Better mpg: manual or automatic transmission?

Jacky Wang

September 26, 2015

Contents

Introduction

In this report, we will analysis the *mtcars* data set to answer:

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

Preproces and Explore the data

We use categorical variable to denote:

- *vs* ~ the engine type
- *am* ~ whether manual or automatic

```
library(ggplot2); library(GGally); library(dplyr, warn.conflicts=FALSE); data(mtcars)
raw_data <- mtcars
processed_data <- mutate(mtcars, am = factor(ifelse(am == 0, "automatic", "manual")),
                         vs = factor(ifelse(vs == 0, "v", "s")))
```

We use *ggpairs* to explore relation bewtten *mpg* and *am*. The first figure in Appendix shown that: manual cars seems have a better mpg than automatic ones.

Regression Analysis

Previous figure only shows the relationship between *mpg* and *am*, other feature are ignored. In this section, we will use linear regression model to answer the target question.

We will use *step()* to choose the model by AIC in a stepwise algorithm, which suggests using *wt*, *am* and *qsec*.

```
model <- step(lm(mpg ~ ., processed_data), trace=FALSE)
coefficients(model)
```

```
(Intercept)          wt        qsec    ammanual
   9.617781   -3.916504    1.225886    2.935837
```

Add the inteaction terms (all *wt:qsec, wt:am, qsec:am* are tried, based on p-value and adjusted R^2, *am:wt* is used.):

```
final_model <- lm(mpg ~ wt + am + qsec + am:wt, processed_data)
coefficients(final_model)
confint(final_model)
```

```
(Intercept)          wt    ammanual        qsec wt:ammanual
   9.723053   -2.936531   14.079428    1.016974   -4.141376
                 2.5 %    97.5 %
(Intercept) -2.3807791 21.826884
wt          -4.3031019 -1.569960
ammanual     7.0308746 21.127981
qsec         0.4998811  1.534066
wt:ammanual -6.5970316 -1.685721
```

Since the 95% confidence interval of intercept contains zero(large p value), we are not able to reject $H_{NULL} : \beta_0 = 0$. So the final model is given by:

