Miles per Gallon for Different Type of Transimissions

yangru1q1

2019-07-07

Synopsis

In this project, we try to anwer questions regard whether or not cars with different transimission perform differently in mile/gallon with the *mtcars* dataset. Using a unpaired two sample t-test we can conclude there is a difference in mean, munual transmission has a better performance in miles per gallon. After building a linear model with mpg (miles per gallon) as outcome and transmission as one of the predictor, the model tells us the expected miles per gallon for different type of transmissions. For the model and it's interpretation please check the **Conclusion** part.

Data Exploratory

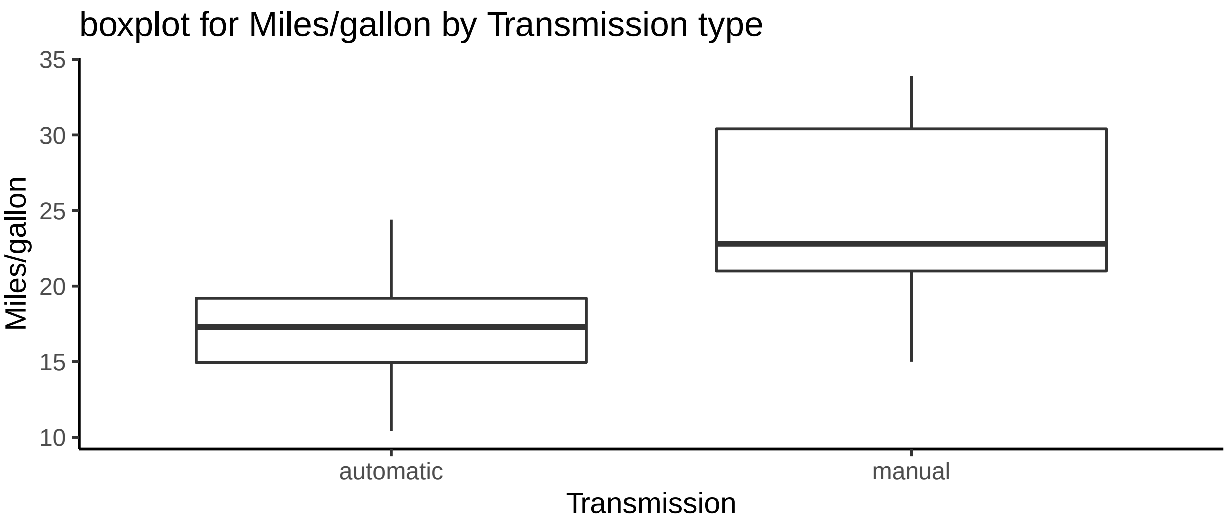
```r
require(tidyverse)
```

```r
data("mtcars")
mtcars_tb <- as_data_frame(mtcars)
mtcars_tb <- mtcars_tb %>% mutate(vs = factor(vs), am = factor(am))
glimpse(mtcars_tb)
```

```
## Observations: 32
## Variables: 11
## $ mpg  <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19....
## $ cyl  <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4, ...
## $ disp <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 1...
## $ hp   <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, ...
## $ drat <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.9...
## $ wt   <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3...
## $ qsec <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 2...
## $ vs   <fct> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, ...
## $ am   <fct> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, ...
## $ gear <dbl> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, ...
## $ carb <dbl> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, ...
```

Since *vs* (0 = V-shaped, 1= straight) describe the type of engine, and *am* (0 = automatic, 1 = manual) describe type of Transmission, we can transform these to variables into factor. We should to convert the column *cyl* as factor as well, but we actually don't interest in this variable, **and it make no difference if we choose to convert it into factor**. Since the way I perform my analysis, I choose to leave *cyl* as a continuous variable.

```r
mtcars_tb %>% mutate(am_type = ifelse(am == 0, "automatic", "manual")) %>%
        ggplot() + theme_classic() + geom_boxplot(aes(x = am_type, y = mpg)) +
        labs(title = "boxplot for Miles/gallon by Transmission type",
             x = "Transmission", y = "Miles/gallon")
```



We can see from above plot, the equal variance for Miles per gallon between different Transmission group may not hold, let's explore more.

```
kable(
mtcars_tb %>% mutate(am_type = ifelse(am == 0, "automatic", "manual")) %>%
        group_by(am_type) %>% summarize(mean = mean(mpg), sd = sd(mpg), count = n()))
```

am_type	mean	sd	count
automatic	17.14737	3.833966	19
manual	24.39231	6.166504	13

Is an automatic or manual transmission better for MPG?

