linear regression practice

main body

let's review the data first

```
library(dplyr)
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
```

we are interested in whether an automatic or manual transmission is better for MPG, so our variable of interest is am Transmission (0 = automatic, 1 = manual), now we are going to perform a regression analysis

```
fit<-lm(mpg~am, data = mtcars)
coef(fit)
```

```
## (Intercept)          am
##   17.147368    7.244939
```

The interpretation of the coefficients is as follows: the average for the automatic transmission is 17.147368 MPG, while for the manual transmission it is higher by 7.244939 MPG, let's do a hypothesis test to disprove the difference of means

$$H_0 : \mu_0 = \mu_1 \qquad vs \qquad H_1 : \mu_0 \neq \mu_1$$

```
automatic <- subset(mtcars,am==0)
manual <- subset(mtcars,am==1)
t.test(automatic$mpg, manual$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  automatic$mpg and manual$mpg
## 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
```

let's consider a significance of $\alpha = 0.05$,since the p-value is 0.001374, then $0.001374 < \alpha$ and the null hypothesis can be rejected. Now we are going to add more variables to our model and see what happens

```
fit2<-lm(mpg~cyl+ disp+hp +drat +   wt + qsec+am, data = mtcars)
coef(fit2)
```

```
## (Intercept)         cyl        disp          hp        drat          wt
## 15.30918956 -0.34192099  0.01458808 -0.02057733  0.81836607 -3.99345102
##        qsec          am
##  0.85996253  2.72022025
```

As we can see the difference in the average estimate when using manual transmission and when using automatic transmission is of 2.72022025 , being the manual transmission the one that produces more MPG, the difference when adding more variables is smaller but still significant, let's compare if there is an improvement between the models

```
anova(fit,fit2)
```

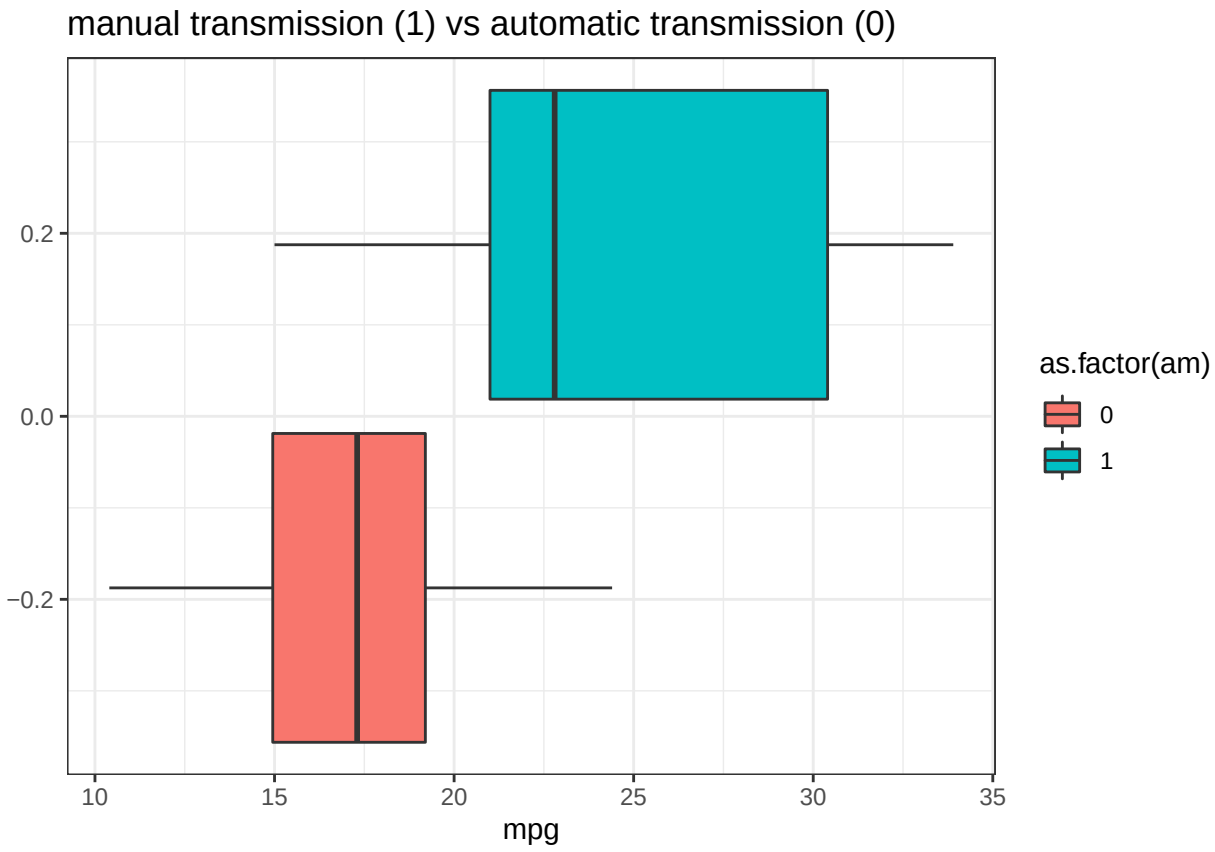
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     24 149.09  6    571.81 15.341 3.648e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis is that the aggregated regressors are not significant, also the null hypothesis is rejected at the 0.001 level of significance, therefore it can be said that fit2 is a "better" model.

