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)
```

```
## Analysis of Variance Table
##
## Model  1: mpg ~ am
## Model  2: mpg ~ am + cyl
## Model  3: mpg ~ am + cyl + disp
## Model  4: mpg ~ am + cyl + disp + hp
## Model  5: mpg ~ am + cyl + disp + hp + drat
## Model  6: mpg ~ am + cyl + disp + hp + drat + wt
## Model  7: mpg ~ am + cyl + disp + hp + drat + wt + qsec
## Model  8: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs
## Model  9: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear
## Model 10: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear + carb
##    Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1      30 720.90
## 2      29 271.36  1    449.53 64.0039 8.231e-08 ***
## 3      28 252.08  1     19.28  2.7452   0.11241
## 4      27 216.37  1     35.71  5.0849   0.03493 *
## 5      26 214.50  1      1.87  0.2663   0.61121
## 6      25 162.43  1     52.06  7.4127   0.01275 *
## 7      24 149.09  1     13.34  1.8999   0.18260
## 8      23 148.87  1      0.22  0.0309   0.86214
## 9      22 147.90  1      0.97  0.1384   0.71365
## 10     21 147.49  1      0.41  0.0579   0.81218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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
```

Also, we do step-wise regression for feature selection:

```
step(lm(mpg~am,mtcars),direction = "forward",scope=list(lower=formula(lm(mpg~1,mtcars)),upper=formula(lm(mpg~.,m
```

```
## Start:  AIC=103.67
## mpg ~ am
##
##        Df Sum of Sq    RSS     AIC
## + hp    1    475.46 245.44  71.194
## + cyl   1    449.53 271.36  74.407
## + wt    1    442.58 278.32  75.217
## + disp  1    420.62 300.28  77.647
## + carb  1    387.22 333.68  81.022
## + qsec  1    368.26 352.63  82.790
## + vs    1    367.41 353.49  82.867
## + drat  1    147.26 573.64  98.361
## <none>               720.90 103.672
## + gear   1      0.05 720.85 105.670
##
## Step:  AIC=71.19
## mpg ~ am + hp
##
##         Df Sum of Sq    RSS    AIC
## + wt     1    65.148 180.29 63.323
## + vs     1    26.560 218.88 69.529
## + cyl    1    24.886 220.55 69.773
## + disp   1    19.336 226.10 70.568
## + carb   1    16.264 229.18 71.000
## <none>              245.44 71.194
## + drat   1    13.089 232.35 71.440
## + gear   1     7.458 237.98 72.207
## + qsec   1     6.879 238.56 72.284
##
## Step:  AIC=63.32
## mpg ~ am + hp + wt
##
##         Df Sum of Sq    RSS    AIC
## + qsec   1   20.2246 160.07 61.515
## + vs     1   11.3276 168.96 63.246
## <none>              180.29 63.323
## + cyl    1   10.2933 170.00 63.442
## + carb   1    5.3658 174.93 64.356
## + drat   1    3.3262 176.97 64.727
## + gear   1    0.9507 179.34 65.154
## + disp   1    0.3835 179.91 65.255
##
## Step:  AIC=61.52
## mpg ~ am + hp + wt + qsec
##
##         Df Sum of Sq    RSS    AIC
## <none>              160.07 61.515
## + disp   1    6.6287 153.44 62.162
## + carb   1    3.2272 156.84 62.864
## + drat   1    1.4278 158.64 63.229
## + cyl    1    0.2490 159.82 63.465
## + vs     1    0.2486 159.82 63.466
## + gear   1    0.1711 159.90 63.481
```

```
##
## Call:
## lm(formula = mpg ~ am + hp + wt + qsec, data = mtcars)
##
## Coefficients:
## (Intercept)           am           hp           wt         qsec
##    17.44019      2.92550     -0.01765     -3.23810      0.81060
```