From both boxplot and table above, we can see mean Miles/gallon for manual type of transimission is greater than automatic type of transimission. Let's test how significant this mean difference is. Since the huge difference in standard deviation for each transmission, I prefer a two sample t-test with different variance.

```
t.test(mtcars_tb$mpg[mtcars_tb$am == 0], mtcars_tb$mpg[mtcars_tb$am == 1],
       var.equal = FALSE,paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars_tb$mpg[mtcars_tb$am == 0] and mtcars_tb$mpg[mtcars_tb$am == 1]
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

We use the two sided t-test above. However, we can still conclude that manual transmission better for MPG at a 95% confidence level.
Notice that we also construct a simple linear model use mpg as outcome with only am as independent dummy variable, the result is the same as a two sided paired t-test with equal variance, which will agree with our conclusion as well.

Quantify the MPG difference between automatic and manual transmission

In order to answer this question we need to build a model with mpg as outcome and transimission as one of the predictor. The strategy I will use is **Backward Elimination**, which is fit a model with all variable except mpg as predictor, and delete one variable with the largest p-value for coefficient at a time till all coefficient significant.

Model Generation

```
model <- function(mdl){
        # backward elimination method to generate the correct model with all
        # coeffecients significant for mtcars dataset
        #       INPUT:     a simple linear regression (Full model)
        #       OUTPUT:    a simple linear regression model with all coeffecient
        #                  significant
        data <- mtcars_tb
        pvals <- summary(mdl)$coefficients[-1, 4]
        maxp <- max(pvals)
        while(maxp > 0.05){
                # which column of the new dataset need to be deleted
                dlt <- which.max(summary(mdl)$coef[, 4])
                data <- data[, -dlt]
                mdl <- lm(mpg~., data)
                pvals <- summary(mdl)$coefficients[-1, 4]
                maxp <- max(pvals)
        }
        mdl
}
mdl <- model(lm(mpg~., mtcars_tb))
```

In order to reduce the amount space used for the correct model generation process, I wrote a function called **model** to perform the backward elimination for us. With the pre-analysis experince with the *mtcars* dataset, I know the final model with backward elimination method will have the transimission as a predictor. Once again, **it makes no difference if we convert cyl as a factor**. Let's explore our model.

```
kable(summary(mdl)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
wt	-3.916504	0.7112016	-5.506882	0.0000070
qsec	1.225886	0.2886696	4.246676	0.0002162
am1	2.935837	1.4109045	2.080819	0.0467155

As we can noticed, the p-value for intercept is not significant, it means the constant in our model will not significantlly differ from zero if we hold other variable 0. I don't think we should worry too much about the intercept, since we only have 32 observations in the *mtcars* dataset, and we don't have a single observation has weight and 1/4 mile time 0 with a automatic transmission.

Analysis of the Intersection term

Intersection between weight and transmission

```
mdl_int1 <- update(mdl, .~. + wt:am, data = mtcars_tb)
kable(summary(mdl_int1)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.723053	5.8990407	1.648243	0.1108925
wt	-2.936531	0.6660253	-4.409038	0.0001489
qsec	1.016974	0.2520152	4.035366	0.0004030
am1	14.079428	3.4352512	4.098515	0.0003409
wt:am1	-4.141376	1.1968119	-3.460340	0.0018086

Intersection between 1/4 Mile time and transmission

```
mdl_int2 <- update(mdl, .~. + qsec:am, data = mtcars_tb)
kable(summary(mdl_int2)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.5289983	7.2670657	2.274508	0.0310965
wt	-3.7771965	0.6710599	-5.628703	0.0000057
qsec	0.8169222	0.3298625	2.476554	0.0198240
am1	-15.6137466	8.6237370	-1.810555	0.0813526
qsec:am1	1.0600292	0.4869562	2.176847	0.0384079

```
adjustR.table <- data.frame("No intersection" = summary(mdl)$adj.r.squared,
                        "Weight and Transmission" = summary(mdl_int1)$adj.r.squared,
                        "qsec and Transmission" = summary(mdl_int2)$adj.r.squared,
                         row.names = "Adjusted R-squared")
kable(adjustR.table)
```

	No.intersection	Weight.and.Transmission	qsec.and.Transmission
Adjusted R-squared	0.8335561	0.8804219	0.8531624

Observe from above summary tables, both intersections term are significant. The model with intersection between wight and transimissions is better since a higher adjusted R-squared value. Furthermore, in the model with qsec (1/4 mile time) and transimission, the non-significant of coefficent of munual (am = 1) makes us hard to interpretate what we interested in.
Last thing we should consider is if a model with 2 intersection term help improve the model.

