Relationship of variables with miles per gallon

wenlarry

1/4/2017

Executive Summary

An analysis of the relationship of variables with miles per gallon (mpg) using the 'mtcars' dataset. We use exploratory analysis and regression models to a) determine whether an automatic or manual transmission is better for mpg and b) quantify the mpg difference between automatic and manual transmissions.

The results are that manual transmissions have higher mpgs. Our best case (step model) explains 83% of the variance of the mpg with p-values significant at 0.05 level. Also, the step model meets the basic assumptions of a linear regression through a residual plot analysis.

Exploratory Data Analysis

Load data (mtcar) and change some variables to factor class. Also, change 'am' to 2 levels (automatic and manual).

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

Plot 1 in Appendix shows the relationship of the 2parameters (am and mpg). The manual transmissions have higher miles per gallon (mpg). As there could be bias in the dataset, we need to explore what other parameters have higher correlations to mpg than am.

Plot 2 in Appendix shows that mpg has other strong correlations than just 'am'. So a model based on 'mpg' alone is inaccurate.

Regression Models

Use a base model with only 'am' as the predictor.

```
#M1
basefit<-lm(mpg~am,mtcars)
summary(basefit)

##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

M1shows that Adjusted R-squared is 0.3385. This means that the base model explains 34% of the variance of 'mpg'.

Given the need to include more predictors/parameters, we next use a full model with all the parameters

```
#M2
fullfit<-lm(mpg~.,mtcars)
summary(fullfit)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

M2 shows Adjusted R-squared is 0.8066 an improvement to the base model as 81% of the variance of 'mpg ' is explained. However, many of the p-values are not significant at 0.05.

We next use a step model to include significant variables.

```
#M3
stepfit<-step(fullfit, direction="both", trace=FALSE)
summary(stepfit)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

M3 is 'mpg ~ wt +qsec + am'. The Adjusted R-squared is 0.8336. This is an improvement to the full model as 83% of the variance of 'mpg' is explained. Also, all of the p-values are significant at 0.05

Plot 3 in Apendix is a residual plot to test the step model. It shows that:

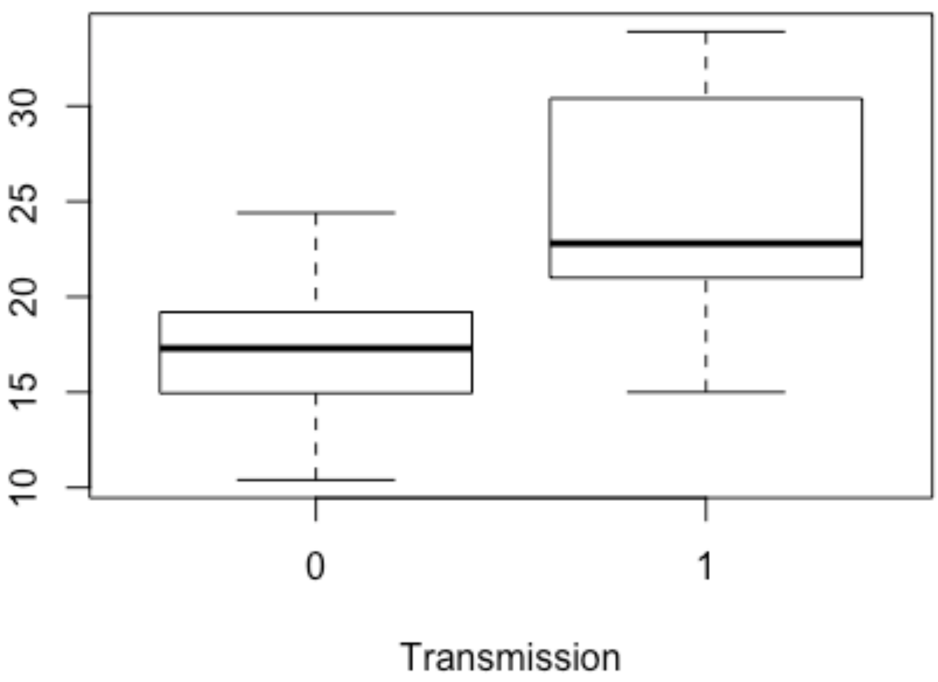
1. Normal Q-Q plot shows that the points are close to the line, indicating a normal distribution
2. Residual vs Leverage plot shows no outliers, with all points within 0.5 bands
3. Residual vs Fitted plot shows no consistent pattern
4. Scale -Location plot shows points that are randomnly distributed confirming the constant variance

Therefore, the basic assumptions of linear regression have been met in the step model (the best case) .

Appendix

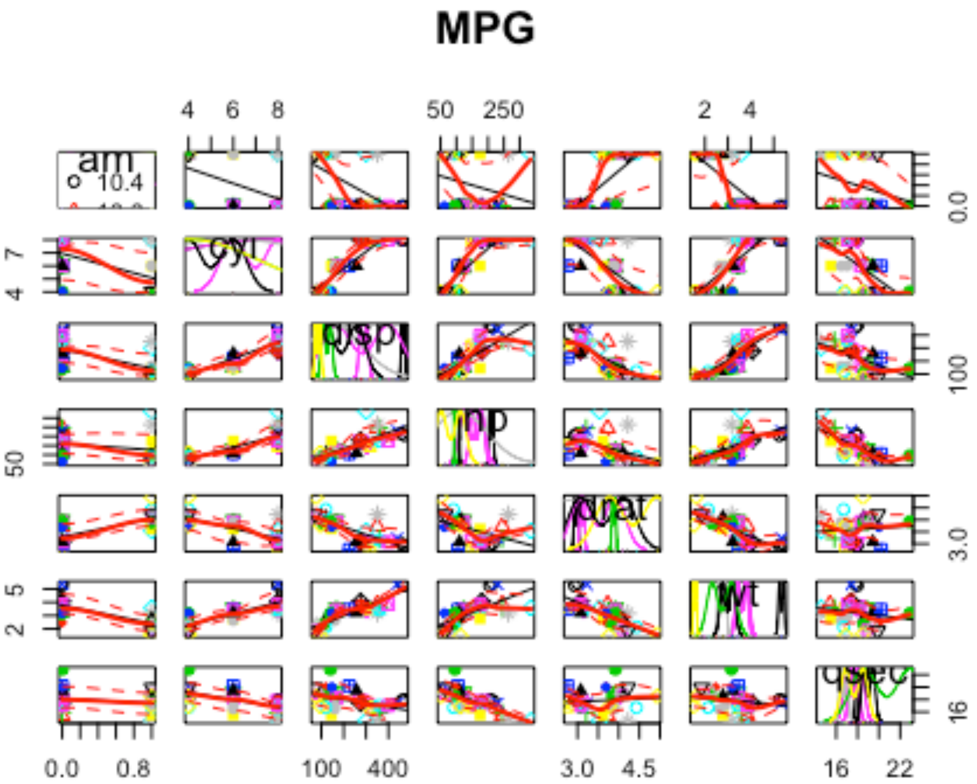
Plot 1 (transmission vs mpg)

