Regression Models Course Project

kosw

2023-11-25

Summary

In this report, we explored the regression model with car's mpg(Miles/gallon) as reponse variable using 'mtcars' dataset. We excluded several variables among 10 variables(except mpg, response variable) that mtcars dataset contains, using correlation between variables and vif. After that, we diagnosed assumptions about errors with some diagnostic plots. Conclusionally, we got significant multivariate regression model explaining mpg variable with 3 variables : am, vs, carb (also with intercept)

Load Data

First, Load some packages we will use in this research : ggplot2, GGally, car

And Load Dataset we will use : mtcars

```
x <- mtcars
```

Exploratory Data Analysis

This dataset comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). And using this data, we wanna make regression model that has 'mpg' variable as reponse. And the most interested variable as regressor is 'am'. For exploration of the relationship between all variables briefly, let's make an pair plot!(see plot1)

First Model

For our first regression model, we will exclude one of two variables that has correlation over 0.8. It is because two variables that has strong correlation may explain reponse variable similarly, and including both variable can decrease the estimation accuracy of each others. When two variable has correlation over 0.8, we will remove one that has less correlation with 'am' variable because 'am' variable is variable we interested in. In this way, here we will exclude 2 variables : disp, cyl

Then we can make first model including all of the rest of variables.

```
model1 <- lm(mpg ~ am+hp+drat+wt+qsec+vs+gear+carb, x)
```

```
## (Intercept)          am          hp        drat          wt        qsec
##  0.29502745  0.21759174  0.46498344  0.56451388  0.03416499  0.31851450
##          vs        gear        carb
##  0.96332041  0.61292695  0.31059143
```

Above are p-values of each regressors. In this model, only one variable among 9 is siginificant - surely problem!!! It is presumably because, we include too many variables, so they decrease the accuracy of estimation. So We need to exclude a few more variables. Which variable will we exclude? To get significant conclusion, we will exclude variable that has the strongest correlation with other variables. We can see it through vif.

```
##       am       hp     drat       wt     qsec       vs     gear     carb
## 4.285815 6.015788 3.111501 6.051127 5.918682 4.270956 4.690187 4.290468
```

Let's exclude the largest of vif : wt

Second Model

By excluding 'wt' variable, we can make our second model.

```
model2 <- lm(mpg ~ am+hp+drat+qsec+vs+gear+carb, x)
```

```
## (Intercept)          am          hp        drat        qsec          vs
##  0.31340611  0.10103332  0.11309559  0.37232257  0.99717616  0.43964206
##        gear        carb
##  0.30324418  0.06686465
```

And, also not all variables are significant. So, by repeating this method until all variables are signigicant, we get out final model.

```
##       am       hp     drat     qsec       vs     gear     carb
## 4.052249 5.075011 3.020934 4.713681 3.791803 4.372869 3.641534
```

Final Model

Final model is following.

```
model3 <- lm(mpg ~ am+vs+carb, x)
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 19.517399  1.6090815 12.129528 1.155904e-12
## am           6.797956  1.1014890  6.171606 1.154742e-06
## vs           4.195736  1.3245867  3.167581 3.695735e-03
## carb        -1.430783  0.4081085 -3.505890 1.552505e-03
```

In this model, all regressors are significant.

Let’s try anova to check each variables are significant once more compared with several nested models.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + vs
## Model 3: mpg ~ am + vs + carb
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 353.49  1    367.41 41.878 5.198e-07 ***
## 3     28 245.65  1    107.83 12.291  0.001553 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All coefficient are significant so, It seems better to use our final model.

Our model’s coefficient’s 95 percent confidence interval is following.

```
##                 2.5 %    97.5 %
## (Intercept) 16.221345 22.813453
## am           4.541658  9.054254
## vs           1.482443  6.909029
## carb        -2.266756 -0.594811
```

Residual Plot and Diagnostics

Let’s make some diagnostic plot including residual plot and check several things about residuals and data points.

First, from residual plot(see plot2),it seems that the residuals have don’t have specific pattern. It meets the assumption of the linear model.

And from residual Q-Q plot(see plot3), it seems that the distrbution of residuals are almost normal.

