Regression Models Course Project

Roshan Riazi

Executive Summary

Looking at a data set of a collection of cars ("mtcars" dataset), we want to explore the relationship between a set of variables and miles per gallon (MPG) (outcome). If we just use "am" variable to fit a linear model on "mpg", we will find that it has a significant p-value and manual cars have 7.2449 more mpg than automatic cars, but this model has a low adjusted R-sqared of 0.3385. When we find the best linear model with 5 predictor that explains 0.847 of the variance in "mpg", "am" variable has an insignificant p-value of 0.05421, and if every variable is going to be constant, manual cars will have 2.6211 more expceted value of "mpg" than automatic cars.

Exploratory Data Analyses

We first load the "mtcars" data set and look at its structure. By looking at its structure we find that there are some variables that should be factor variables, but are numeric. We change their type, so that our regressions will treat them appropriately.

```
data(mtcars)
str(mtcars)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("automatic", "manual")
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
```

We have ploted some exploratory plots, which can be found in the appendix. By looking at these plots we can see that mpg of automatic cars is lower than mpg of manual cars. So we will fit a linear regression using just "am" as the predictor to see its individual effect on "mpg".

Fitting Linear Regression with "am" as the only predictor

```
fit.lm1 <- lm(mpg ~ am, data = mtcars)
summary(fit.lm1)$coef
```

```
##              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
```

We can see that both intercept and "ammanual" have very low p-values and are significant. By just considering "am" as predictor, automatic cars have expected value mpg of 17.1474 and expected change in mpg for manual cars is 7.2449. So we expect that manual cars have 7.2449 more mpg than automatic cars. But we should be aware that this is a model with just "am" as predictor, with adjuste R-sqared of 0.3385, which explains just 0.3385 of variation in mpg and there may be better predictors for mpg in this dataset. (Also, residual plots show some patterns.)

Finding the Best Linear Model

We will use "regsubset" function in "leaps" package to find the best linear regression model in this dataset. We will use full search method for model selection.

```
library(leaps)
regfit.full <- regsubsets(mpg ~ ., data = mtcars, nvmax = 16)
reg.summary <- summary(regfit.full)
maxAdjR2 <- which.max(reg.summary$adjr2)
#coef(regfit.full, maxAdjR2)
```

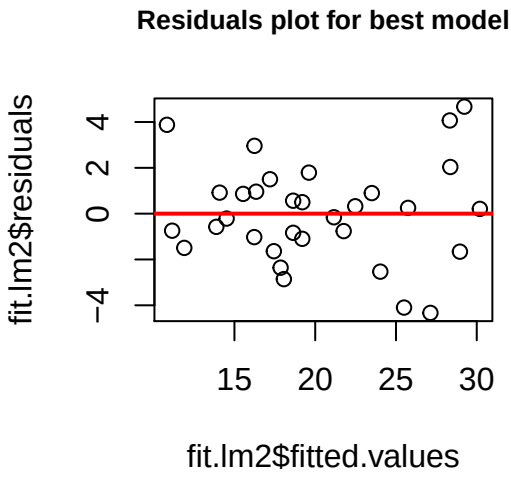
By using "regsubset" and full search method, we found that the best model consists of 5 predictors (and intercept), which has an adjusted R-squared of 0.8478 (explained variance of "mpg"). It's important to note that each level of variables is considered a distinct predictor! For example cyl variable with level of 6 is one of this 5 predictors.

```
fit.lm2 <- lm(mpg ~ I(cyl == 6) + hp + wt + vs + am, data = mtcars)
summary(fit.lm2)$coef
```

```
##                    Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     31.28240981 3.19020789  9.805759 3.184553e-10
## I(cyl == 6)TRUE -2.20519611 1.03213033 -2.136548 4.221406e-02
## hp              -0.03393442 0.01044009 -3.250396 3.178741e-03
## wt              -2.36781111 0.86855157 -2.726161 1.131631e-02
## vs1              1.87741318 1.24809114  1.504228 1.445745e-01
## ammanual         2.62111773 1.29995401  2.016316 5.420741e-02
```

By looking at coefficients of this model, we can see that "ammanual" has a p-value of 0.05421, which isn't significant, but is very close to 0.05! Considering this somehaw large p-value, if every variable is going to be constant, manual cars will have 2.6211 more expceted value of mpg than automatic cars. But we should say again that with other variables in the model, p-value of "am" is somehaw large and insignificant.

So, lets see the residual plot of this model.