```
step(lm(mpg~.,mtcars),direction = "backward")
```

```
## Start:  AIC=70.9
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##         Df Sum of Sq    RSS    AIC
## - cyl    1    0.0799 147.57 68.915
## - vs     1    0.1601 147.66 68.932
## - carb   1    0.4067 147.90 68.986
## - gear   1    1.3531 148.85 69.190
## - drat   1    1.6270 149.12 69.249
## - disp  1    3.9167 151.41 69.736
## - hp    1    6.8399 154.33 70.348
## - qsec  1    8.8641 156.36 70.765
## <none>              147.49 70.898
## - am    1   10.5467 158.04 71.108
## - wt    1   27.0144 174.51 74.280
##
## Step:  AIC=68.92
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - vs    1    0.2685 147.84 66.973
## - carb  1    0.5201 148.09 67.028
## - gear  1    1.8211 149.40 67.308
## - drat  1    1.9826 149.56 67.342
## - disp  1    3.9009 151.47 67.750
## - hp    1    7.3632 154.94 68.473
## <none>              147.57 68.915
## - qsec  1   10.0933 157.67 69.032
## - am    1   11.8359 159.41 69.384
## - wt    1   27.0280 174.60 72.297
##
## Step:  AIC=66.97
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  1    0.6855 148.53 65.121
## - gear  1    2.1437 149.99 65.434
## - drat  1    2.2139 150.06 65.449
## - disp  1    3.6467 151.49 65.753
## - hp    1    7.1060 154.95 66.475
## <none>              147.84 66.973
## - am    1   11.5694 159.41 67.384
## - qsec  1   15.6830 163.53 68.200
## - wt    1   27.3799 175.22 70.410
##
## Step:  AIC=65.12
## mpg ~ disp + hp + drat + wt + qsec + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  1     1.565 150.09 63.457
## - drat  1     1.932 150.46 63.535
## <none>              148.53 65.121
## - disp  1    10.110 158.64 65.229
## - am    1    12.323 160.85 65.672
## - hp    1    14.826 163.35 66.166
## - qsec  1    26.408 174.94 68.358
## - wt    1    69.127 217.66 75.350
##
## Step:  AIC=63.46
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1     3.345 153.44 62.162
## - disp  1     8.545 158.64 63.229
## <none>              150.09 63.457
## - hp    1    13.285 163.38 64.171
## - am    1    20.036 170.13 65.466
## - qsec  1    25.574 175.67 66.491
## - wt    1    67.572 217.66 73.351
##
```

```
## Step:  AIC=62.16
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 61.515
## <none>              153.44 62.162
## - hp    1    12.572 166.01 62.682
## - qsec  1    26.470 179.91 65.255
## - am    1    32.198 185.63 66.258
## - wt    1    69.043 222.48 72.051
##
## Step:  AIC=61.52
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 61.307
## <none>              160.07 61.515
## - qsec  1    20.225 180.29 63.323
## - am    1    25.993 186.06 64.331
## - wt    1    78.494 238.56 72.284
##
## Step:  AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 61.307
## - am    1    26.178 195.46 63.908
## - qsec  1   109.034 278.32 75.217
## - wt    1   183.347 352.63 82.790

##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Coefficients:
## (Intercept)           wt         qsec           am
##       9.618       -3.917        1.226        2.936
```