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model3)
## W = 0.946, p-value = 0.1109
```

For more accurate test, we can do shapiro.test. In shapiro.test, p-value is over 0.05, so we fail to reject the null hypothesis : this distribution is normally distributted. So we can say our residuals are normal.

From Residual vs Leverage plot(see plot4), The largest leverage point has residual near zero. It conforms to regression line. And there is no point that has cook’s distance over 1. So there is no particularly influential point.

Conclusion

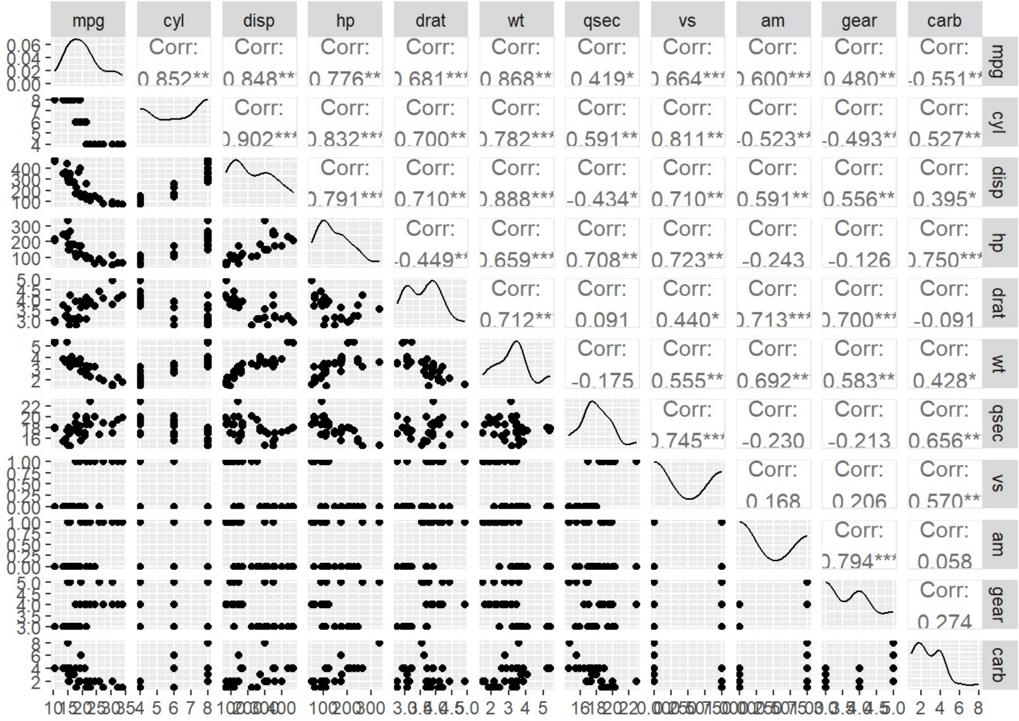
From final model, we can say 4 thing from our conclusion of 4 coefficients.

- First, all regressor are zero(‘automatic’/‘V-shaped Engine’/’0 carburetors), the expected value of mpg is 19.5174 Miles/gallon
- Second, When our am variable(Transmisson) goes automatic to manual holding other variables constant,the expected change in mpg value is increase 6.7980 Miles/gallon
- Third, When our vs variable(Engine) goes ‘V-shaped’ to ‘straight’ holding other variables constant, the expected change in mpg value is increase 4.1957 Miles/gallon.
