regression model

eric

March 29, 2016

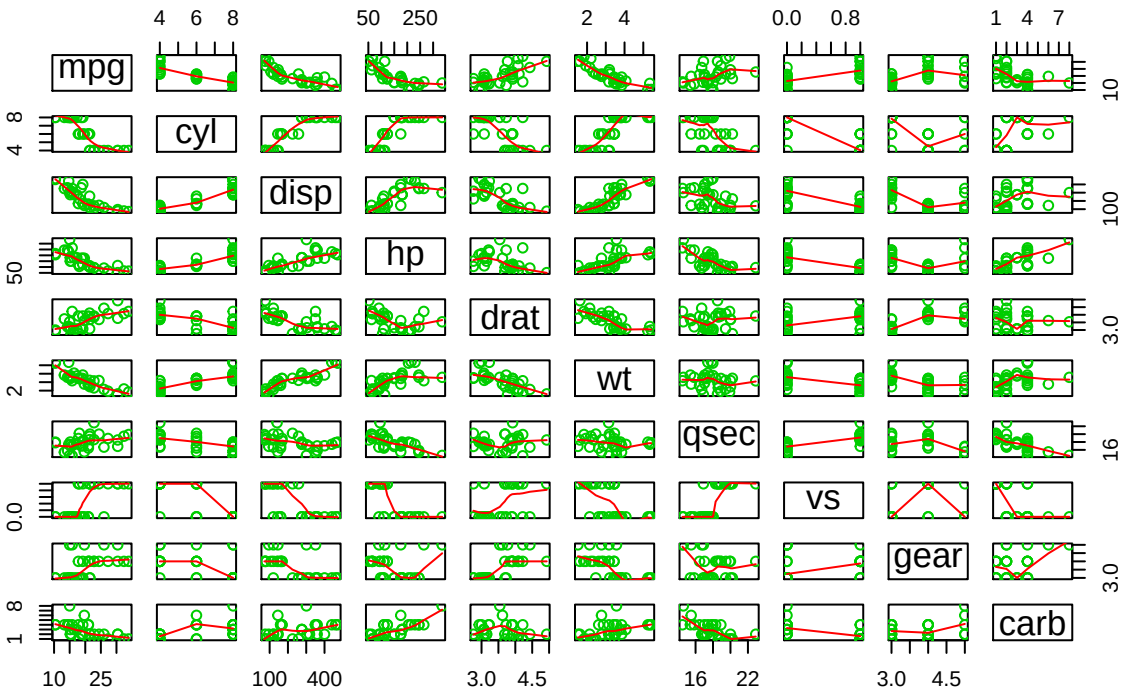
Exploratory

Because in liner regression there is a lot of assumption that depend on Gaussian distribution.So at the first ,we need to see there is highly skewed or not

```
data("mtcars")
library(ggplot2)
mtcars$am=as.factor(mtcars$am)
mtcars$am=ifelse(mtcars$am==1,"manual","automatic")
ggplot(mtcars,aes(x=mpg,fill=am))+geom_density(bw=2)+facet_grid(am~.)+ggtitle("Distribution of am in mt
```

It seems that the variate that we are most interested in is ok.And there is some different between two types. Next:we check the other variates.And we can see that which variates should we alter to factors.And the distribution can also help we select the model.

```
pairs(mtcars[,-9], panel = panel.smooth, main = "variates scatter plot in car", col = 3)
```



Make the models

first see the just one variate that we are interested.

```
lm.fit.just.one<-lm(mpg~factor(am),data=mtcars)
summary(lm.fit.just.one)
```

```
##
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)manual    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
```