appendix

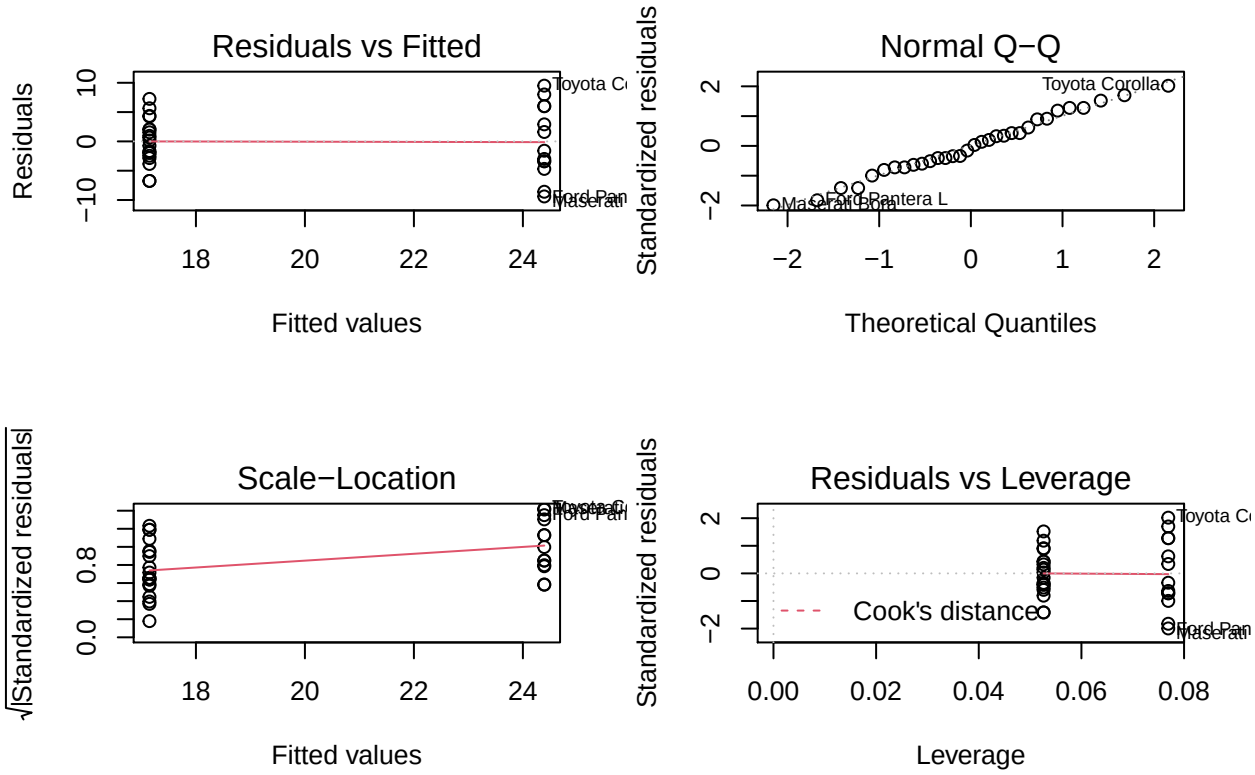
```
library(ggplot2)

g<-ggplot(data=mtcars,aes(x= mpg,fill=as.factor(am)))
g<-g+geom_boxplot()
g<-g+theme_bw()
g<-g+ggtitle("manual transmission (1) vs automatic transmission (0)")
g
```