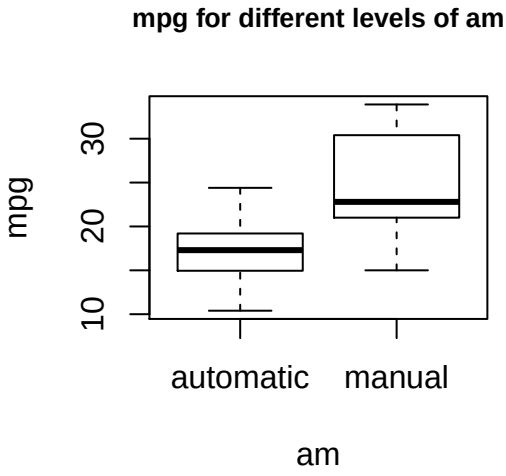


Residuals don't seem to have any obvious pattern in this plot, which is a good sign and indicates that there isn't an obvious problem in this model. More residual plots can be found in the appendix.

Appendix

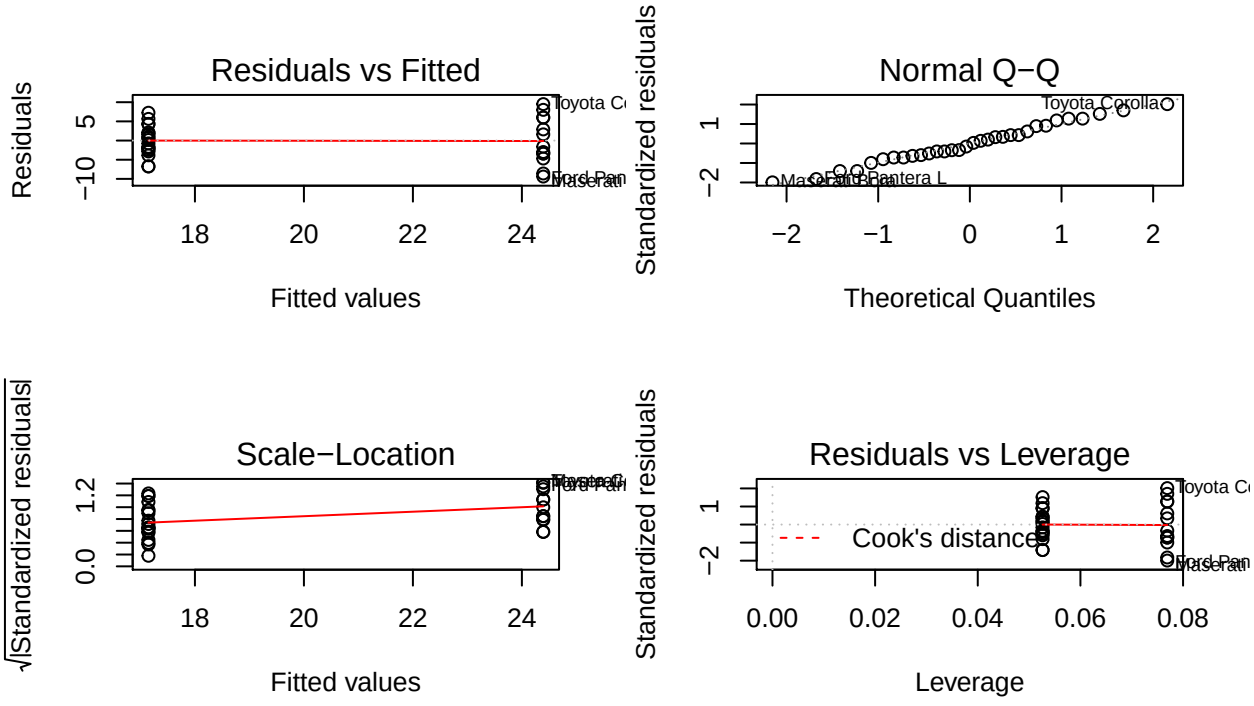
Plot of mpg for different levels of am

```
plot(mpg ~ am, data = mtcars, main = "mpg for different levels of am", cex.main = .8)
```



