Regression Models Course Project

kimnewzealand

28 June 2017

Load packages

```
library(ggplot2)
library(dplyr)
library(knitr)
options(digits=3)
devtools::install_github("rstudio/rmarkdown")
library(car)
```

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 perform some EDA then fit three models, a linear model and two multivariable linear models and show that the third one using model selection may be a better model fit based on adjusted R-squared.

2. Load data

```
data(mtcars)
```

3. Perform basic exploratory data analysis

```
## Create a summary of the top 2 records from mtcars dataset
kable(head(mtcars,2), caption="Summary of first rows of mtcars Dataset",align = c("c", "c"))
```

Table 1: Summary of first rows of mtcars Dataset

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.5	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.88	17.0	0	1	4	4

The mtcars data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles. Each row in the dataset is a make of car, with each column as a aspect of design and performance.

From the violin plot (Figure 1 in the Appendix) it appears that the *mpg* for the manual transmissions have greater *mpg* than for the automatic transmissions. See also the summary statistics (Table 2).

We will take a look at fitting different models based on a hypothesis test.

4. Regression Models

HYPOTHESIS TEST

First set the hypothesis test for the question "Is an automatic or manual transmission better for MPG?":

The null hypothesis $H_0 : \beta_1 = 0$ the manual transmission is not a significant predictor for *mpg*.
The alternative hypothesis $H_A : \beta_1 \neq 0$ is that manual transmission is a significant predictor for *mpg*.

We assume for the test that the sampled car types are independent of each other.

SIMPLE LINEAR REGRESSION MODEL

The first model we will apply is the simple linear model using the lm function in R on the factor of the categorical predictor variable *am* with levels automatic transmissions (0) and manual transmissions (1), of the numerical response variable *mpg*.

```
##  Create a linear regression model lm1
lm1 <- lm(mpg~factor(am),mtcars)
```

From the coefficient summary in the Appendix, the p-value is $< 5\%$, therefore we would reject the null hypothesis in favour of the alternative hypothesis that the manual transmission is a significant predictor given no other variables are present in the model.

The adjusted R squared is 0.34 which is not very high so this may not be the best model yet. There may be other variables that impact *mpg* so we will investigate with a multivariable linear model.

MULTIVARIABLE LINEAR REGRESSION MODEL

The second model we will apply is the multivariable linear model, to view if the transmission type (*am*) is a significant predictor, when other significant variables are included in the model.

```
# Create a multivariable linear model of mpg to all the other 10 variables
lm2 <- lm(mpg~.,mtcars)
```

The p-values for the included variables are 0.52, 0.92, 0.46, 0.33, 0.64, 0.06, 0.27, 0.88, 0.23, 0.67, 0.81 which are all greater than 0.05%, so we would not reject the null hypothesis, given all other variables included in the model.

We will apply stepwise backward model selection in a third model.

```
# Calculate model using stepwise backwards model selection.
sw <- step(object = lm2,direction = "backward",trace =FALSE)
```

The adjusted R squared for this third model is 0.83.

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

MODEL COMPARISON AND DIAGNOSTICS

The adjusted R squared for this third model using stepwise backwards is 0.83 which is higher than the adjusted R squared for the first model with only 1 factor, with 0.34, so the third may represent a better model fit. Also see the Appendix for diagnostics with residuals plots of each model, and an ANOVA comparison.

