Regression Models Course Project

linhns

6/3/2020

Packages

```r
library(ggplot2)
library(dplyr)
library(knitr)
library(car)
options(digits = 3)
```

1. Executive summary

Motor Trend are interested in exploring the relationship between a set of variables and miles per gallon (MPG). Using a data set of a collection of cars, we take a look at answering the following questions:
- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions.

We will conduct some EDA then fit several models.

2. Load data

```r
data("mtcars")
```

3. Basic exploratory data analysis

```r
## Create a summary of the first 3 entries of mtcars
kable(head(mtcars, 3), caption = "First 3 rows of the mtcars dataset")
```

Table 1: First 3 rows of the mtcars dataset

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.62	16.5	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.88	17.0	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.6	1	1	4	1

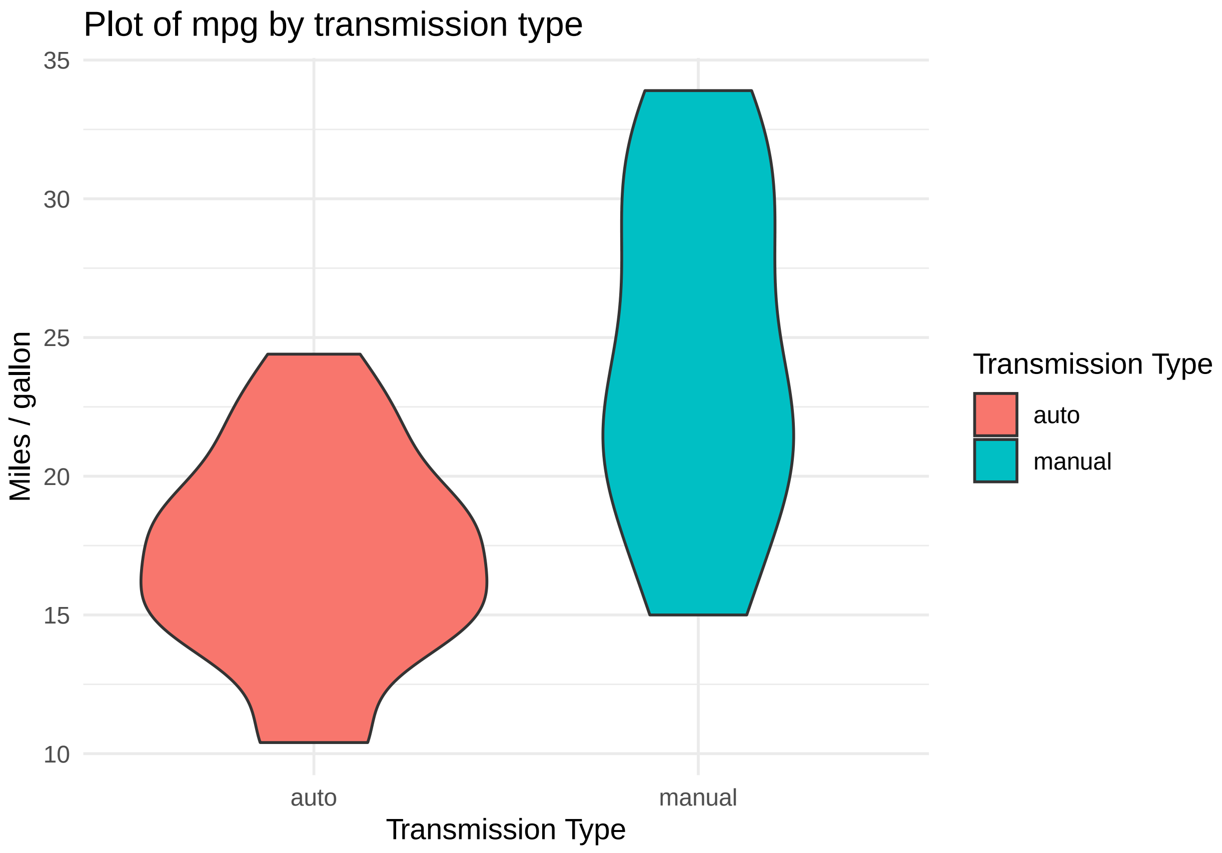
From the help file of this dataset, which was extracted from the 1974 Motor Trend US magazine, we see that it comprises fuel consumption and 10 other aspects of automobile design for 32 automobiles.

```r
# Summary statistics for the mpg
kable(t(as.matrix(summary(mtcars$mpg))),
      caption = "Summary Statistics mpg",align = c("c", "c"))
```

