MPG Analysis - Regression Models Project

Executive Summary

In this project, we conduct analysis for for *Motor Trend* automobile industry magazine, to evaluate MPG (miles per gallon) model based on R's `mtcars` dataset. Specifically, the analysis tries to answer the following question:

- "Is an automatic or manual transmission better for MPG?"
- "Quantify the MPG difference between automatic and manual transmissions"

The analysis result shows that there is indeed MPG difference between Manual and Automatic transmission; the value is about `1.8` more MPG for Manual compared to Automatic. However, transmissions is not the only significant factor affecting MPG.

Explaratory Data Analyses

Load Data

Load `mtcars` dataset, and convert the following variables into Factor: `cyl`, `vs`, `gear`, `carb`, `am`.

Null Hypothesis

Assuming the dataset are normally distributed, t-test is performed on Null Hypothesis that there is no difference between Manual and Automatic transmission on MPG - which is rejected, as follow:

```
t.test(mpg ~ am, data = mtcars)[c('p.value', 'conf.int')]
```

```
## $p.value
## [1] 0.001373638
##
## $conf.int
## [1] -11.280194  -3.209684
## attr(,"conf.level")
## [1] 0.95
```

Pair Plots and Initial Regression

Figure 1 in Appendix shows relationships between variables in pair.

MPG (`mpg`) and Transmission (`am`) has the following linear regression property:

```
fit0 <- lm(mpg ~ am, data=mtcars)
summary(fit0)$coefficients
```

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

The above (`mpg` ~ `am`) will be our **first** model.

Linear Regression Analysis

Best Model Selection

We make use of function **bestmodel** to fit multiple models automatically. `bestmodel` uses Akaike information criterion (AIC) as the measure of quality of the models.

```
fit1 <- lm(mpg ~ ., data=mtcars)
bestmodel <- step(fit1, direction = "both")
```

```
summary(bestmodel)[c('call', 'coefficients', 'r.squared')]
```

```
## $call
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## $coefficients
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## ammanual     1.80921138 1.39630450  1.295714 2.064597e-01
##
## $r.squared
## [1] 0.8658799
```

The finding (`mpg` ~ `cyl` + `hp` + `wt` + `am`) is our **second** model. To confirm that the confounders `cyl`, `hp`, `wt`, `am` are indeed significant in estimating `mpg`, we perform the ANOVA:

```
anova(fit0, bestmodel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Diagnostics

Figure 2 in Appendix display the diagnostics plot of the chosen model. The *Residuals vs. Fitted* plot shows randomly scattered points indicating independence condition. The points on *Normal Q-Q* plot mostly fall on the line indicating that the residuals are normally distributed. The points on *Scale-Location* plot are scattered in a constant band pattern, indicating constant variance. Some distinct points of interest on *Residuals vs Leverage* plot (top & right) may indicate increased leverage of outliers.

Conclusion

- The ANOVA model confirms the **second model** (`mpg` ~ `cyl` + `hp` + `wt` + `am`) explaining about ~85% MPG variation.
- Manual transmission gives better MPG with 1.809 more miles per gallon with ~20% error within the model.

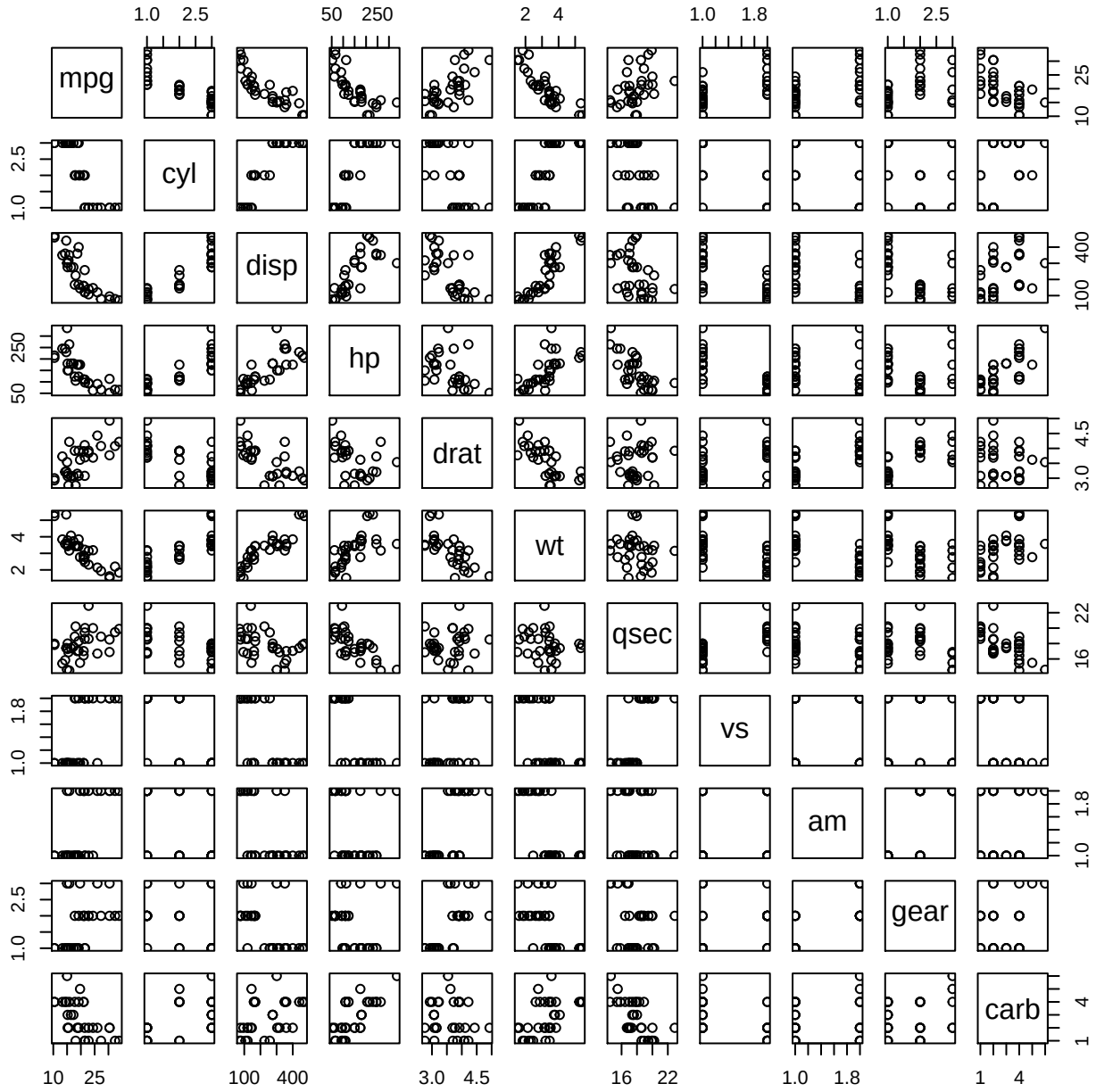

Figure 1: mtcars Pair Plot

Appendix