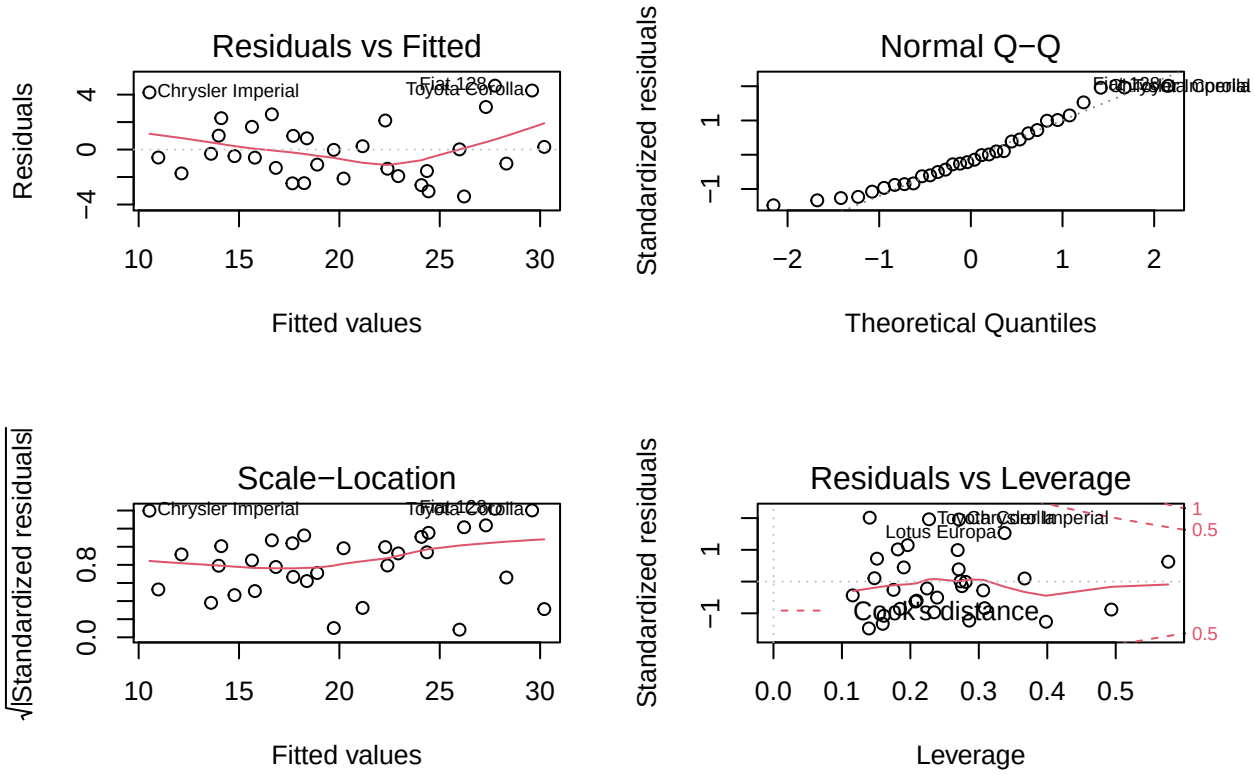


We can see that the whole body of the box for the manual transmission is above the body of the fall of the automatic transmission,now let's check the residuals

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




```
par(mfrow = c(2,2))
plot(fit2)
```



In conclusion we can say that the manual transmission produces more MPG