Regression Models Course Project

Mohit Shukla

Sunday, May 24, 2015

Summary This report includes an analysis of mtcars data set which is a collection of cars, main interest of exploration is to understand the relationship between a set of variables and miles per gallon (MPG) (outcome). Main focus of the analysis is the two following aspects:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

This analysis may follow many models depending upon the number of covariates being included. Hence A model comparison test is performed to test the relative reliablility of different models. A number of exploratory analysis tasks have been performed to check the validity of models.

The analysis shwoed that all covariates produces more reliable model than single covariate model. Since, a motar car has so many influencing factors other than transmission, so this result seems logical. From including two covariates to including all the covariates, results of the model didn't show much difference. Hence **all covariate model has been selected** over all the others assuming that it represents all the other models and dominates the single covariate model.

Analysis of mpg based on factor variable am

```
data(mtcars)
data <-  mtcars[,c("mpg","am")]
data[,"am"] <-  as.factor(data[,"am"])
attach(data)
summary(mpg[am == 1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

```
summary(mpg[am == 0])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
```

```
fit <- lm(mpg~am)
summary(fit)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```

Above analysis shows dominance of am==1 i.e. Manual transmission. Since the data were coded as 0/1, the coefficient is directly interpreted as the manual (1) - auto (0) difference. In other words, manual on average were 7.245 points greater than auto, a significant difference.

Analysis of mpg based on all the covariates

```r
data(mtcars)
model <- lm(mpg~., data=mtcars)
summary(model)$coef
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

The above model shows that manual transmission on average is 2.5202269 points better than the auto transmission. Now this is serious **conflict** between sigle covariate model and all covariate model. To decide between these two and all the other available models a model selection test is performed below.

Model Selection

```r
data(mtcars)
fit10 <- lm(mpg~ am, data=mtcars)
fit9 <- update(fit10, mpg~  cyl + am, data=mtcars)
fit8 <- update(fit10, mpg~  cyl + disp + am, data=mtcars)
fit7 <- update(fit10, mpg~  cyl + disp + hp + am, data=mtcars)
fit6 <- update(fit10, mpg~  cyl + disp + hp + drat + am, data=mtcars)
fit5 <- update(fit10, mpg~  cyl + disp + hp + drat + wt + am, data=mtcars)
fit4 <- update(fit10, mpg~  cyl + disp + hp + drat + wt + qsec + am , data=mtcars)
fit3 <- update(fit10, mpg~  cyl + disp + hp + drat + wt + qsec + vs+ am , data=mtcars)
fit2 <- update(fit10, mpg~  cyl + disp + hp + drat + wt + qsec + vs+ am + gear, data=mtcars)
fit1 <- update(fit10, mpg~ cyl + disp + hp + drat + wt + qsec + vs+ am + gear + carb, data=mtcars)
anova(fit10,fit9,fit8,fit7,fit6,fit5,fit4,fit3,fit2,fit1)
```

```
## Analysis of Variance Table
##
## Model  1: mpg ~ am
## Model  2: mpg ~ cyl + am
## Model  3: mpg ~ cyl + disp + am
## Model  4: mpg ~ cyl + disp + hp + am
## Model  5: mpg ~ cyl + disp + hp + drat + am
## Model  6: mpg ~ cyl + disp + hp + drat + wt + am
## Model  7: mpg ~ cyl + disp + hp + drat + wt + qsec + am
## Model  8: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
## Model  9: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
## Model 10: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##    Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1      30 720.90
## 2      29 271.36  1    449.53 64.0039 8.231e-08 ***
## 3      28 252.08  1     19.28  2.7452   0.11241
## 4      27 216.37  1     35.71  5.0849   0.03493 *
```