```
# Pair Plots:
pairs(mpg ~ ., data=mtcars)
```

```
# Best Model Diagnostics:
layout(matrix(1:4,2,2))
plot(bestmodel)
```

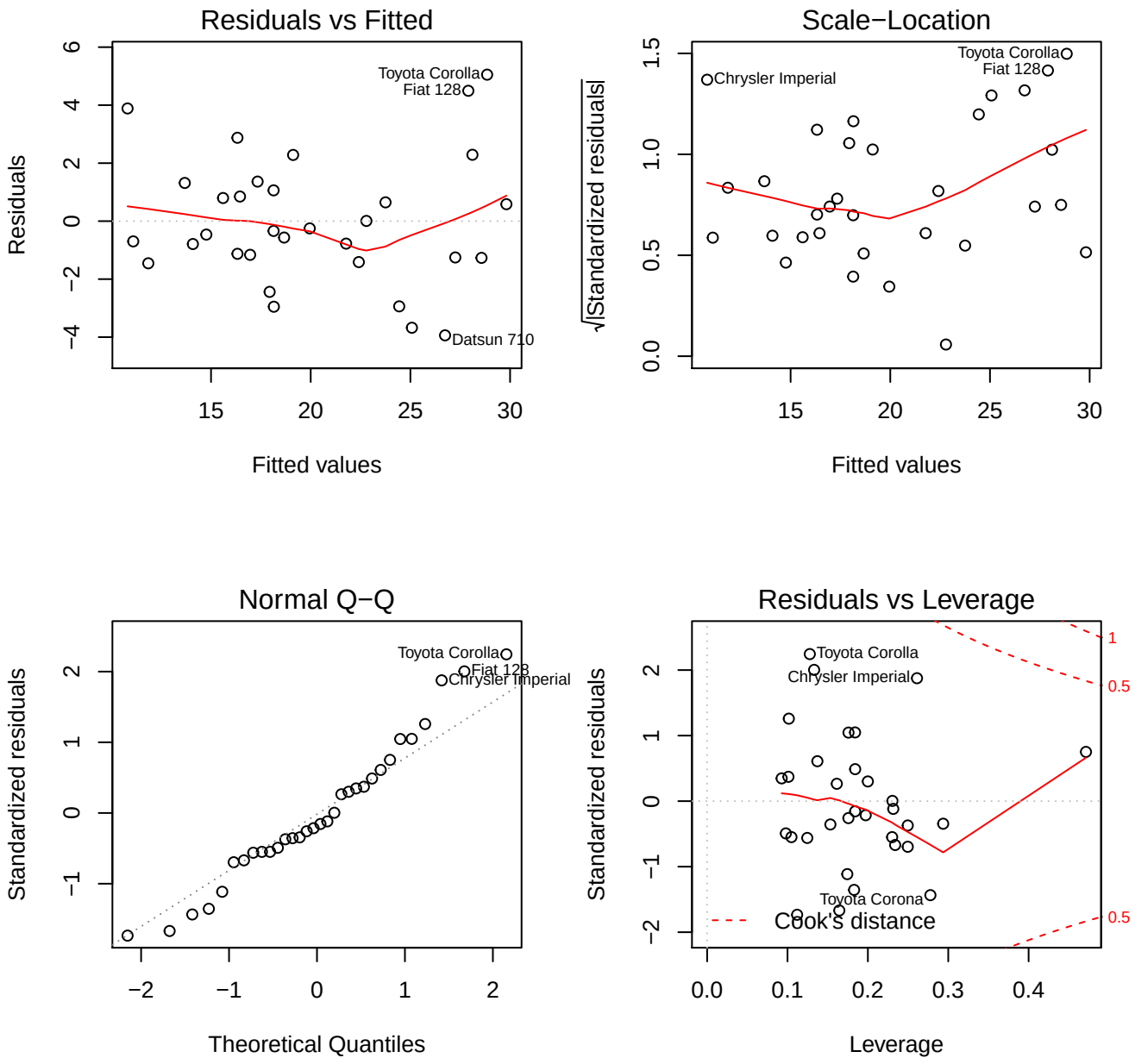

Figure 2: Diagnostics of model: mpg = cyl + hp + wt + am