Table 2: Summary Statistics mpg

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
10.4	15.4	19.2	20.1	22.8	33.9

```r
mtcars <- mtcars %>% mutate(am = as.factor(am))
levels(mtcars$am) <- c("auto", "manual")
mtcars %>% ggplot(aes(am, mpg)) +
    geom_violin(aes(fill = am)) +
    labs(title = "Plot of mpg by transmission type", x = "Transmission Type",
         y = "Miles / gallon") +
    scale_fill_discrete("Transmission Type") +
    theme_minimal()
```



From the violin plot, it is notable that manual transmission is associated with greater *mpg* than automatic tranmission. See also the summary statistics.

Now we look at fitting different models based on a hypothesis test.

4. Regression models

Hypothesis test

Let's remind ourselves of our initial question: "Is an automatic or manual transmission better for mpg?".

Null hypothesis: $H_0 : \beta_1 = 0$ manual transmission is not a significant predictor for mpg.
Alternative hypothesis: $H_a : \beta_1 \neq 0$ manual transmission is a significant predictor for mpg.

Assume car models sample are independent of each other.

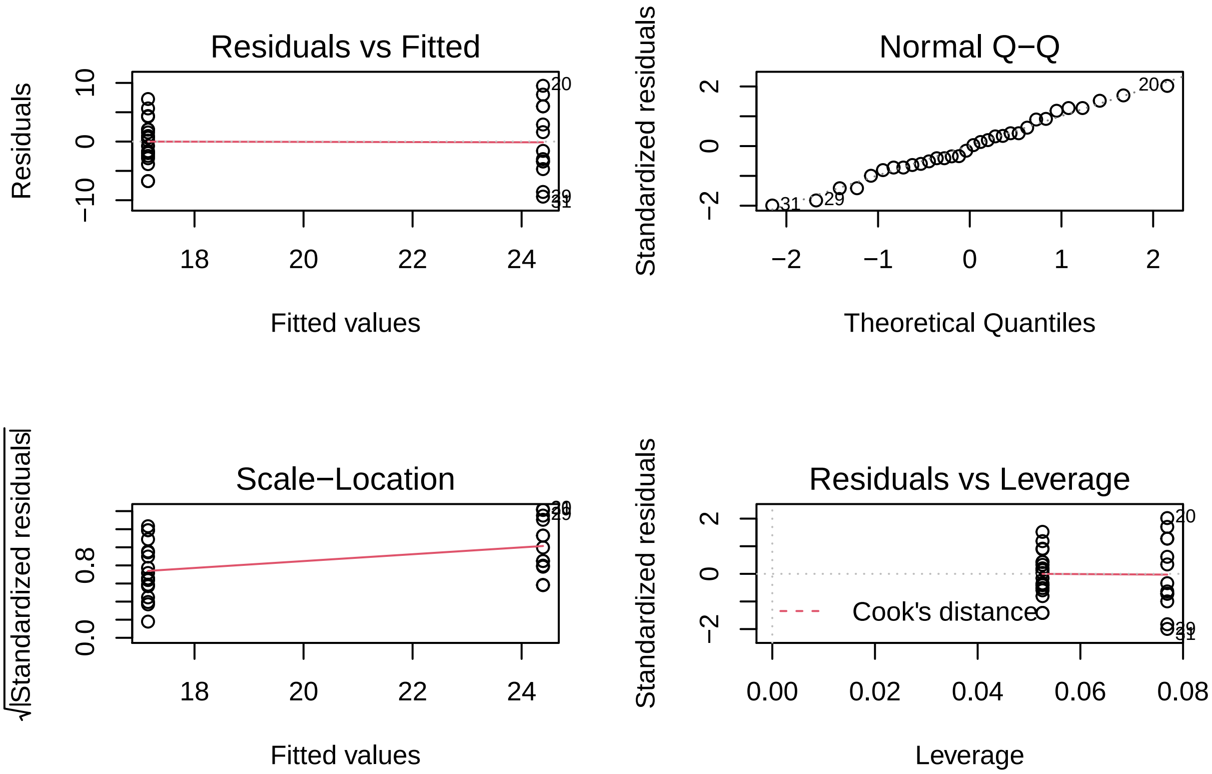
Simple linear regression model

We apply the simple linear model first because it is intuitive and easy to do in R.

```
simple_linear <- lm(mpg ~ am, mtcars)
coef(summary(simple_linear))
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    17.15       1.12   15.25 1.13e-15
## ammanual        7.24       1.76    4.11 2.85e-04
```

```
par(mfrow=c(2,2))
plot(simple_linear)
```



Looking at the coefficients table, we would reject the null hypothesis since the p-value is less than 5%.

