Motor Trend Cars Fuel Efficiency Analysis

Executive Summary

By studying the dataset `mtcars`, we found that the manual transmission is better for MPG than the automatic one, by a margin of 1.809. We reach this conclusion by finding the best model and exploring the relationship between a set of variables as predictors and the MPG as the output.

Exploratory Data Analysis

We begin the exploratory analysis by checking the structure and data class of the data frame `mtcars`.

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
head(mtcars, 2)
```

```
##               mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4      21   6  160 110  3.9 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

The data set has 32 observed samples and 11 variables, among which the "mpg" is the outcome and the other 10 are the predictors. It is noted that some variables, which should be considered as categorical factors, are listed as numeric. The conversion is conducted first.

```
mtcars$cyl<-as.factor(mtcars$cyl)
mtcars$vs<-as.factor(mtcars$vs)
mtcars$am<-as.factor(mtcars$am)
mtcars$gear<-as.factor(mtcars$gear)
mtcars$carb<-as.factor(mtcars$carb)
```

A boxplot is performed between "mpg" and "am" (transmission type, 0 for automatic, 1 for manual)[Fig-1], and it's obvious that the manual transmission has higher MPG. We also plot the "mpg" as the function of all other 10 predictors in Fig-2 and found that it has either positive or negative dependency on all predictors except the "gear".

Model Fitting and Selection

The visually observed dependency between "mpg" and almost all other predictors indicates the possible multicollinearity between the predictors. It is checked by calculating the variance inflation factor (vif) of each predictor in the complete fitting model.

```
library(car)
fitAll<-lm(mpg~., mtcars)
vif(fitAll)
```

```
##            GVIF Df GVIF^(1/(2*Df))
## cyl  128.120962  2        3.364380
## disp  60.365687  1        7.769536
## hp    28.219577  1        5.312210
## drat   6.809663  1        2.609533
## wt    23.830830  1        4.881683
## qsec  10.790189  1        3.284842
## vs     8.088166  1        2.843970
## am     9.930495  1        3.151269
## gear  50.852311  2        2.670408
## carb 503.211851  5        1.862838
```

The high VIF values of "disp", "hp" and "wt" suggest that a correlation exists between them and other predictors. The complete model with all 10 predictors is obviously not optimum and the number of predictors could be reduced. We use the Adjusted R-Squared, P values and backward elimination method to find the best model.

```
library(leaps)
subset<-regsubsets(mpg~., mtcars, nbest=2)
```

Fig-3 shows all the possible models and their corresponding Adjusted R-Squared values. We pick some top models with highest Adjusted R-Squared values for analysis, and try to find the best model among them.

```
fit1<-lm(mpg~cyl+hp+wt+am, mtcars)
fit2<-lm(mpg~cyl+hp+wt+am+vs, mtcars)
fit3<-lm(mpg~cyl+hp+wt+am+qsec, mtcars)
```

The "best" model is the one with fewest predictors yet still with one of the largest Adjusted R-Squared value among all models. The complete `fitAll` model has an Adjusted R-Squared value of ~0.807, and the `fit1`, `fit2`, and `fit3` models have very similar Adjusted R-Squared value of ~0.84. Therefore we intend to go with the `fit1` model. Now let's do the `anova` analysis bewteen `fit1` and `fit2`, and bewteen `fit1` and `fit3`, to test if they are substantially different.

```
anova(fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + hp + wt + am
## Model 2: mpg ~ cyl + hp + wt + am + vs
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 151.03
## 2     25 143.68  1    7.3459 1.2782  0.269
```

```
anova(fit1, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + hp + wt + am
## Model 2: mpg ~ cyl + hp + wt + am + qsec
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     26 151.03
## 2     25 143.98  1    7.0439 1.223 0.2793
```

When comparing the fit1 and fit2 models, the P value of 0.269 is larger than 0.05, so the null hypothesis that

the two models have no difference is tested false. Therefore there is a good reason to go with `fit1` model which has fewer predictors. The same reasoning is behind the `anova` analysis of `fit1` and `fit3` models, which has a p value of 0.279. **fit1 is the best model**.

```
summary(fit1)$coef
```

```
##                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
## am1          1.80921138 1.39630450  1.295714 2.064597e-01
```

```
summary(fit1)$adj.r.squared
```

```
## [1] 0.8400875
```

The slope coefficient 1.809 of the predictor `am`, the transmission type, shows that the `mpg` will increase by 1.809 when using a manual transmission instead of an automatic one. The p value of 2.065e-1 is not significant, so there is good reason to believe that the slope value is different from 0. The adjusted R squared value of 0.84 show that the model can explain 84% of the total variance.

