Week4Assignment

Ken Gu

2017年3月4日

Executive Summary: Background: A magazine (Motor Trend) about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

“Is an automatic or manual transmission better for MPG” “Quantify the MPG difference between automatic and manual transmissions”

Conclusion: By using hypothesis testing and linear regression, it is concluded that there is a yawning gap between automatic and manual transmission for the MPG. In addition, to adjust for other variables - consisting of displacement, number of cylinders, weight, the result shows that these observations have significant impact on the mpg consumption, rather than only transmission.

Multivariable model selection strategy: We use nested likelikhood ratio tests for model selection. In detail: Given a coefficient that I’m interested in, I like to use covariate adjustment and multiple models to probe that effect to evaluate it for robustness and to see what other covariates knock it out or amplify it. In other words, if I have an effect, or absence of an effect, that I’d like to report, I try to first come up with criticisms of that effect and then use models to try to answer those criticisms.

Step 1: load library and data

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.2.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.2.5
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
data(mtcars)
datamtcars <- mutate(mtcars, amvar=as.factor(am))
```

Step 2: Data Cleaning

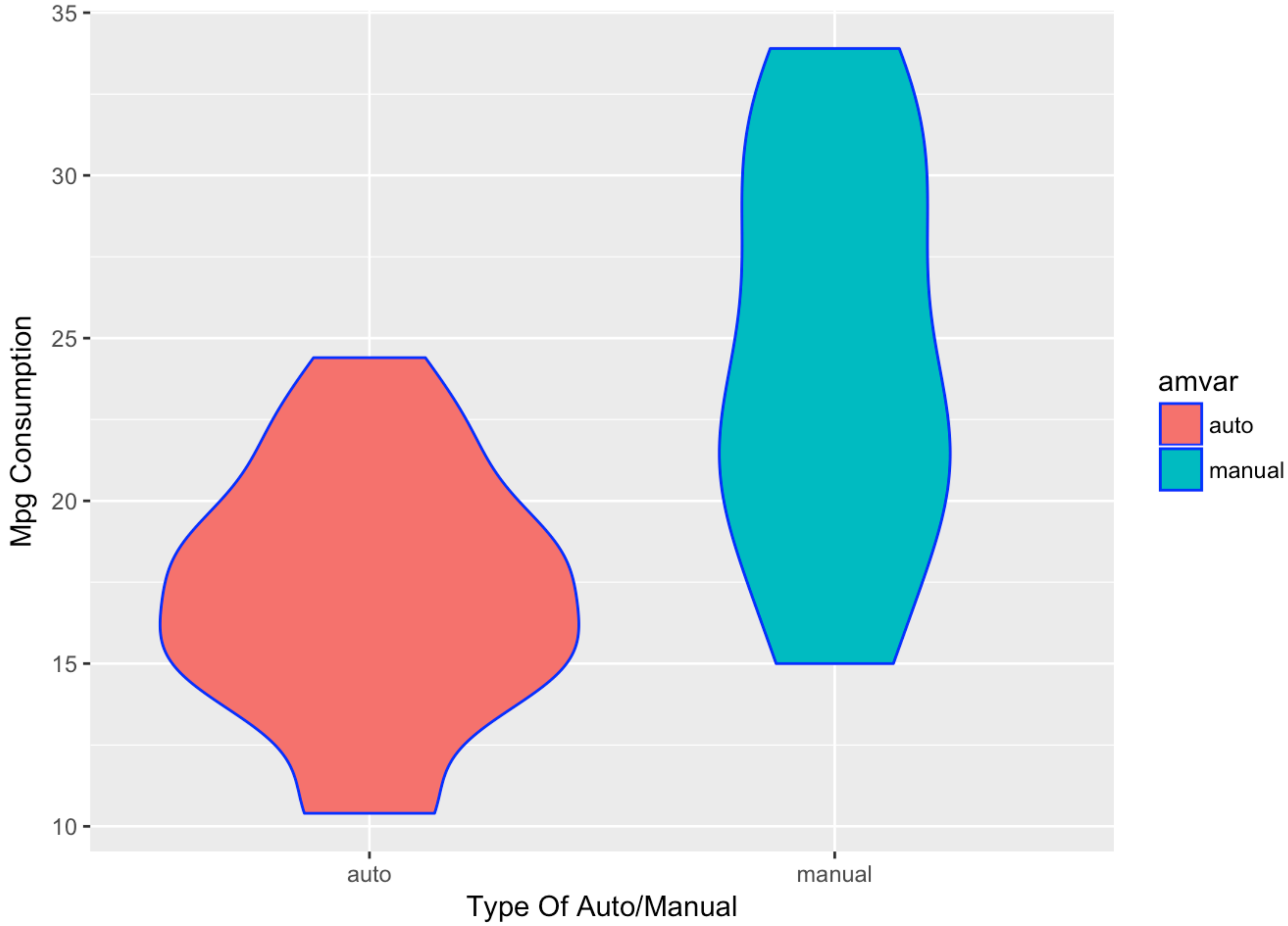
```
# clean the data - add new variable - amvar - as factor
dataautocars <- filter(datamtcars, amvar==0)
datamanucars <- filter(datamtcars, amvar==1)
dataautocars$amvar = "auto"
datamanucars$amvar = "manual"
datamtcars <- data.frame()
datamtcars <- rbind(dataautocars, datamanucars)
datamtcars <- select(datamtcars, -am)
```

Step 3: Use explortory analysis to compare the automatic/manual vs mpg