```
step(lm(mpg~.,mtcars),direction = "both")
```

```
## Start:  AIC=70.9
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - cyl   1    0.0799 147.57 68.915
## - vs    1    0.1601 147.66 68.932
## - carb  1    0.4067 147.90 68.986
## - gear  1    1.3531 148.85 69.190
## - drat  1    1.6270 149.12 69.249
## - disp  1    3.9167 151.41 69.736
## - hp    1    6.8399 154.33 70.348
## - qsec  1    8.8641 156.36 70.765
## <none>              147.49 70.898
## - am    1   10.5467 158.04 71.108
## - wt    1   27.0144 174.51 74.280
##
## Step:  AIC=68.92
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - vs    1    0.2685 147.84 66.973
## - carb  1    0.5201 148.09 67.028
## - gear  1    1.8211 149.40 67.308
## - drat  1    1.9826 149.56 67.342
## - disp  1    3.9009 151.47 67.750
## - hp    1    7.3632 154.94 68.473
## <none>              147.57 68.915
## - qsec  1   10.0933 157.67 69.032
## - am    1   11.8359 159.41 69.384
## + cyl   1    0.0799 147.49 70.898
## - wt    1   27.0280 174.60 72.297
##
## Step:  AIC=66.97
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  1    0.6855 148.53 65.121
## - gear  1    2.1437 149.99 65.434
## - drat  1    2.2139 150.06 65.449
## - disp  1    3.6467 151.49 65.753
## - hp    1    7.1060 154.95 66.475
## <none>              147.84 66.973
## - am    1   11.5694 159.41 67.384
## - qsec  1   15.6830 163.53 68.200
## + vs    1    0.2685 147.57 68.915
## + cyl   1    0.1883 147.66 68.932
## - wt    1   27.3799 175.22 70.410
##
## Step:  AIC=65.12
## mpg ~ disp + hp + drat + wt + qsec + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  1     1.565 150.09 63.457
## - drat  1     1.932 150.46 63.535
## <none>              148.53 65.121
## - disp  1    10.110 158.64 65.229
## - am    1    12.323 160.85 65.672
## - hp    1    14.826 163.35 66.166
## + carb  1     0.685 147.84 66.973
## + vs    1     0.434 148.09 67.028
## + cyl   1     0.414 148.11 67.032
## - qsec  1    26.408 174.94 68.358
## - wt    1    69.127 217.66 75.350
##
## Step:  AIC=63.46
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1     3.345 153.44 62.162
## - disp  1     8.545 158.64 63.229
## <none>              150.09 63.457
## - hp    1    13.285 163.38 64.171
## + gear  1     1.565 148.53 65.121
## + cyl   1     1.003 149.09 65.242
## + vs    1     0.645 149.45 65.319
## + carb  1     0.107 149.99 65.434
## - am    1    20.036 170.13 65.466
## - qsec  1    25.574 175.67 66.491
## - wt    1    67.572 217.66 73.351
##
## Step:  AIC=62.16
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 61.515
## <none>              153.44 62.162
## - hp    1    12.572 166.01 62.682
## + drat  1     3.345 150.09 63.457
## + gear  1     2.977 150.46 63.535
## + cyl   1     2.447 150.99 63.648
## + vs    1     1.121 152.32 63.927
## + carb  1     0.011 153.43 64.160
## - qsec  1    26.470 179.91 65.255
## - am    1    32.198 185.63 66.258
## - wt    1    69.043 222.48 72.051
##
## Step:  AIC=61.52
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 61.307
## <none>              160.07 61.515
## + disp  1     6.629 153.44 62.162
## + carb  1     3.227 156.84 62.864
## + drat  1     1.428 158.64 63.229
## - qsec  1    20.225 180.29 63.323
## + cyl   1     0.249 159.82 63.465
## + vs    1     0.249 159.82 63.466
## + gear  1     0.171 159.90 63.481
## - am    1    25.993 186.06 64.331
## - wt    1    78.494 238.56 72.284
##
## Step:  AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 61.307
## + hp    1     9.219 160.07 61.515
## + carb  1     8.036 161.25 61.751
## + disp  1     3.276 166.01 62.682
## + cyl   1     1.501 167.78 63.022
## + drat  1     1.400 167.89 63.042
## + gear  1     0.123 169.16 63.284
## + vs    1     0.000 169.29 63.307
## - am    1    26.178 195.46 63.908
## - qsec  1   109.034 278.32 75.217
## - wt    1   183.347 352.63 82.790
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Coefficients:
## (Intercept)           wt         qsec           am
##       9.618       -3.917        1.226        2.936
```

The final model is agreed to be mpg~am+hp+wt+qsec.

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
```