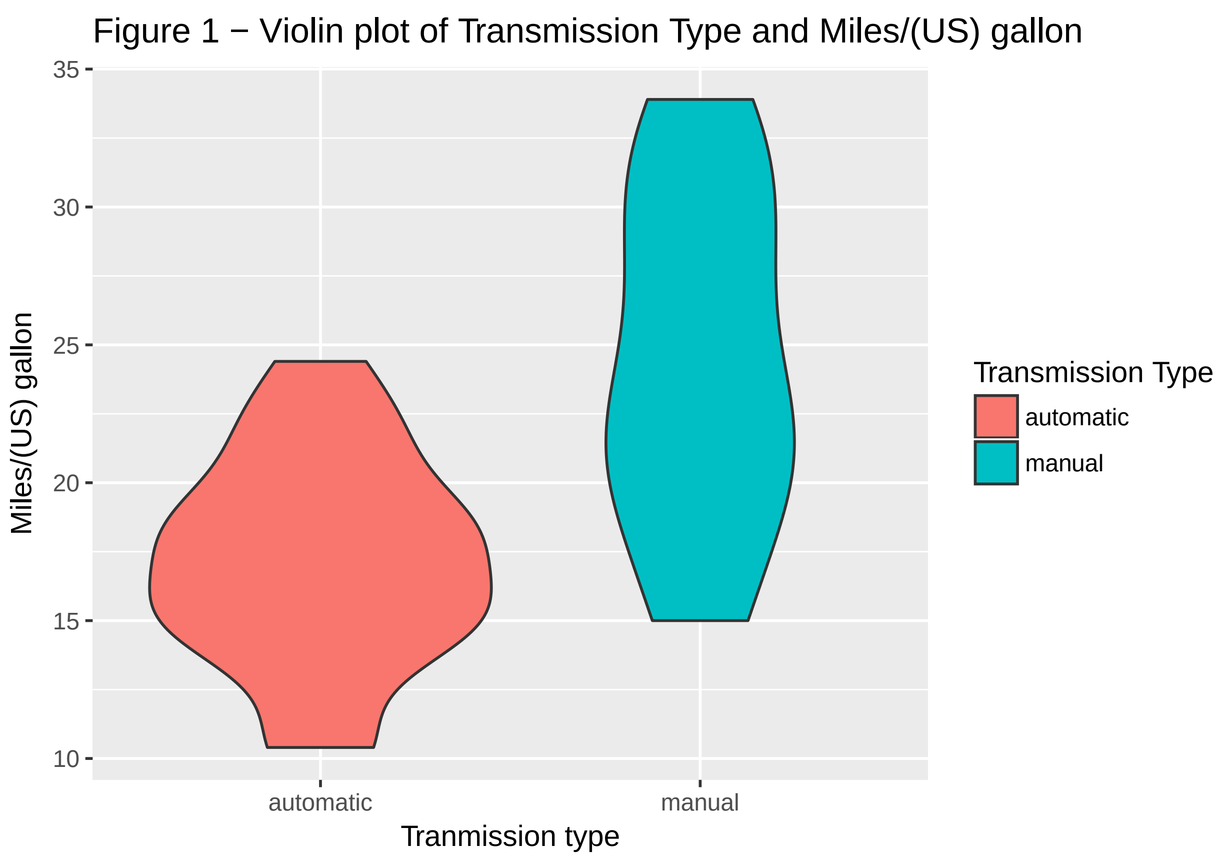
5. APPENDIX

```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
# Plot a violinplot to see the transmission types by mpg
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("automatic", "manual")
g <- ggplot(mtcars,aes(x=am,y=mpg))
g + geom_violin(aes(fill=am)) +
  labs(title="Figure 1 - Violin plot of Transmission Type and Miles/(US) gallon",
       x="Tranmission type",y="Miles/(US) gallon") +
  scale_fill_discrete("Transmission Type")
```



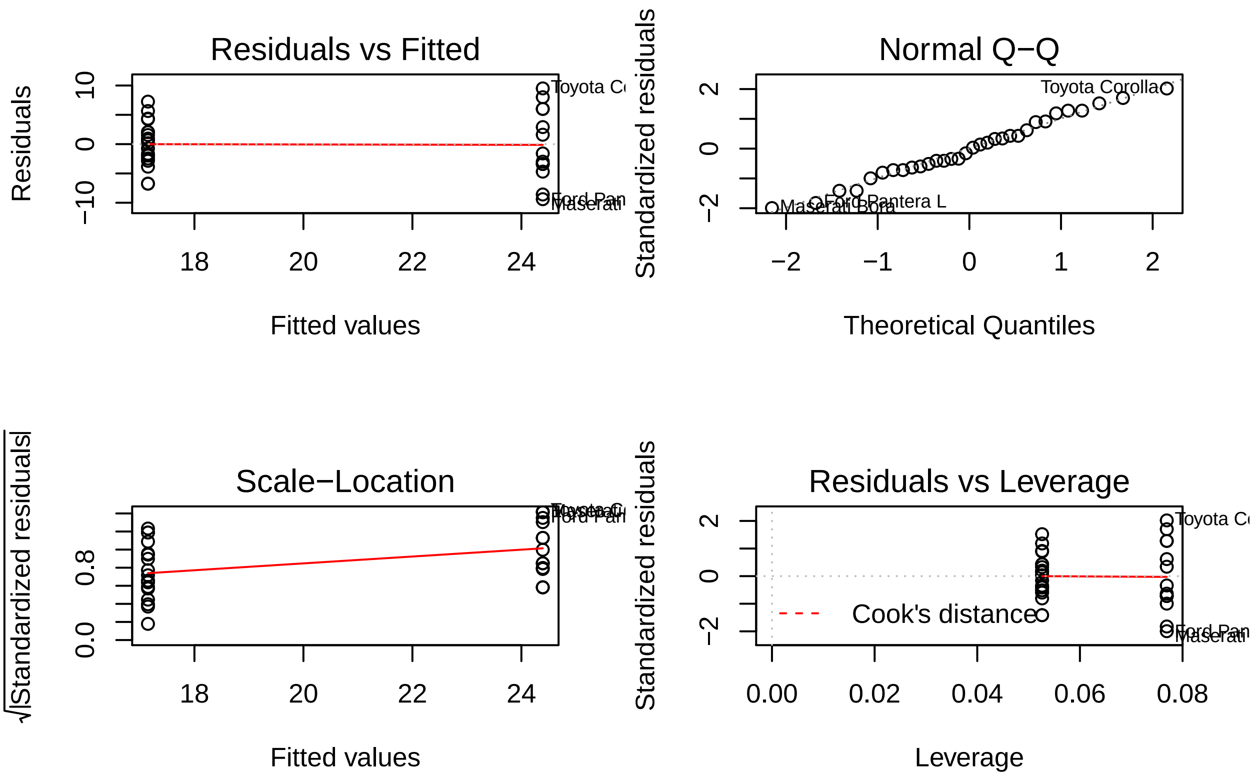
SIMPLE LINEAR REGRESSION MODEL

```r
##  Produce a summary of lm1
summary(lm1)[4]
```

```
## $coefficients
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     17.15       1.12   15.25 1.13e-15
## factor(am)1      7.24       1.76    4.11 2.85e-04
```

```
## Model diagnostics Figure 2
par(mfrow=c(2,2))
plot(lm1)
```



