markdown

Ji

2022-10-20

## Regression Models Course Project

### Introduction

This project is an analysis of the **mtcars** data.The goal is to answer two questions:  
1. Is an automatic or manual transmission better for MPG?  
2. Quantify the MPG difference between automatic and manual transmissions.  
### Read the data mtcars is in the “datasets” package. It has 32 observation on 11 variables. The unit of mpg is miles/US gallon. Several parameters can affect mpg, including number of cylinders (cyl), displacement (disp), gross horsepower (hp), rear axle ratio (drat), weight (wt), 1/4 mile time (qsec), engine (vs), transmission (am), number of forward gears (gear), number of carburetors (carb).These parameters may have positive or negative correlation with mpg.

data(mtcars)  
str(mtcars)

## '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 ...

### Is an automatic or manual transmission better for MPG?

length(which(mtcars$am==0))

There are 19 samples where am = 0, and 13 samples where am = 1. Since the population standard deviation is unknown, a t-test is used to verify the difference between automatic and manual transmission. The **null** hypothesis is that automatic and manual transmission has the **same** mpg.

t.test(mpg ~ am, mtcars)

##   
## Welch Two Sample t-test  
##   
## data: mpg by am  
## t = -3.7671, df = 18.332, p-value = 0.001374  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -11.280194 -3.209684  
## sample estimates:  
## mean in group 0 mean in group 1   
## 17.14737 24.39231

Here group 0 and 1 are automatic and manual transmission. The mean of manual transmission is larger than that of automatic transmission. The p-value is 0.001374, smaller than 0.05. So we should reject null hypothesis. The difference is significant.

### Quantify the MPG difference between automatic and manual transmissions

A linear regression should be figured out to answer this question. We need to find the best fit. Fit mpg ~ am and mpg ~ all parameters and compare their R-squared.

fit\_1<-lm(mpg ~ am, mtcars)  
fit\_all<-lm(mpg ~ ., mtcars)  
summary(fit\_1)

##   
## 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

summary(fit\_all)

##   
## 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

When using am as the single regressor, the R-squared is 0.3598, meaning that the fitting is not very good. Adding all parameters increases the R-squared to 0.869. However, there might be some regressors that are correlated to am. These regressors may cause variance inflation. Use **step** function to find the best fitting. **Step** selects a model by different criteria in a stepwise algorithm. The default criteria is AIC, which find the most likelihood fitting with the least number of regressors.

fit\_s <- step(fit\_all, direction = "backward", trace = FALSE)  
summary(fit\_s)

##   
## 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

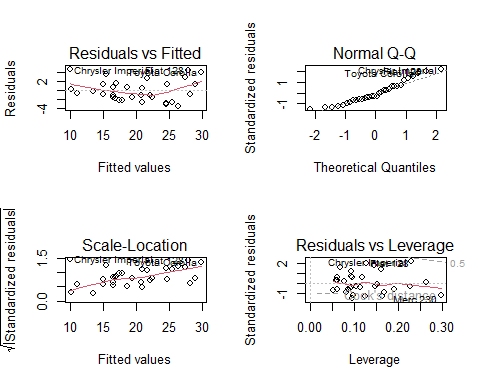
confint(fit\_s, "am")

## 2.5 % 97.5 %  
## am 0.04573031 5.825944

This fitting gives R-squared = 0.8497, which is comparable with fitting all regressors. The slope negative for wt and positive for qsec, in agreement with common understanding. The Pr for wt, qsec and am are smaller than 0.05, meaning that the regressors are necessary. The coefficient of am is 2.9358, meaning that changing transmission from automatic to manual will cause 2.9358 mpg increase for a fixed wt and qsec. The 95% confidence interval is (0.04,5.82).

### Plot fit\_s

par(mfrow=c(2,2))  
plot(fit\_s,which=1)  
plot(fit\_s,which=2)  
plot(fit\_s,which=3)  
plot(fit\_s,which=5)

 Residuals vs Fitted shows residuals around 0. There are some outliers at 15 and 28.  
Normal Q-Q shows almost normal distribution of the residuals.  
The Scale-location is almost horizontal, the spread around the red line does not vary much with the fitted values.  
All the points lies within the cook’s distance, meaning that they are not influential. So we don’t need to remove points.

### Conclusion

1. Manual transmission has a higher mpg than automatic transmission.
2. The mpg is linearly determined by am, wt and qsec. The mpg difference is 2.9358 between manual transmission and automatic transmission when wt and qsec is fixed.The 95% confidence interval is (0.04,5.82)