The adjusted R squared is 0.34 which indicates this may not be the best model.

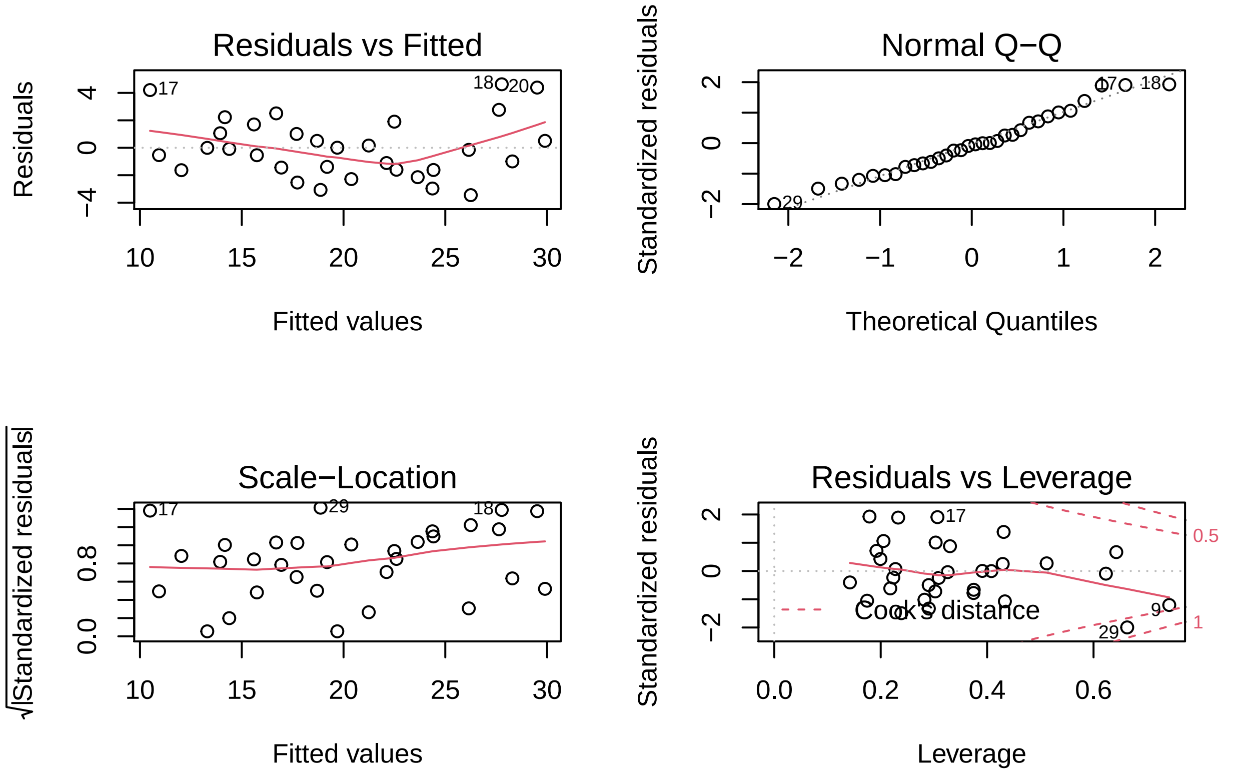
There may be other variables influencing *mpg* so we will investigate with a multivariable linear model.

Multivarible regression model

```
mvr <- lm(mpg ~ ., mtcars)
coef(summary(mvr))
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.3034    18.7179   0.657   0.5181
## cyl          -0.1114     1.0450  -0.107   0.9161
## disp          0.0133     0.0179   0.747   0.4635
## hp           -0.0215     0.0218  -0.987   0.3350
## drat          0.7871     1.6354   0.481   0.6353
## wt           -3.7153     1.8944  -1.961   0.0633
## qsec          0.8210     0.7308   1.123   0.2739
## vs            0.3178     2.1045   0.151   0.8814
## ammanual      2.5202     2.0567   1.225   0.2340
## gear          0.6554     1.4933   0.439   0.6652
## carb         -0.1994     0.8288  -0.241   0.8122
```

```
par(mfrow=c(2,2))
plot(mvr)
```