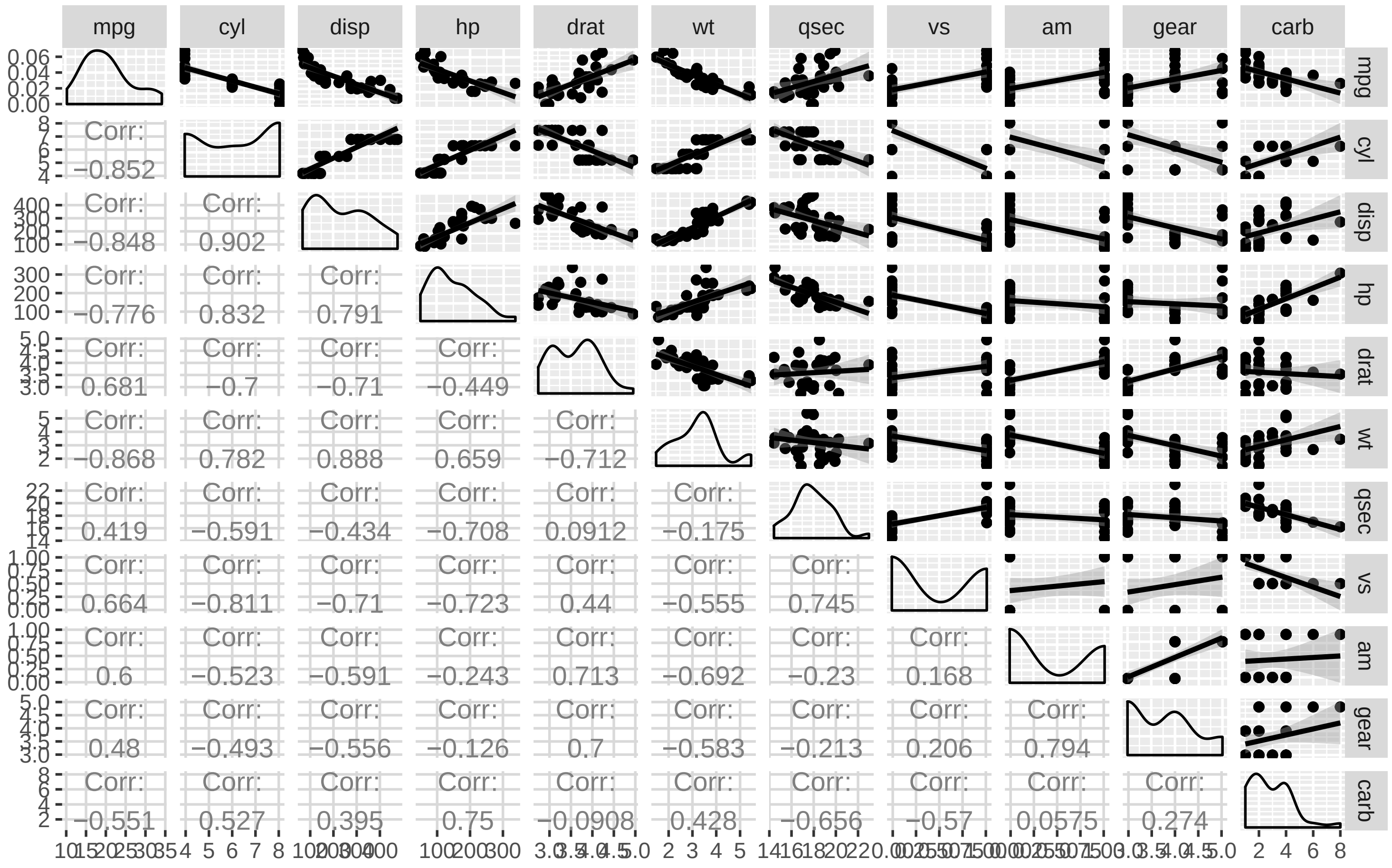
In comparison with model1, the coefficient of transmission type is dramatically dicreased. According to its high p value, transmission type is not a significant factor for mpg therefore. In other words, we cannot conclude that automatic transmission has lower mpg than manual transmission. At last, we 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)
