Regression_models_assignment

Hang_YU

February 3, 2017

Executive summary

In this report, the fuel efficiency between automatic transmission and manual transmission are analyzed. As a conclusion, the fuel efficiency is impacted by multiple predictors and transmission type is not a significant factor for the change of mpg.

Exploratory data analysis

Conduct exploratory data analysis using pair plot for a global view of the mtcars dataset.

```
ggpairs(mtcars,upper=list(continuous="smooth"),lower=list(continuous="cor"))
```

From the pair plot (see attachment), there are some relationships between mpg and those 10 predictors. Roughly, the current plot shows that automatic transmission (0) results in fewer mpg than manual gear (1).

Hypothesis testing

The conclusion is validated through the single-sided student test with alternative hypothesis that automatic transmission has fewer mpg than manual cars:p value of the student test is smaller than 0.05 so we choose the alternative hypothesis.

```
t.test(subset(mtcars,am==0)$mpg,subset(mtcars,am==1)$mpg,alternative = "less")$p.value
```

```
## [1] 0.0006868192
```

Then we quantify the mpg difference using a single-variable linear model:

```
model1 <- lm(mpg~am,data=mtcars);summary(model1)$coeff
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

From the coefficients, the mean of mpg increases by 7.244939 units if the transmission changes from auto to manual. However it is incorrect to conclude the relationship yet. We need to check if this relationship is the result of interference from residuals of other predictors. In other words, to eliminite the bias due to omitting significant variables.

Model selection

At this point, we select appropriate predictors to build the linear regression for mpg prediction. This is done by using anova for nested models. We focus on the p values of anova test to determine the variable significance.

```
model2 <- update(model1,mpg~am+cyl,data=mtcars)
model3 <- update(model1,mpg~am+cyl+disp,data=mtcars)
model4 <- update(model1,mpg~am+cyl+disp+hp,data=mtcars)
model5 <- update(model1,mpg~am+cyl+disp+hp+drat,data=mtcars)
model6 <- update(model1,mpg~am+cyl+disp+hp+drat+wt,data=mtcars)
model7 <- update(model1,mpg~am+cyl+disp+hp+drat+wt+qsec,data=mtcars)
model8 <- update(model1,mpg~am+cyl+disp+hp+drat+wt+qsec+vs,data=mtcars)
model9 <- update(model1,mpg~am+cyl+disp+hp+drat+wt+qsec+vs+gear,data=mtcars)
model10 <- update(model1,mpg~am+cyl+disp+hp+drat+wt+qsec+vs+gear+carb,data=mtcars)
anova(model1,model2,model3,model4,model5,model6,model7,model8,model9,model10)$Pr
```

```
##  [1]           NA 8.231226e-08 1.124136e-01 3.493497e-02 6.112088e-01
##  [6] 1.275295e-02 1.826036e-01 8.621415e-01 7.136533e-01 8.121787e-01
```

According to the p values of anova test, am,cyl,hp, and wt and selected for the regression model due to their high variability contribution. Next, the variance inflation of the model with those four predictors is checked:

```
model <- lm(mpg~am+cyl+hp+wt,data=mtcars);vif(model)
```

```
##       am      cyl       hp       wt
## 2.546159 5.333685 4.310029 3.988305
```

The variable cyl has a high variance inflation factor of 5.333686 (generally the threshold is 5) so we remove this variable since it is correlated with at least one of the other variables (maybe hp and wt). After doing so the variance inflation becomes acceptable.

```
model <- lm(mpg~am+hp+wt,data=mtcars);vif(model)
```

```
##       am       hp       wt
## 2.271082 2.088124 3.774838
```

Model diagnosis

Next, we check the coefficient of the final model:

```
summary(model)$coef
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## am           2.08371013 1.376420152  1.513862 1.412682e-01
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
```

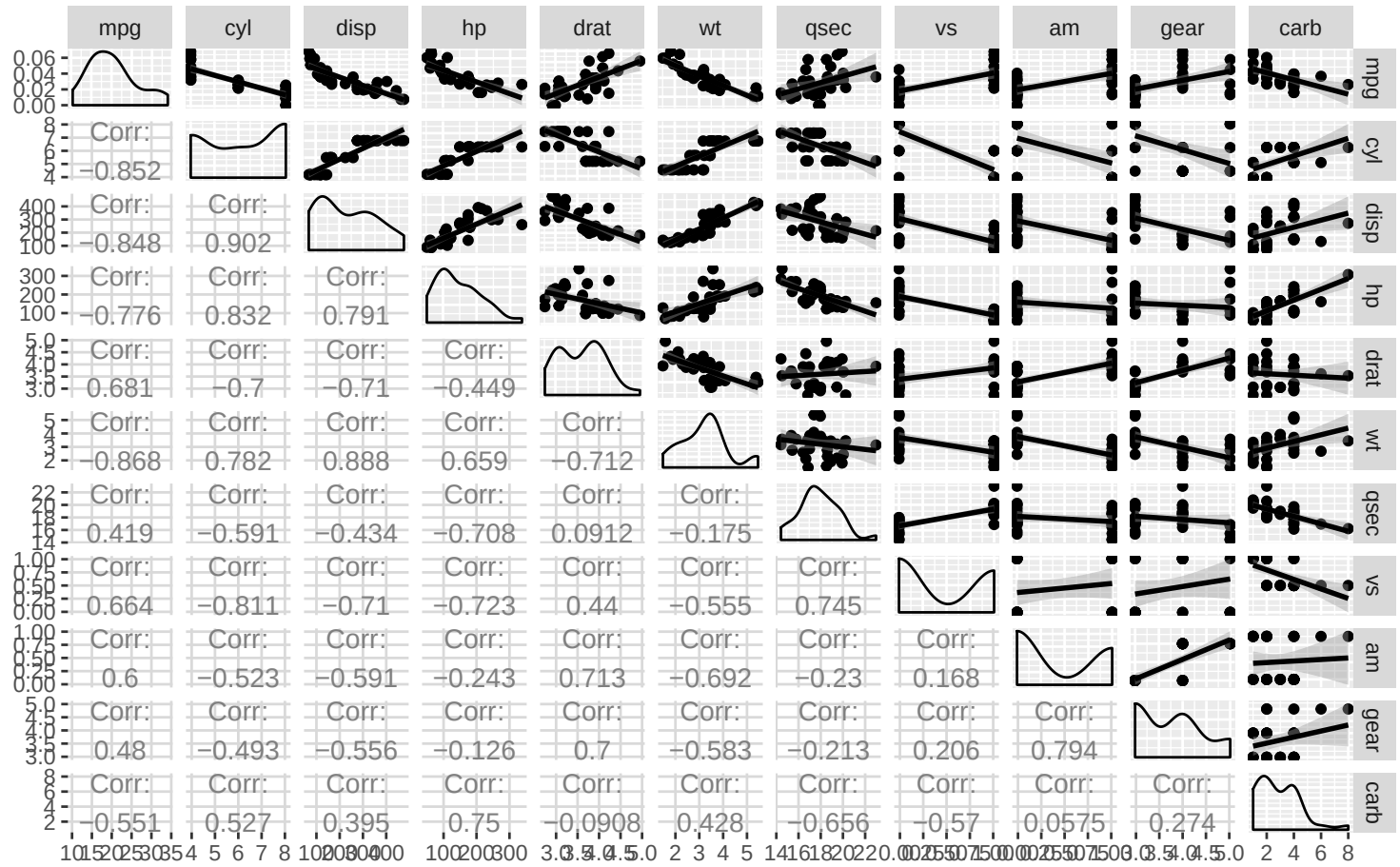
In comparison with model1, the coefficient of transmission type is dramatically dicreased. According to its high p value, transmission type is not a significant factor for mpg therefore. In other words, we cannot conclude that automatic transmission has lower mpg than manual transmission. At last, we plot the final model (see attachment) to check the residuals and high-influence points, which shows that our model is basically reliable however with non-uniform residuals for fitted values caused by omitted variables.

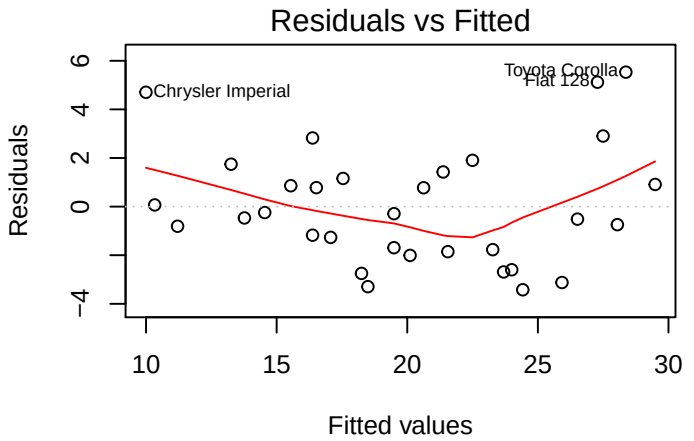
```
par(mfrow=c(4,1));plot(model)
```

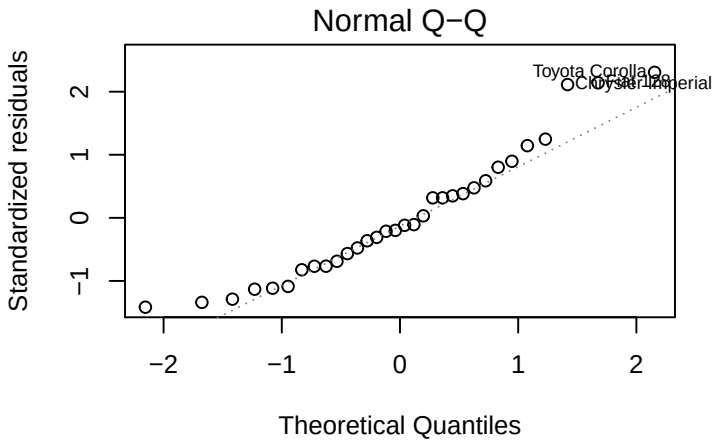
Conclusions

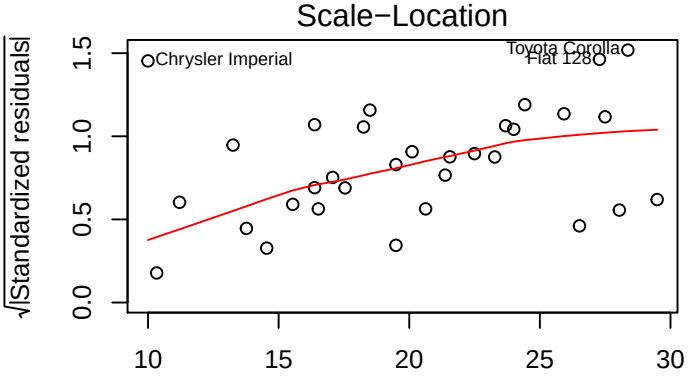
Basically we can conclude that manual transmission has slightly higher but not significant mpg (2.08371013) than automatic transmission. Transmission type is not a siginificant variable for mpg values. Mpg values are influenced by multiple variables such as horse power and weight or even some other potential variables not shown in the dataset.