1) Linear relationship between each (numerical) explanatory variables
   In the Simple linear regression there is only 1 categorical variable so we cannot check the residuals plot that residuals are scattered around 0.
2) Nearly normal distribution of residuals
   The qqplot is almost a straight line we can say that this condition is met.
3) Constant variability of residuals
   Since these are randomly scattered around 0, this condition is met.
4) Leverage
   There are no patterns therefore this condition appears to be met

```
# Summary of dfbetas for model1
kable(dfbetas(lm1)[,2],caption ="Dfbetas for Model lm",
      align = c("c", "c"))
```

Table 3: Dfbetas for Model lm

Mazda RX4	-0.159
Mazda RX4 Wag	-0.159

Datsun 710	-0.074
Hornet 4 Drive	-0.133
Hornet Sportabout	-0.048
Valiant	-0.030
Duster 360	0.089
Merc 240D	-0.234
Merc 230	-0.179
Merc 280	-0.064
Merc 280C	-0.020
Merc 450SE	0.023
Merc 450SL	-0.005
Merc 450SLC	0.060
Cadillac Fleetwood	0.216
Lincoln Continental	0.216
Chrysler Imperial	0.076
Fiat 128	0.391
Honda Civic	0.287
Toyota Corolla	0.475
Toyota Corona	-0.137
Dodge Challenger	0.051
AMC Javelin	0.060
Camaro Z28	0.120
Pontiac Firebird	-0.064
Fiat X1-9	0.136
Porsche 914-2	0.075
Lotus Europa	0.287
Ford Pantera L	-0.423
Ferrari Dino	-0.222
Maserati Bora	-0.468
Volvo 142E	-0.140