Then use the nested models (put in the variable randomly.At the frist,I think the weight may influence most.So I add it) And other way to make model and compare it(some steps it's for convenient to save time).

```r
lm.1<-lm(mpg~factor(am)+factor(cyl)+wt,data=mtcars)
lm.2<-lm(mpg~factor(am)+factor(cyl)+disp+hp+drat+wt+qsec,data=mtcars)
lm.multi<-lm(mpg~factor(cyl)+disp+hp+drat+wt+qsec+factor(vs)+factor(am)+factor(gear)+factor(carb),data=m
best_model <- step(lm.multi, direction = "both")
```

```
## Start:  AIC=76.4
## mpg ~ factor(cyl) + disp + hp + drat + wt + qsec + factor(vs) +
##     factor(am) + factor(gear) + factor(carb)
##
##                Df Sum of Sq    RSS    AIC
## - factor(carb)  5   13.5989 134.00 69.828
## - factor(gear)  2    3.9729 124.38 73.442
## - factor(am)    1    1.1420 121.55 74.705
## - qsec          1    1.2413 121.64 74.732
## - drat          1    1.8208 122.22 74.884
## - factor(cyl)   2   10.9314 131.33 75.184
## - factor(vs)    1    3.6299 124.03 75.354
## <none>                      120.40 76.403
## - disp          1    9.9672 130.37 76.948
## - wt            1   25.5541 145.96 80.562
## - hp            1   25.6715 146.07 80.588
##
## Step:  AIC=69.83
## mpg ~ factor(cyl) + disp + hp + drat + wt + qsec + factor(vs) +
##     factor(am) + factor(gear)
##
##                Df Sum of Sq    RSS    AIC
## - factor(gear)  2    5.0215 139.02 67.005
## - disp          1    0.9934 135.00 68.064
## - drat          1    1.1854 135.19 68.110
## - factor(vs)    1    3.6763 137.68 68.694
## - factor(cyl)   2   12.5642 146.57 68.696
## - qsec          1    5.2634 139.26 69.061
## <none>                      134.00 69.828
## - factor(am)    1   11.9255 145.93 70.556
## - wt            1   19.7963 153.80 72.237
## - hp            1   22.7935 156.79 72.855
## + factor(carb)  5   13.5989 120.40 76.403
##
## Step:  AIC=67
## mpg ~ factor(cyl) + disp + hp + drat + wt + qsec + factor(vs) +
##     factor(am)
##
##                Df Sum of Sq    RSS    AIC
## - drat          1    0.9672 139.99 65.227
## - factor(cyl)   2   10.4247 149.45 65.319
## - disp          1    1.5483 140.57 65.359
## - factor(vs)    1    2.1829 141.21 65.503
## - qsec          1    3.6324 142.66 65.830
## <none>                      139.02 67.005
## - factor(am)    1   16.5665 155.59 68.608
## - hp            1   18.1768 157.20 68.937
## + factor(gear)  2    5.0215 134.00 69.828
```

```
## - wt            1   31.1896 170.21 71.482
## + factor(carb)  5   14.6475 124.38 73.442
##
## Step:  AIC=65.23
## mpg ~ factor(cyl) + disp + hp + wt + qsec + factor(vs) + factor(am)
##
##                Df Sum of Sq    RSS    AIC
## - disp          1    1.2474 141.24 63.511
## - factor(vs)    1    2.3403 142.33 63.757
## - factor(cyl)   2   12.3267 152.32 63.927
## - qsec          1    3.1000 143.09 63.928
## <none>                      139.99 65.227
## + drat          1    0.9672 139.02 67.005
## - hp            1   17.7382 157.73 67.044
## - factor(am)    1   19.4660 159.46 67.393
## + factor(gear)  2    4.8033 135.19 68.110
## - wt            1   30.7151 170.71 69.574
## + factor(carb)  5   13.0509 126.94 72.095
##
## Step:  AIC=63.51
## mpg ~ factor(cyl) + hp + wt + qsec + factor(vs) + factor(am)
##
##                Df Sum of Sq    RSS    AIC
## - qsec          1     2.442 143.68 62.059
## - factor(vs)    1     2.744 143.98 62.126
## - factor(cyl)   2    18.580 159.82 63.466
## <none>                      141.24 63.511
## + disp          1     1.247 139.99 65.227
## + drat          1     0.666 140.57 65.359
## - hp            1    18.184 159.42 65.386
## - factor(am)    1    18.885 160.12 65.527
## + factor(gear)  2     4.684 136.55 66.431
## - wt            1    39.645 180.88 69.428
## + factor(carb)  5     2.331 138.91 72.978
##
## Step:  AIC=62.06
## mpg ~ factor(cyl) + hp + wt + factor(vs) + factor(am)
##
##                Df Sum of Sq    RSS    AIC
## - factor(vs)    1     7.346 151.03 61.655
## <none>                      143.68 62.059
## - factor(cyl)   2    25.284 168.96 63.246
## + qsec          1     2.442 141.24 63.511
## - factor(am)    1    16.443 160.12 63.527
## + disp          1     0.589 143.09 63.928
## + drat          1     0.330 143.35 63.986
## + factor(gear)  2     3.437 140.24 65.284
## - hp            1    36.344 180.02 67.275
## - wt            1    41.088 184.77 68.108
## + factor(carb)  5     3.480 140.20 71.275
##
## Step:  AIC=61.65
## mpg ~ factor(cyl) + hp + wt + factor(am)
##
```

```
##                Df Sum of Sq    RSS    AIC
## <none>                      151.03 61.655
## - factor(am)    1     9.752 160.78 61.657
## + factor(vs)    1     7.346 143.68 62.059
## + qsec          1     7.044 143.98 62.126
## - factor(cyl)   2    29.265 180.29 63.323
## + disp          1     0.617 150.41 63.524
## + drat          1     0.220 150.81 63.608
## + factor(gear)  2     1.361 149.66 65.365
## - hp            1    31.943 182.97 65.794
## - wt            1    46.173 197.20 68.191
## + factor(carb)  5     5.633 145.39 70.438
```

```
anova(lm.fit.just.one,lm.1,lm.2,best_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + factor(cyl) + wt
## Model 3: mpg ~ factor(am) + factor(cyl) + disp + hp + drat + wt + qsec
## Model 4: mpg ~ factor(cyl) + hp + wt + factor(am)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     27 182.97  3    537.93 29.2064 5.085e-08 ***
## 3     23 141.21  4     41.76  1.7006    0.1842
## 4     26 151.03 -3     -9.82  0.5332    0.6641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It seems the model lm.1 is most significant ,and that it means when we added wt changes most!!But after we check Rss is lm.2 best. Check lm.2 solely.

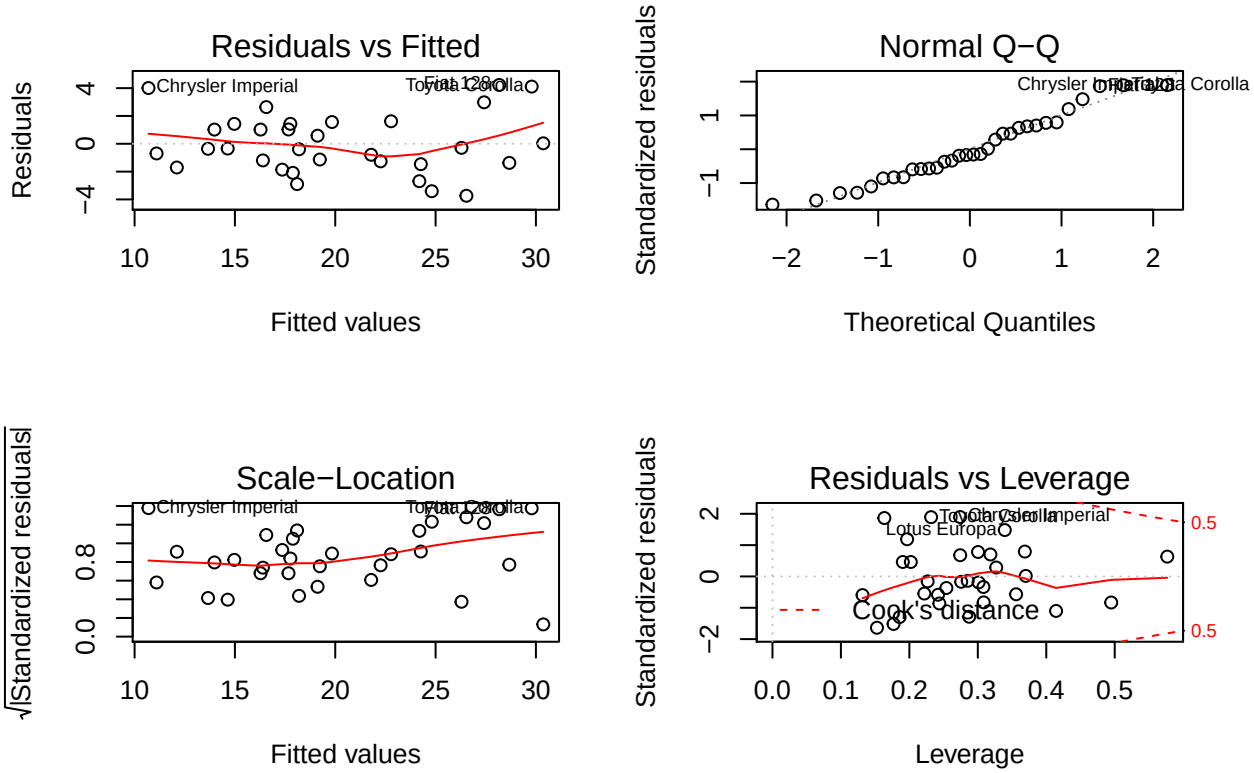
```
summary(lm.2)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + factor(cyl) + disp + hp + drat +
##     wt + qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7362 -1.3973 -0.3513  1.4311  4.2267
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      17.475854  14.027771   1.246   0.2254
## factor(am)manual  2.748596   1.777155   1.547   0.1356
## factor(cyl)6     -1.697049   1.913301  -0.887   0.3843
## factor(cyl)8     -0.522560   3.464300  -0.151   0.8814
## disp              0.007572   0.013236   0.572   0.5728
## hp               -0.025353   0.015699  -1.615   0.1200
## drat              0.633293   1.479625   0.428   0.6726
```