```
grp1 <- group_by(datamtcars, amvar)
datamean <- summarise (grp1, mean(mpg))
datamean
```

```
## # A tibble: 2 × 2
##    amvar `mean(mpg)`
##    <chr>       <dbl>
## 1   auto    17.14737
## 2 manual    24.39231
```

```
g1 <- ggplot(data=datamtcars, aes(y=mpg, x=amvar, fill=amvar)) + geom_violin(colou
r="blue") + xlab("Type Of Auto/Manual") + ylab("Mpg Consumption")
g1
```



From the above plot, its clearly stated that the auto(red color) transmission has less MPG consumption
Step 4: quantify the difference between automatic cars vs. manual cars for the mpg usage

```
quantifyDiff <- datamean[1,2] - datamean[2,2]
quantifyDiff
```

```
##   mean(mpg)
## 1 -7.244939
```

The quantifiy difference between auto vs manual is 7.24 MPG

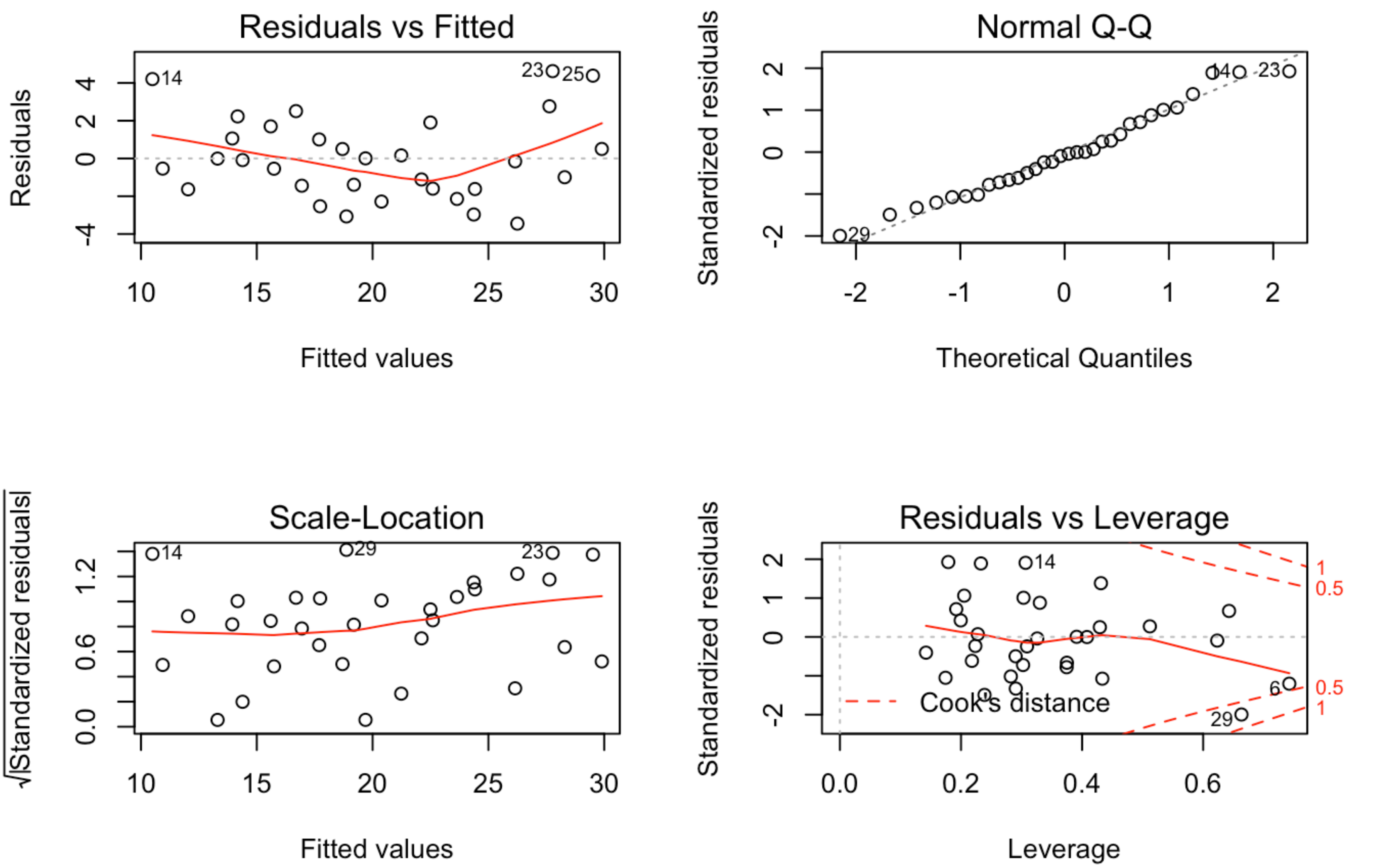
```
fit2 <- lm(mpg~amvar, data=datamtcars)
rSingle <- summary(fit2)