$$mpg = -2.937wt + 1.017qsec + (14.079 - 4.141wt)am_{manual} \tag{1}$$

And the *anova()* shows the prediction is improved by the final model.

```
anova(model, final_model)
```

```
Analysis of Variance Table

Model 1: mpg ~ wt + qsec + am
Model 2: mpg ~ wt + am + qsec + am:wt
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     28 169.29
2     27 117.28  1     52.01 11.974 0.001809 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The residual plot and diagnostics are given in appendix.

Conclusion

From previous model, for **fixed values of** wt **and** $qsec$, the average difference of *mpg* is given by:

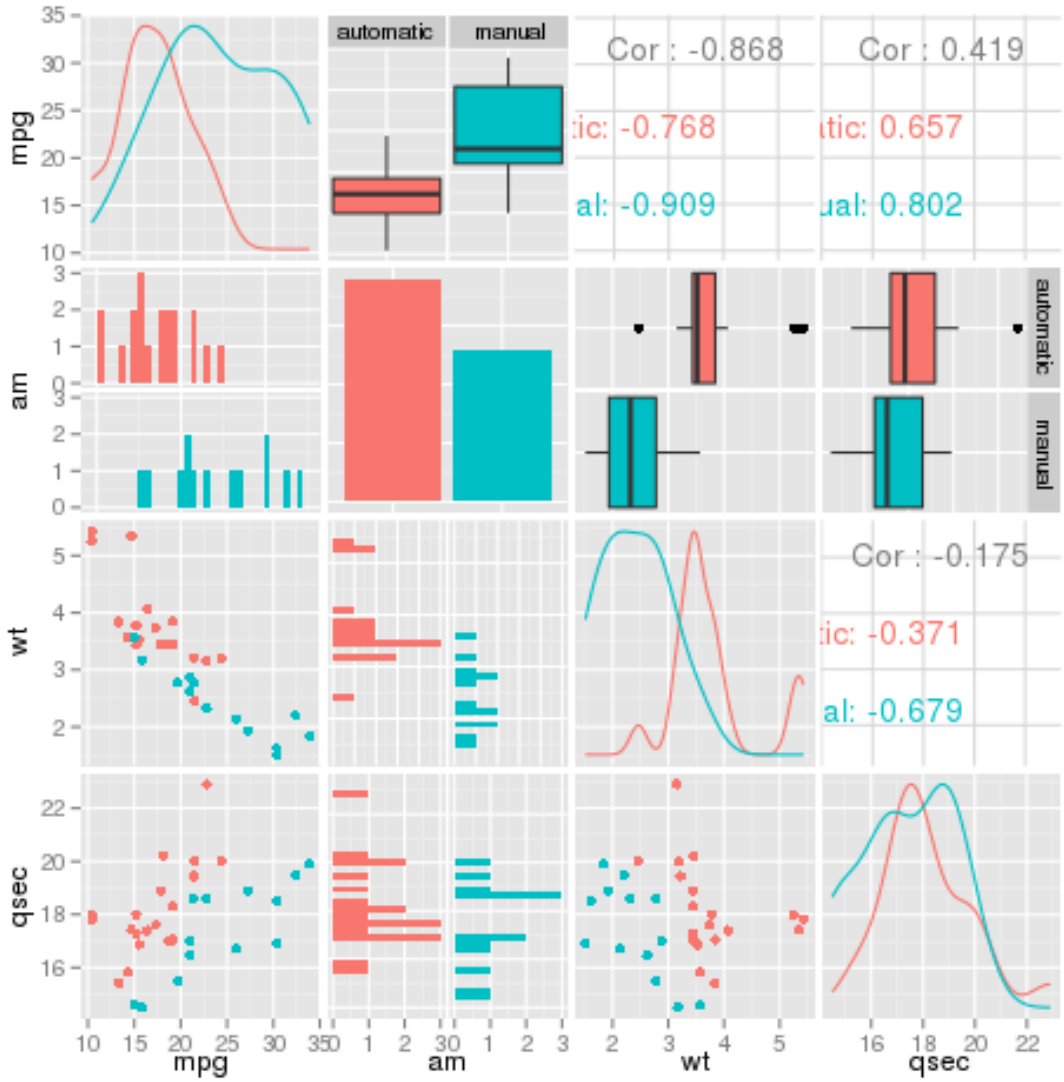
$$mpg_{manual} - mpg_{automatic} = 14.079 - 4.141wt = \begin{cases} > 0 & \text{if } wt \leq 3.40 \\ < 0 & \text{if } wt > 3.40 \end{cases} \tag{2}$$

Thus, the answer of the target quesition can't be answered directly, it dpends on $wt, qsec$.

Appendix

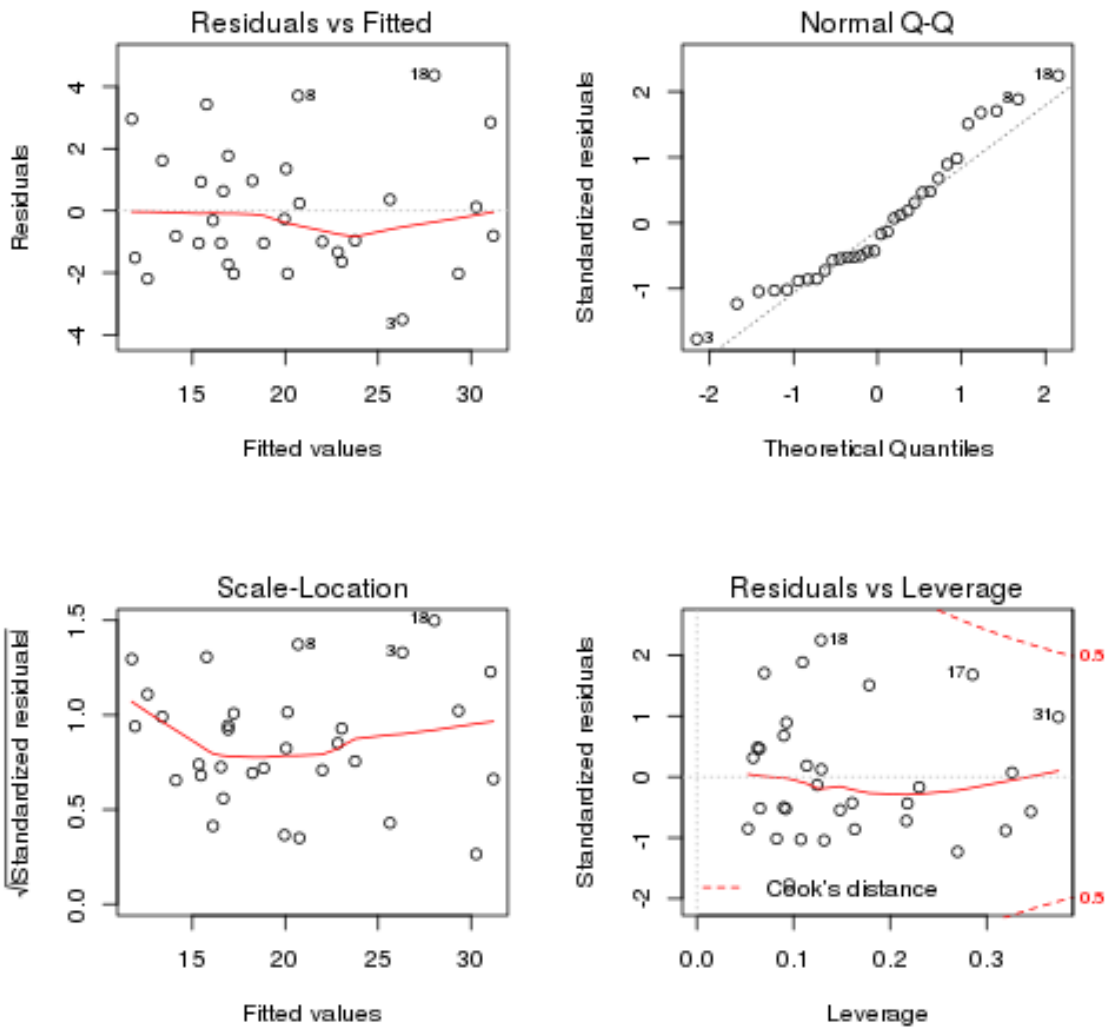
Scatter plot matrix

```
library(GGally); library(dplyr, warn.conflicts=FALSE); options(warn=-1)
ggpairs(select(processed_data, mpg, am, wt, qsec), color='am')
```



Diagnostic

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



Summary of two linear models

```
summary(model)
summary(final_model)
```

```
Call:
lm(formula = mpg ~ wt + qsec + am, data = processed_data)

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 ***
ammanual      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

Call:
lm(formula = mpg ~ wt + am + qsec + am:wt, data = processed_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5076 -1.3801 -0.5588  1.0630  4.3684

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.723      5.899   1.648 0.110893
wt            -2.937      0.666  -4.409 0.000149 ***
ammanual      14.079      3.435   4.099 0.000341 ***
qsec           1.017      0.252   4.035 0.000403 ***
wt:ammanual   -4.141      1.197  -3.460 0.001809 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.084 on 27 degrees of freedom
Multiple R-squared:  0.8959,Adjusted R-squared:  0.8804
F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```