- Finally, One more carburetor holding other variables constant, the expected change in mpg value is decrease -1.4308 Miles/gallon.

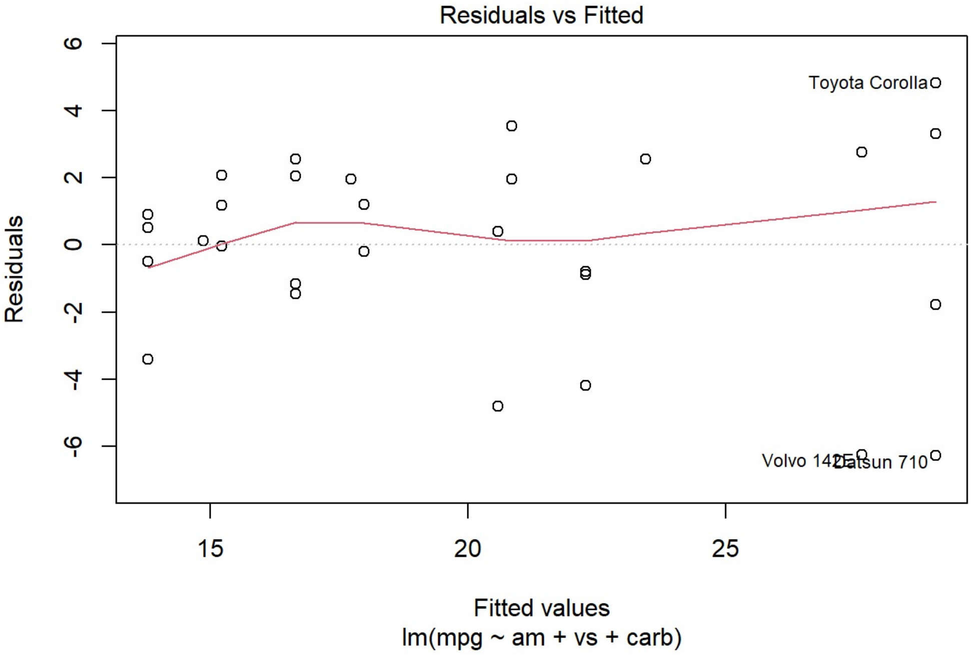
Additionally, And our model’s R Squared value is ’’, which means that our regression model is explaining significant proportion of the Total variance of observed data.

About the following 2 questions of interest : “Is an automatic or manual transmission better for MPG?”, “Quantify the MPG difference between automatic and manual transmissions”, we can say that manual transmission is 6.7980 Miles/gallon better for mpg than automatic when holding other variables constant.

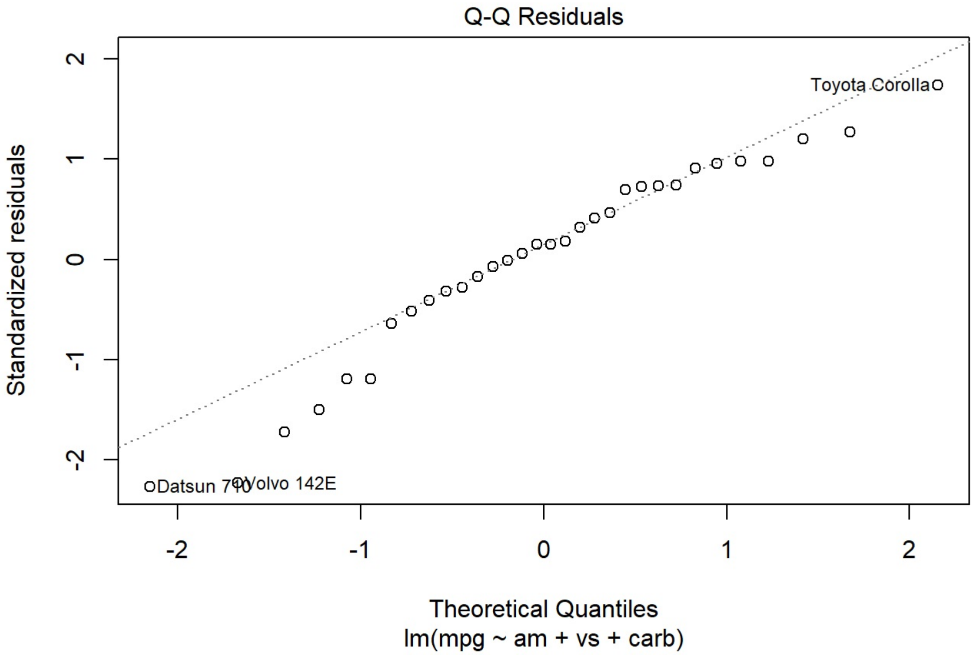
Appendix



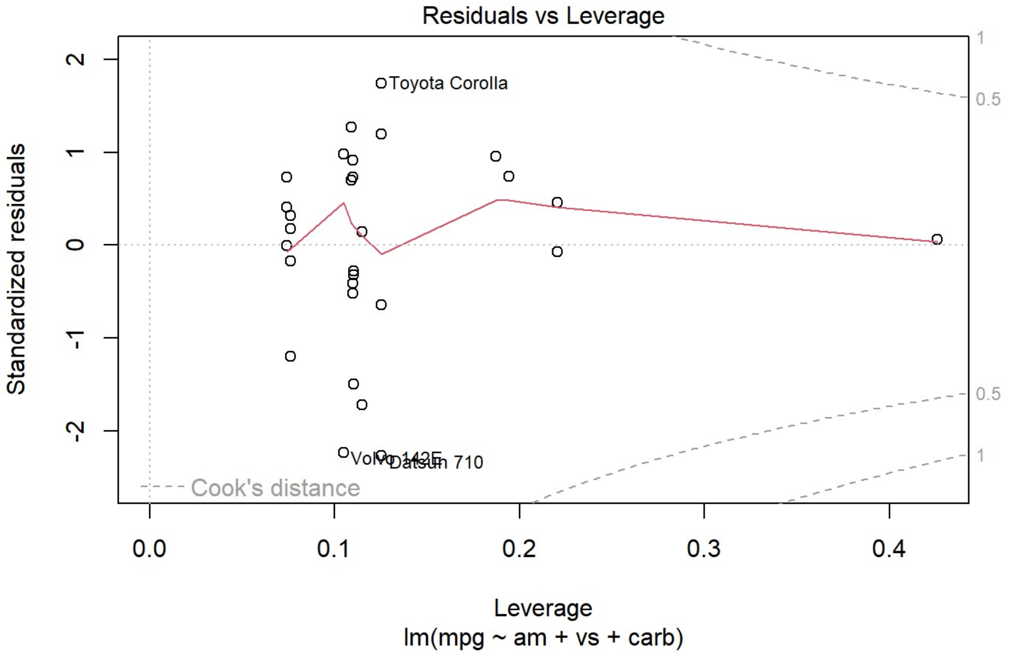
(plot1)



(plot2)


Q-Q Residuals
Standardized residuals
-2
-1
0
1
2
Theoretical Quantiles
lm(mpg ~ am + vs + carb)
Toyota Corolla
Datsun 710
Volvo 142E

(plot3)


Residuals vs Leverage
Standardized residuals
-2
-1
0
1
2
0.0
0.1
0.2
0.3
0.4
Leverage
lm(mpg ~ am + vs + carb)
Toyota Corolla
Volvo 142E
Datsun 710
Cook's distance
1
0.5
0.5
1

(plot4)