resi <- resid(fit2)
testResult <- t.test(dataautocars$mpg, datamanucars$mpg)
```

From the above t-test, we can see the p-value is only 0.001374, therefore we reject the null hypothesis. The assumption is that all other variables inside the mtcars dataset are same while we compare only auto/manual.

Step 5: Evaluation On MultiVariables and finalize the selection

```
fitall <- lm(mpg ~ ., datamtcars)
vifResult <- vif(fitall)
par(mfrow=c(2,2))
plot(fitall)
```




```
# from the vif result, we pick up the most 4 observations - disp, cyl, wt, hp in a
ddition to the amvar
fit3 <- update(fit2, mpg ~ amvar + disp)
fit4 <- update(fit2, mpg ~ amvar + disp + cyl)
fit5 <- update(fit2, mpg ~ amvar + disp + cyl + wt)
fit6 <- update(fit2, mpg ~ amvar + disp + cyl + wt + hp)

anovavalue <- anova(fit2, fit3, fit4, fit5, fit6)
anovavalue
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ amvar
## Model 2: mpg ~ amvar + disp
## Model 3: mpg ~ amvar + disp + cyl
## Model 4: mpg ~ amvar + disp + cyl + wt
## Model 5: mpg ~ amvar + disp + cyl + wt + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 300.28  1    420.62 67.0427 1.141e-08 ***
## 3     28 252.08  1     48.20  7.6827  0.010165 *
## 4     27 188.43  1     63.66 10.1461  0.003736 **
## 5     26 163.12  1     25.31  4.0336  0.055097 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

