(1,1))
```