```
## wt              -3.426768   1.323748  -2.589   0.0164 *
## qsec             0.718097   0.596598   1.204   0.2410
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.478 on 23 degrees of freedom
## Multiple R-squared:  0.8746, Adjusted R-squared:  0.831
## F-statistic: 20.05 on 8 and 23 DF,  p-value: 1.206e-08
```

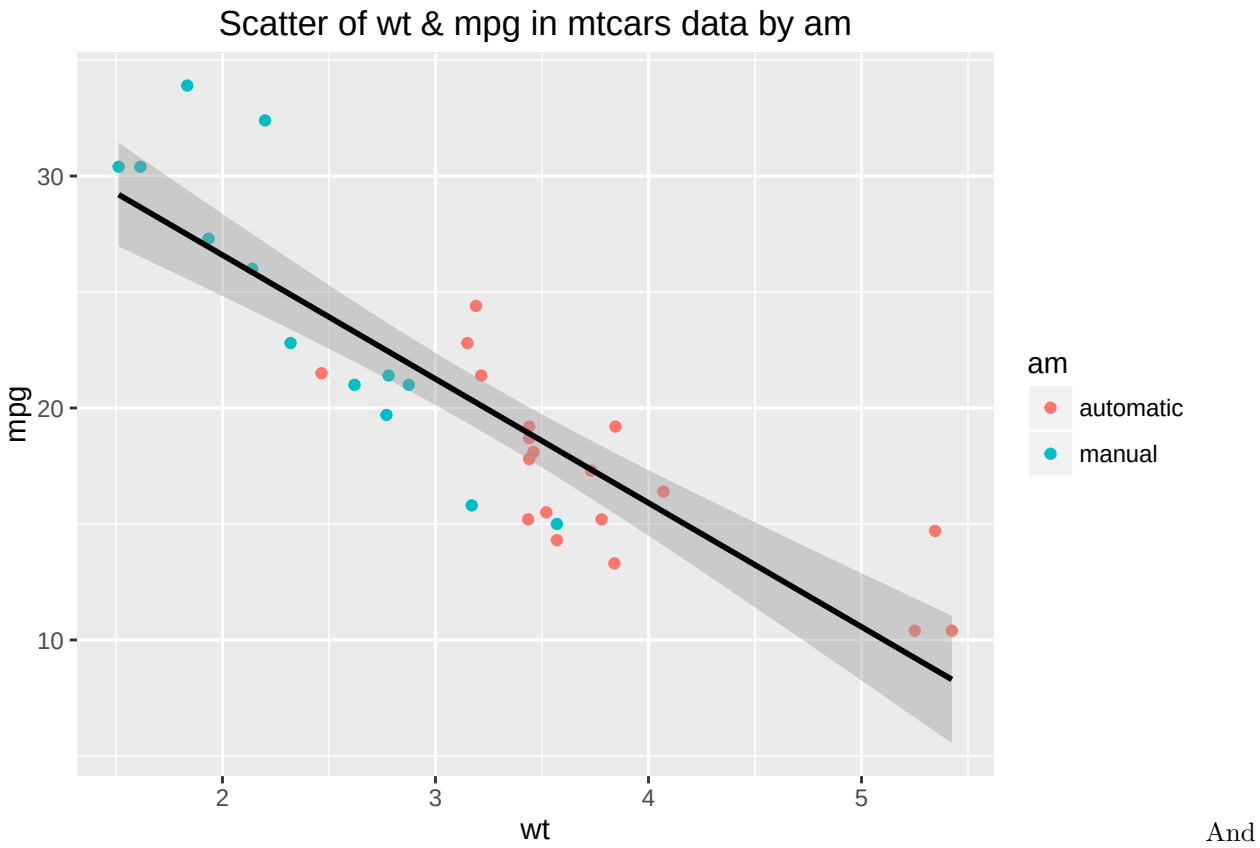
The model is significant,and it also verify that the most significant variate is wt. ##Residual plot