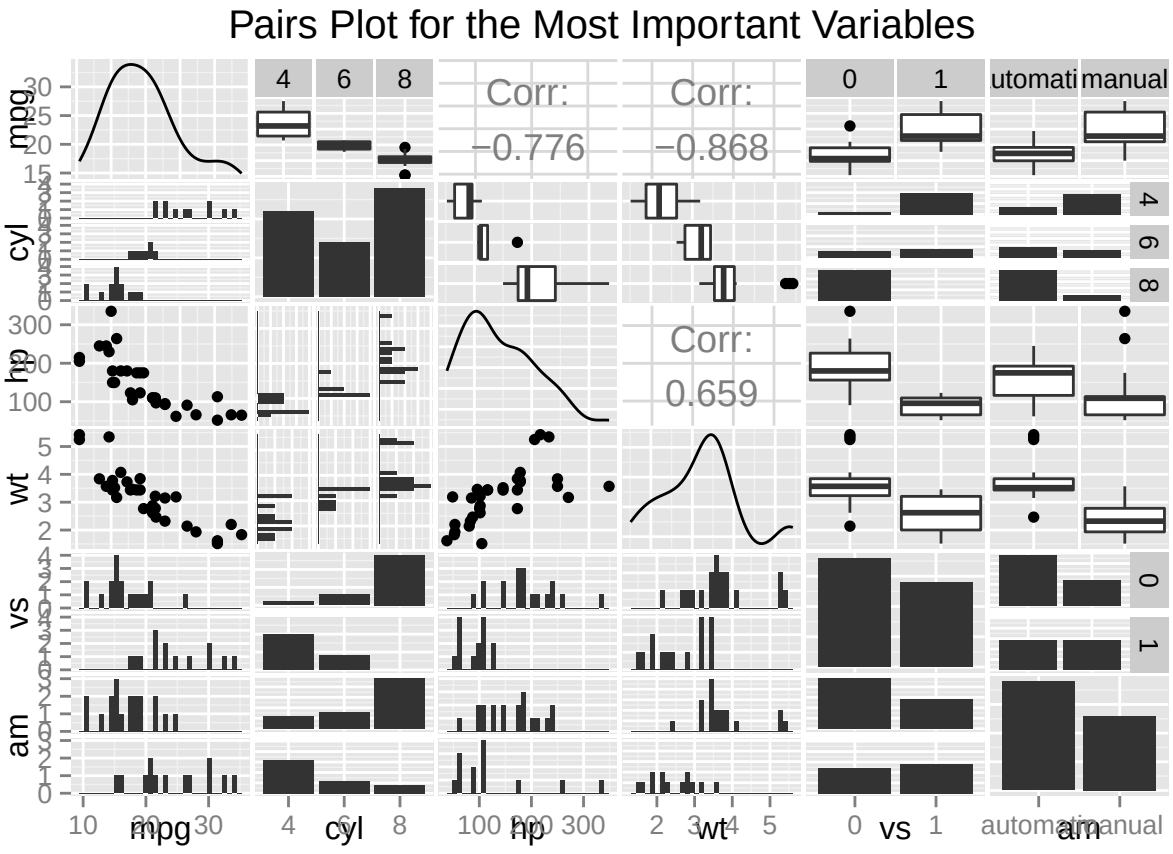
Residual Plots for first model

```
par(mfrow = c(2, 2))
plot(fit.lm1)
```



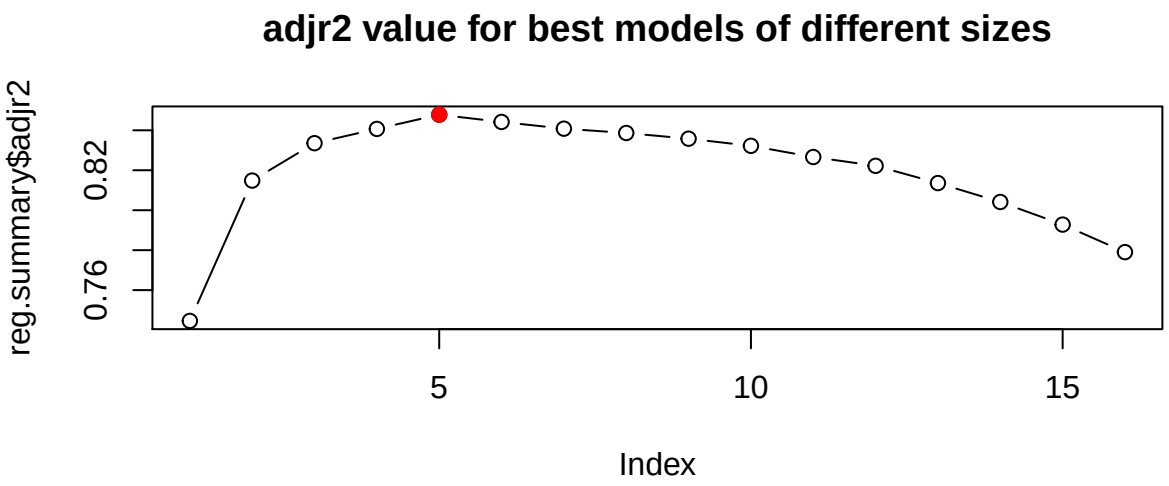
Pairs plot for the Most Important Variables

```
library(GGally)
ggpairs(mtcars, columns = c(1, 2, 4, 6, 8, 9),
        title = "Pairs Plot for the Most Important Variables")
```



Finding the Best Linear Model

```
library(leaps)
regfit.full <- regsubsets(mpg ~ ., data = mtcars, nvmax = 16)
reg.summary <- summary(regfit.full)
maxAdjR2 <- which.max(reg.summary$adjr2)
#coef(regfit.full, maxAdjR2)
plot(reg.summary$adjr2, type = "b", main = "adjr2 value for best models of different sizes")
points(maxAdjR2, reg.summary$adjr2[maxAdjR2], col = "red", pch = 19)
```



Residual Plots for the best model

```
par(mfrow = c(2, 2))
plot(fit.lm2)
```