```

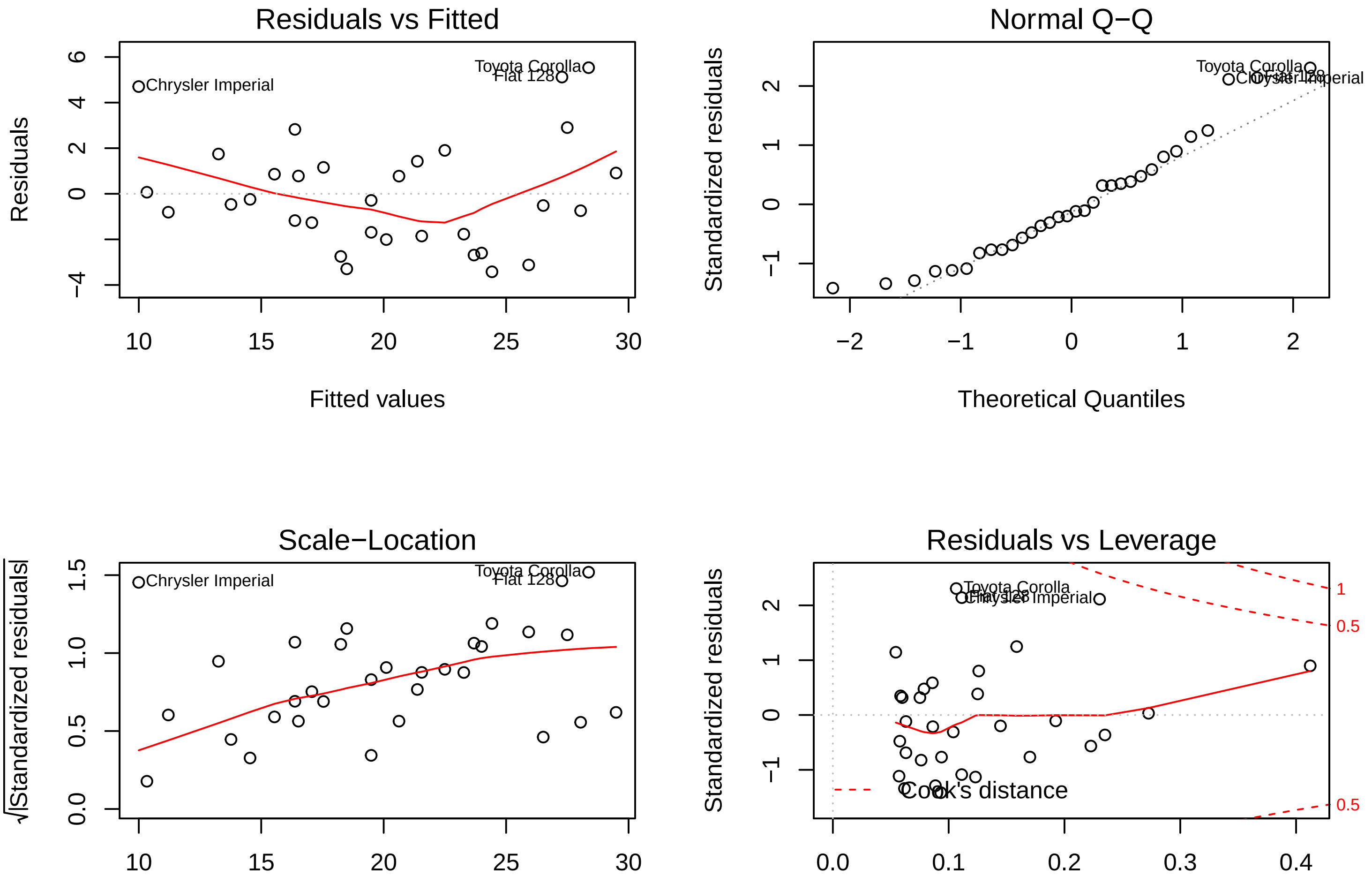
Conclusions

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
Chrysler Imperial
Fiat 128
Scale−Location
√|Standardized residuals|
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Toyota Corolla
Fiat 128
Chrysler Imperial
Cook's distance
1
0.5
0.5