```
par(mfrow = c(2, 2)); plot(lm.2)
```



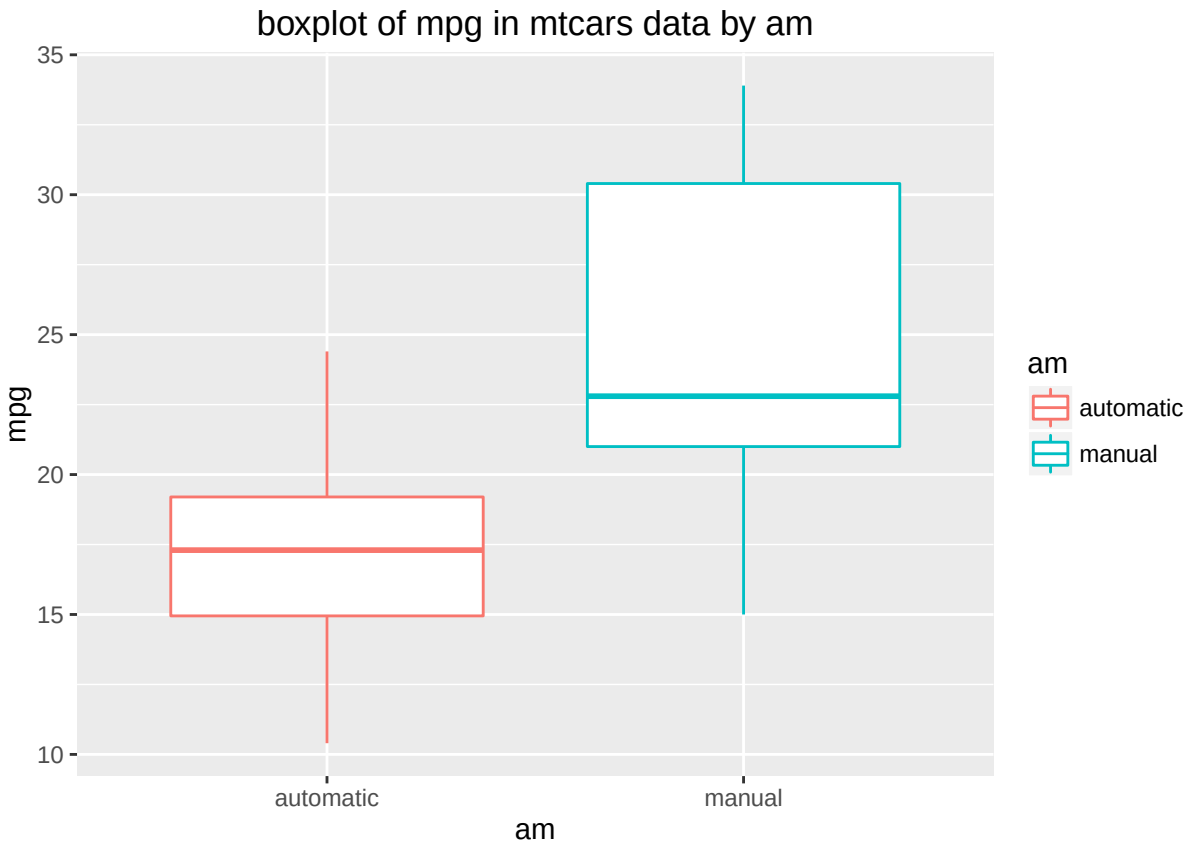
We can see the model is good.Even if there is some points that highly influence data and make it out of normal.But it doesn't harm a lot, just a littlebit. ##Conclusion and Appendix Then we chek most significant variate and the things that we are interested.