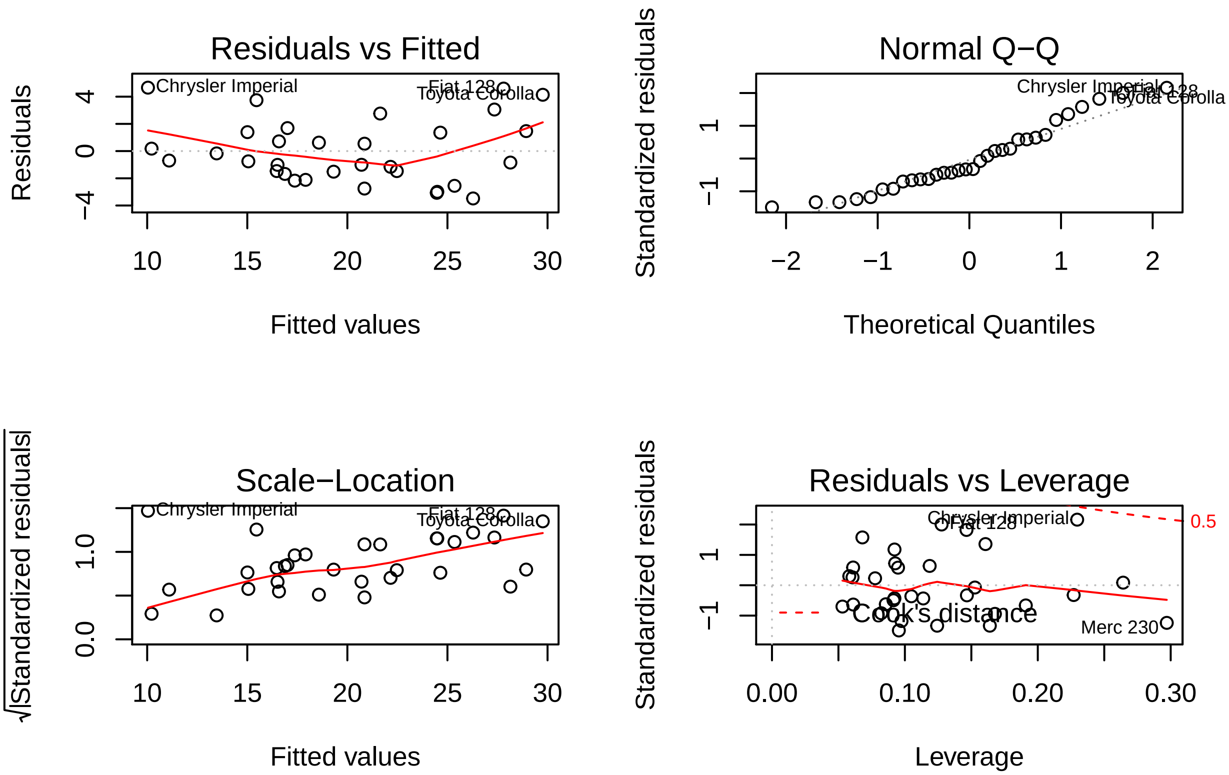
From looking at there at the dfbetas, there do not appear to be any outliers or influence for the first model.

MULTIVARIABLE LINEAR REGRESSION MODEL

```
## Produce a summary of  multivariable linear model using stepwise backwise model selection
summary(sw)[4]
```

```
## $coefficients
##             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
## am              2.94      1.411    2.08 4.67e-02
```

```
## Model diagnostics Figure 3
par(mfrow=c(2,2))
plot(sw)
```



1) Linear relationship between each (numerical) explanatory variables
   These residuals are scattered around 0 so this condition is met.
2) Nearly normal distribution of residuals
   The qqplot is almost a straight line we can say that this condition is met.
3) Constant variability of residuals
   Since these are randomly scattered around 0, but this condition is met.
4) Leverage
   There are no patterns therefore this condition appears to be met.

```
# Take a look at the variance inflation factors
kable(vif(sw),caption= "Table 5 - Variance Inflation factors Model sw",align = c("c", "c"))
```

Table 4: Table 5 - Variance Inflation factors Model sw

wt	2.48
qsec	1.36
am	2.54

We can see that the *cyl*, *disp* and *wt* have high variance in this model.

```
# Summary of dfbetas for sw model
kable(dfbetas(sw)[,2],caption="Table 6 - Dfbetas for sw model",align = c("c", "c"))
```

Table 5: Table 6 - Dfbetas for sw model

Mazda RX4	-0.007
Mazda RX4 Wag	-0.059
Datsun 710	-0.070
Hornet 4 Drive	-0.021
Hornet Sportabout	-0.120
Valiant	-0.030
Duster 360	0.067
Merc 240D	-0.082
Merc 230	-0.126
Merc 280	-0.023
Merc 280C	0.036
Merc 450SE	0.026
Merc 450SL	-0.013
Merc 450SLC	0.005
Cadillac Fleetwood	-0.151
Lincoln Continental	0.045
Chrysler Imperial	1.094
Fiat 128	0.129
Honda Civic	-0.111
Toyota Corolla	-0.051
Toyota Corona	0.407
Dodge Challenger	0.065
AMC Javelin	0.140
Camaro Z28	0.011
Pontiac Firebird	-0.069
Fiat X1-9	0.020
Porsche 914-2	-0.069
Lotus Europa	-0.429
Ford Pantera L	-0.062
Ferrari Dino	0.000
Maserati Bora	-0.132
Volvo 142E	-0.254

From looking at there at the dfbetas, there do not appear to be any outliers or influence for the third model.

Lastly we will compare the two models using anova.

```
anova(lm1,sw)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(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 in favour of an alternate hypothesis that the coefficients are not 0.

OTHER REGRESSION MODELS

We would not use logistic regression since the *mpg* outcome does not have two values but is a numerical outcome. Additionally we would not use Poisson regression since the *mpg* outcome is not a count.