Residual Analysis

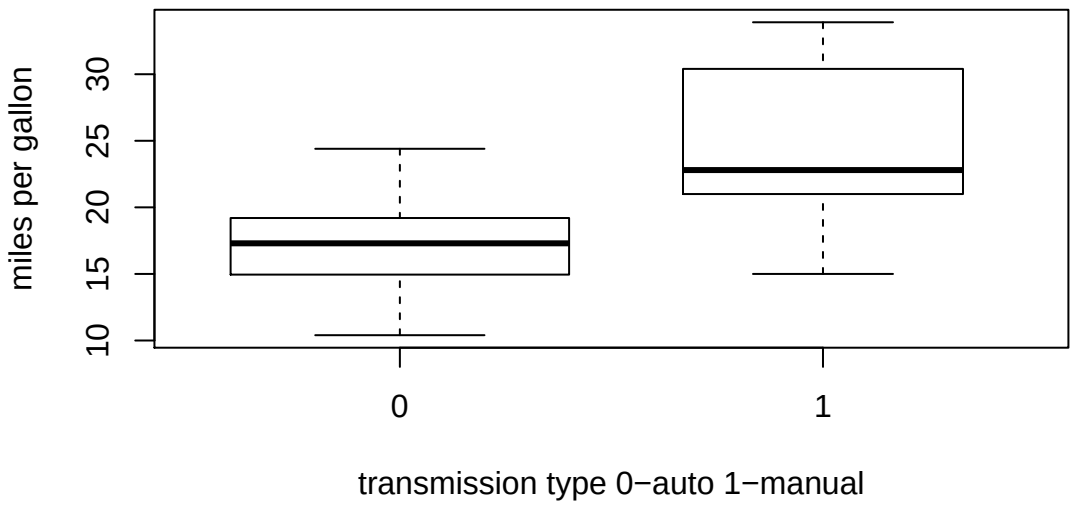
The plots of `fit1` are attached as Fig-4. The "Residual vs Fitted" and "Standard Residual vs Fitted" curves show the residual is distributed randomly around 0 and no specified pattern can be seen. The "QQ" and "Cook's Distance" plots reveal, however, there may be some outliners toward the end of the sample data.

To Answer the Questions

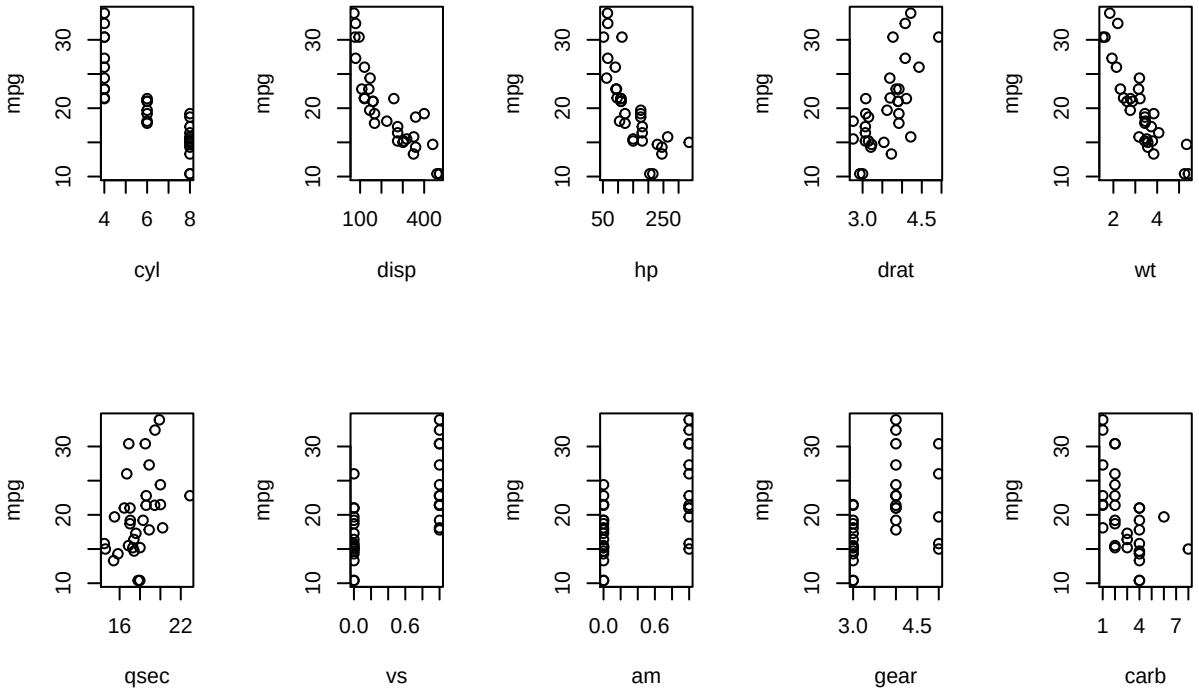
1. The manual transmission is better for MPG
2. The expected MPG difference between the automatic and manual transmission is 1.809, with the manual one being higher.