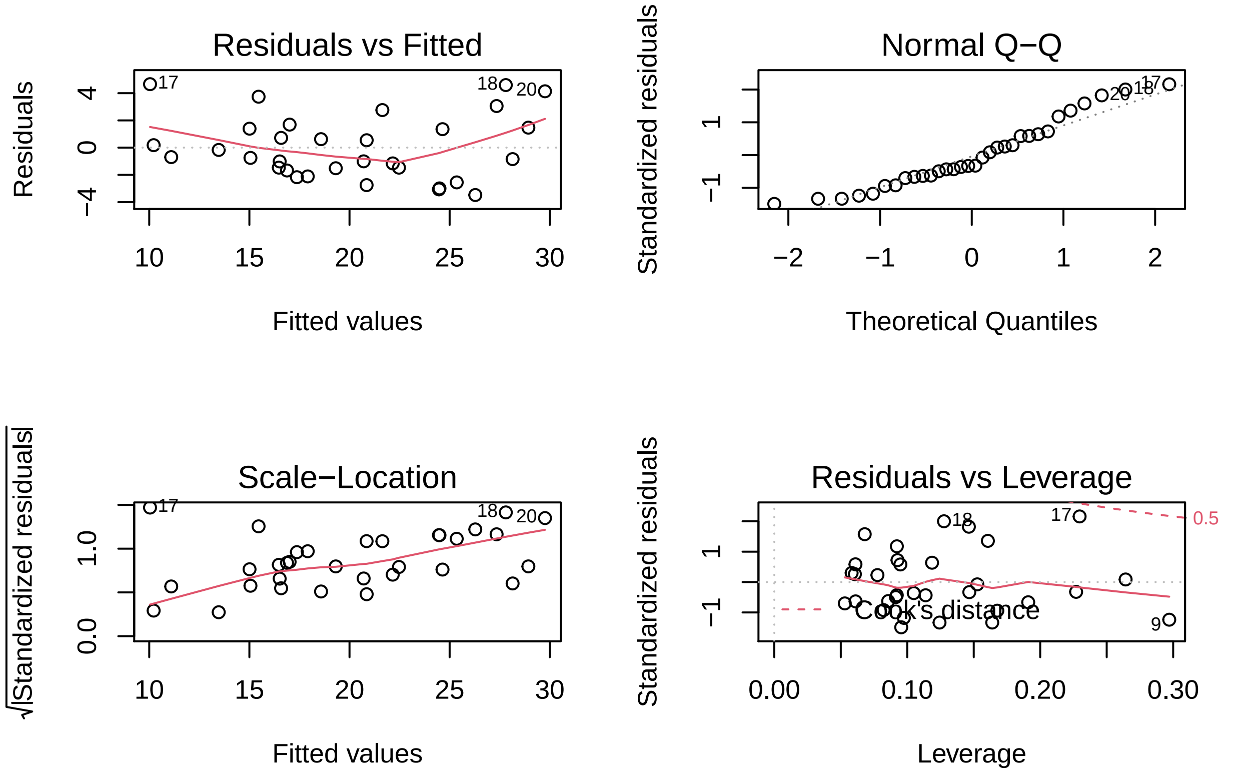


Looking at the coefficient table, we cannot reject the null hypothesis since all p-values are larger than 5%. Hence, we will now try a third model, which is stepwise backward regression.

```
sw <- step(mvr, direction = "backward", trace = FALSE)
coef(summary(sw))
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.62      6.960    1.38 1.78e-01
## wt              -3.92      0.711   -5.51 6.95e-06
## qsec             1.23      0.289    4.25 2.16e-04
## ammanual         2.94      1.411    2.08 4.67e-02
```

```
par(mfrow=c(2,2))
plot(sw)
```



The adjusted R squared for this third model is 0.83.

We can summarise the third model further by saying that the manual transmission appears to be a significant predictor of *mpg* and we expect an increase of 2.94 *mpg* when choosing manual over an automatic transmission, with other variables held constant.

Model diagnostic
As the adjusted R-squared for the third model is much higher than the first one, we can say it is a better fit. We can compare models using anova.

```
anova(simple_linear,sw)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     30 721
## 2     28 169  2       552 45.6 1.6e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is <0.05% we would reject a null hypothesis that the variable coefficients for model sw are 0.

Other models
Since mpg is numerical, the logistic and Poisson regression models are not applicable.