```
#Plot 1
boxplot(mpg~am,mtcars,xlab="Transmission")
```



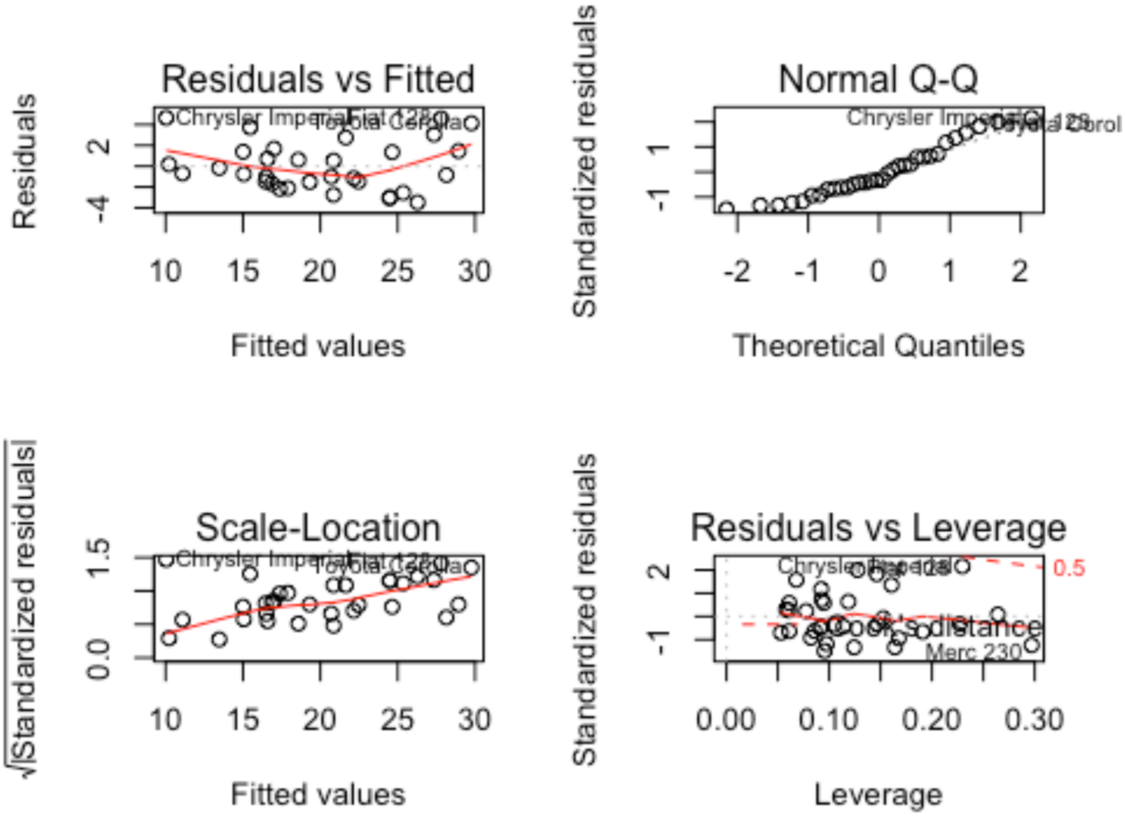
Plot 2 (scatterplot matrix)

```
library(car)
scatterplot.matrix(~am  +cyl+disp  +hp+drat+wt+qsec |mpg, data=mtcars,
main="MPG")
```



Plot 3 (residual plot)

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


Residuals vs Fitted
Residuals
2
-4
10
15
20
25
30
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q-Q
Standardized residuals
1
-1
-2
-1
0
1
2
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Toyota Corol
Scale-Location
√|Standardized residuals|
1.5
0.0
10
15
20
25
30
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals vs Leverage
Standardized residuals
2
-1
0.00
0.10
0.20
0.30
Leverage
Chrysler Imperial
Fiat 128
0.5
Cook's distance
Merc 230

END