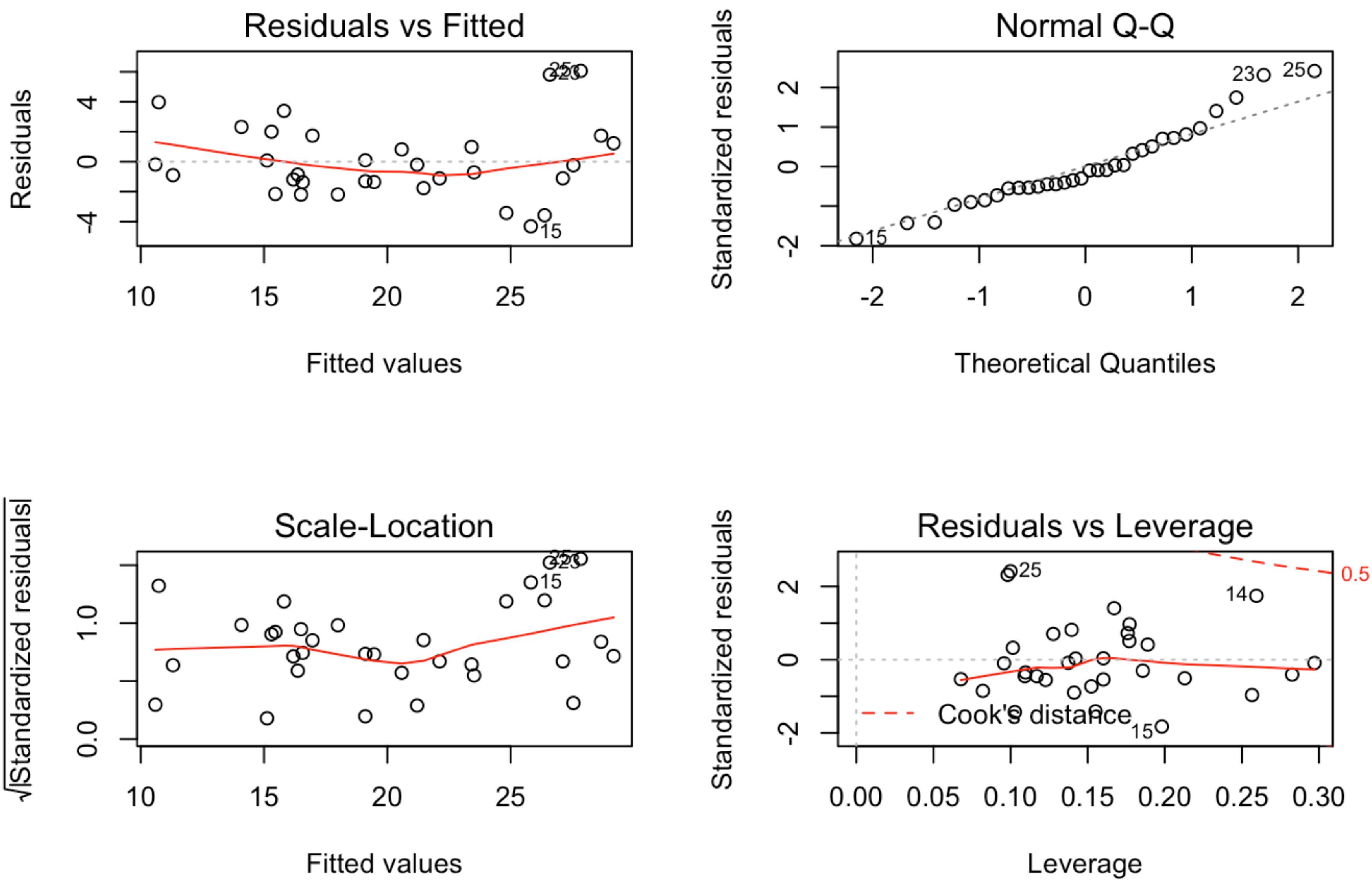
```
newMutliModel <- lm(mpg~amvar + disp + cyl + wt, datamtcars)
summaryMulti <- summary(newMutliModel)
summaryMulti
```

```
##
## Call:
## lm(formula = mpg ~ amvar + disp + cyl + wt, data = datamtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.318 -1.362 -0.479  1.354  6.059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.898313   3.601540  11.356 8.68e-12 ***
## amvarmanual  0.129066   1.321512   0.098  0.92292
## disp         0.007404   0.012081   0.613  0.54509
## cyl         -1.784173   0.618192  -2.886  0.00758 **
## wt          -3.583425   1.186504  -3.020  0.00547 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.642 on 27 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8079
## F-statistic: 33.59 on 4 and 27 DF,  p-value: 4.038e-10
```

From the above code, we use the vif fucntion to help on the model selection, we pick up the most 4 observations - disp, cyl, wt, hp in addition to the amvar From the anova value, we find that disp, cyl, wt tend to have more impacts. As a result, we select this new model

Step 6: Calculate the residuals using plot

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



- From the new model, we see 81% of the variance. The multi variables shows that the question of auto car and manual car is not fully anwsered and the context of displacement, number of cylinders, weight should be taken into total consideration.