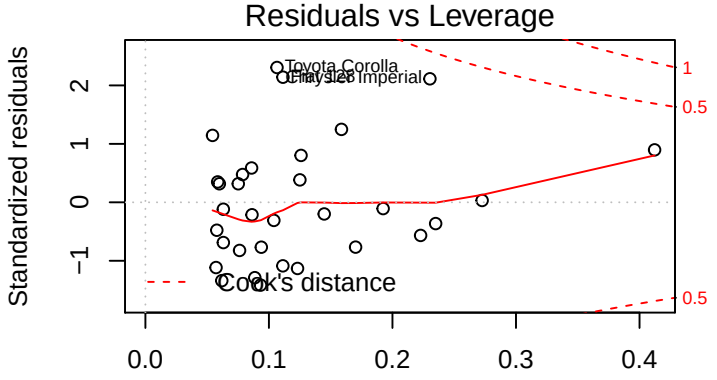
Appendix


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb
Corr: −0.852
Corr: −0.848
Corr: 0.902
Corr: −0.776
Corr: 0.832
Corr: 0.791
Corr: 0.681
Corr: −0.7
Corr: −0.71
Corr: −0.449
Corr: −0.868
Corr: 0.782
Corr: 0.888
Corr: 0.659
Corr: −0.712
Corr: 0.419
Corr: −0.591
Corr: −0.434
Corr: −0.708
Corr: 0.0912
Corr: −0.175
Corr: 0.664
Corr: −0.811
Corr: −0.71
Corr: −0.723
Corr: 0.44
Corr: −0.555
Corr: 0.745
Corr: 0.6
Corr: −0.523
Corr: −0.591
Corr: −0.243
Corr: 0.713
Corr: −0.692
Corr: −0.23
Corr: 0.168
Corr: 0.48
Corr: −0.493
Corr: −0.556
Corr: −0.126
Corr: 0.7
Corr: −0.583
Corr: −0.213
Corr: 0.206
Corr: 0.794
Corr: −0.551
Corr: 0.527
Corr: 0.395
Corr: 0.75
Corr: −0.0908
Corr: 0.428
Corr: −0.656
Corr: −0.57
Corr: 0.0575
Corr: 0.274


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128


Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla


Scale−Location
√|Standardized residuals|
Chrysler Imperial
Toyota Corolla
Fiat 128


Residuals vs Leverage
Standardized residuals
Toyota Corolla
Cook's distance