Anova test 2 intersection terms

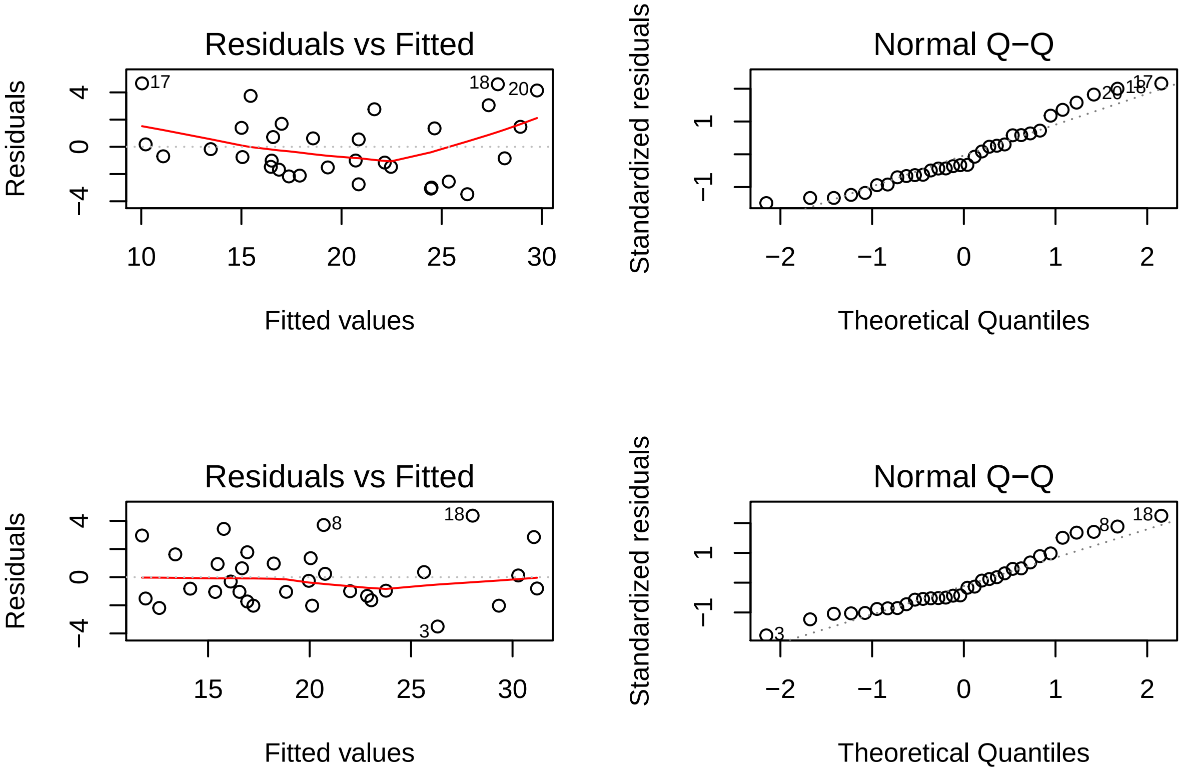
```
mdl_int3 <- update(mdl_int1, .~. + qsec:am, mtcars_tb)
anova(mdl_int1, mdl_int3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + qsec + am + wt:am
## Model 2: mpg ~ wt + qsec + am + wt:am + qsec:am
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 117.28
## 2     26 116.47  1   0.80231 0.1791 0.6756
```

Result from the anova tells us, the model with 2 intersection terms does not perform better than the model with only 1 intersection term (weight and transimission).

Model Diagnostics

```
par(mfrow = c(2, 2))
plot(lm(mpg~wt+qsec+am, mtcars_tb), which = 1)
plot(lm(mpg~wt+qsec+am, mtcars_tb), which = 2)
plot(mdl_int1, which = 1)
plot(mdl_int1, which = 2)
```



One important Gauss-Markov assumption for using OLS is error with mean zero and constant variance, we can check the assumption from residual v.s. fitted value plot. The top left plot is the residual v.s. fitted value for model with no intersection term, we observe there's some tiny trend for the residual mean and variance, but we don't need to worry about the tiny trend since we only have a small amount of observations. This model is actually pretty good.
The bottom left plot is the residual v.s. fitted value for the model with weight and transimission intersection. Observed that the residual performed very well in the model, which means the intersection term does improve the orginal model and have a better performance.
From the normal QQ plot on the right side (top for no intersection, bottom for intersection) we can see both models' residual has some tiny right-skewed problem. Once again, the normality of the residual for the model is actually okay for me since the observation size.

Conclusion

Our final model is $\hat{Y} = 9.723 - 2.937 * X_{wt} + 1.1017X_{qsec} + 14.079 * X_{am=1} - 4.141 * X_{wt} * X_{am=1}$, this means **the change of mpg with different transimission depends on wight**, and if we hold 1/4 miles time of a car constant, per 1000 lbs increase in weight (1 increase in X_{wt}) we expect the car with manual transmission has $14.097 - 4.141 = 9.983$ more miles/gallon than those with automatic transmission.