```
ggplot(mtcars,aes(x=wt,y=mpg,color=am))+geom_point()+ggtitle("Scatter of wt & mpg in mtcars data by am")
```

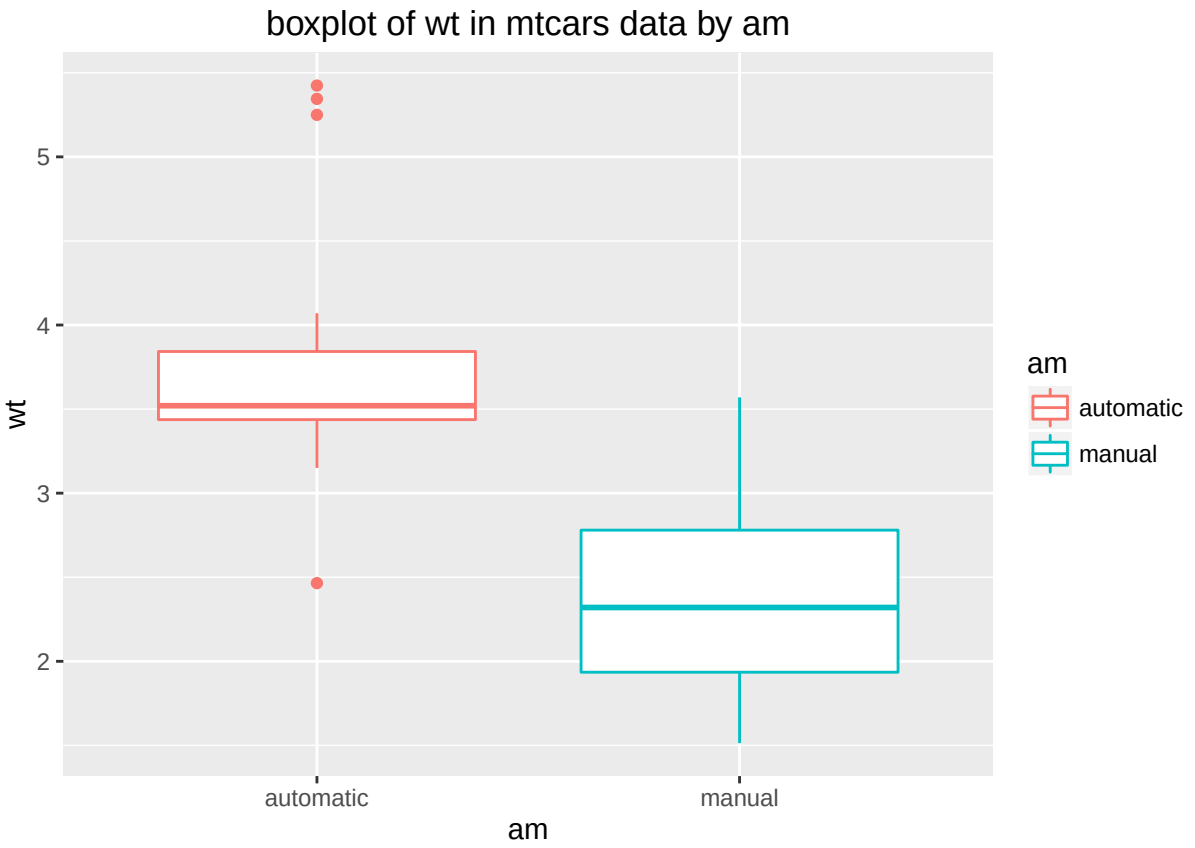


And we also need to see more detail about the different between two types.

```
ggplot(mtcars,aes(x=am,y=mpg,color=am))+geom_boxplot()+ggtitle("boxplot of mpg in mtcars data by am")
```




```
ggplot(mtcars,aes(x=am,y=wt,color=am))+geom_boxplot()+ggtitle("boxplot of wt in mtcars data by am")
```



It seems that can answer the question which we are interested in.So two types(manual,“automatic) are different in mpg,and also in weight.There is not going to discuss deeper problem.But we can guess maybe the mechanical different in am make both variate different.