Appendix

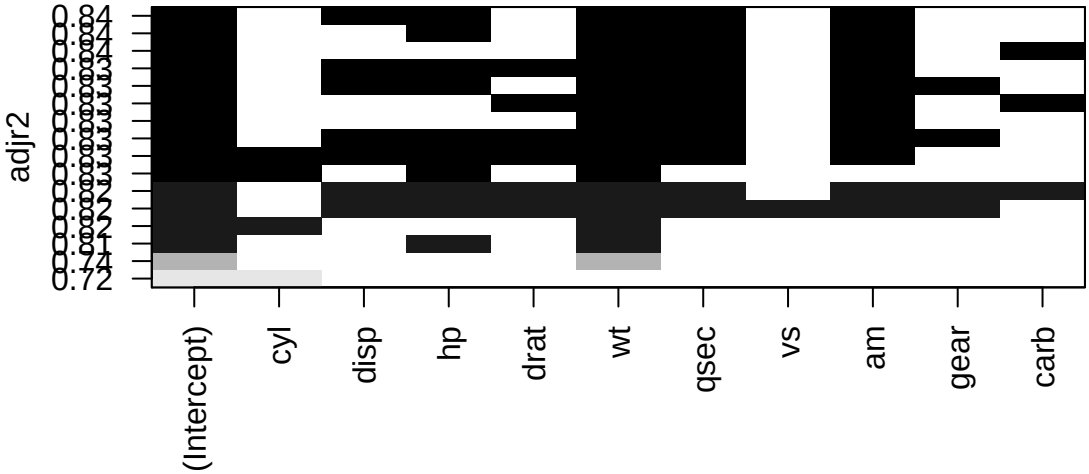
Fig−1 mpg vs transmission type

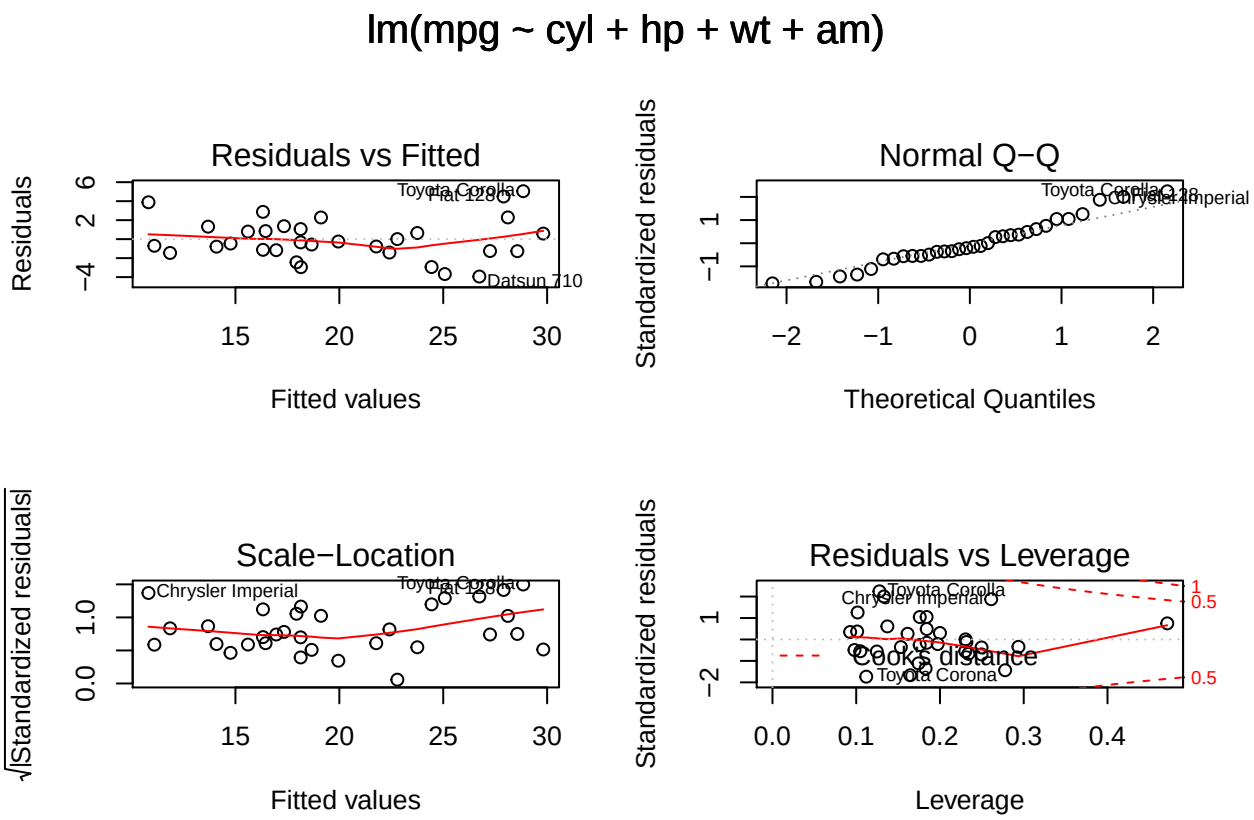


Fig−2 mpg vs all predictors



Fig−3 Adjusted R−Squared Plot for All Models




lm(mpg ~ cyl + hp + wt + am)
Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Toyota Corolla
Chrysler Imperial
Toyota Corona