```
## 5      26 214.50  1      1.87  0.2663   0.61121
## 6      25 162.43  1     52.06  7.4127   0.01275 *
## 7      24 149.09  1     13.34  1.8999   0.18260
## 8      23 148.87  1      0.22  0.0309   0.86214
## 9      22 147.90  1      0.97  0.1384   0.71365
## 10     21 147.49  1      0.41  0.0579   0.81218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fit10,fit1)
```

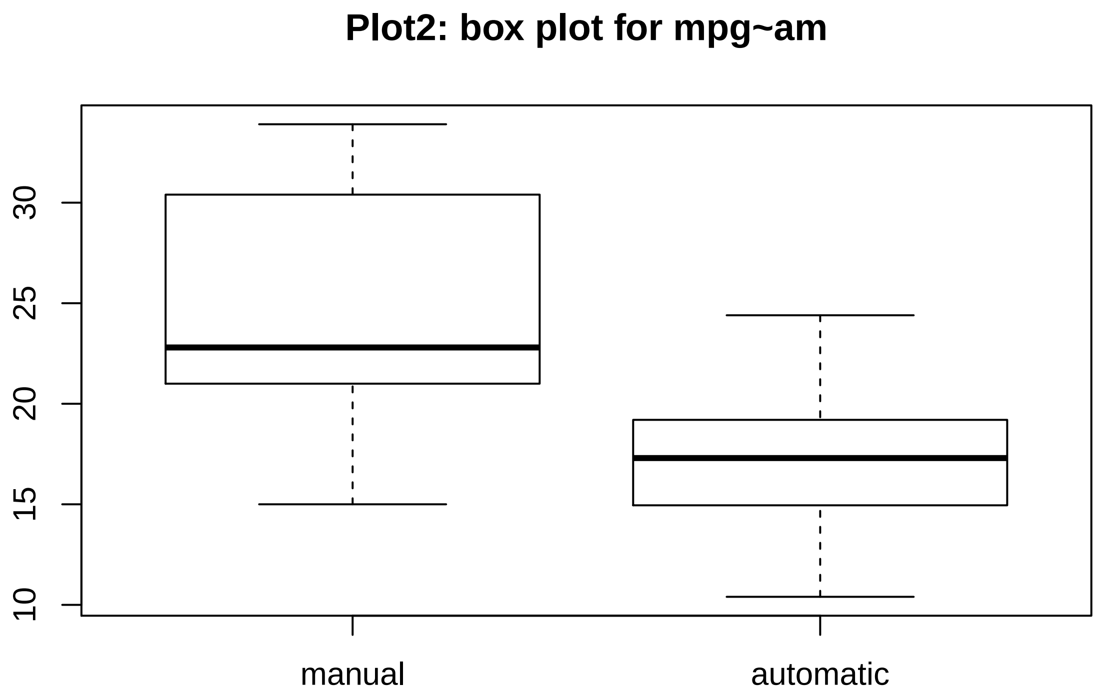
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     21 147.49  9     573.4 9.0711 1.779e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

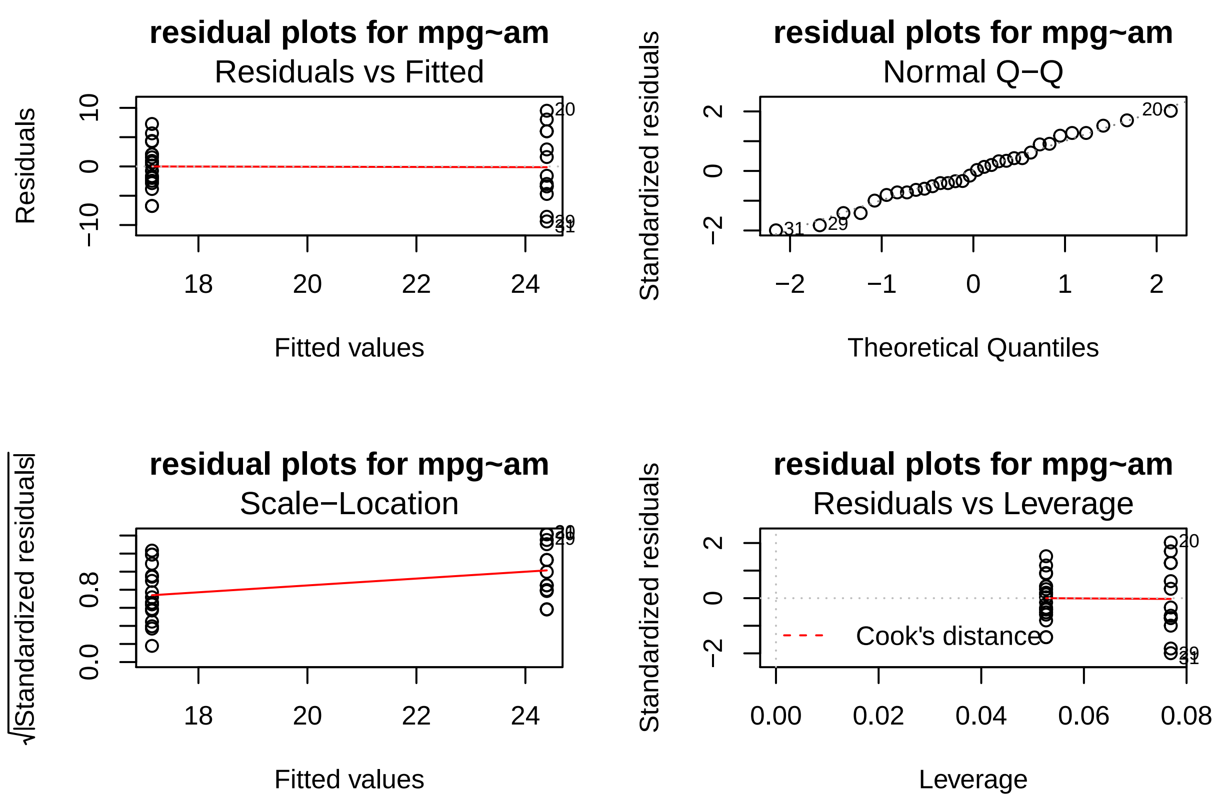
The above analysis shows (by looking at Pr>F values)that all the models from model1 to model9 are better off than model10 which includes only one covariate. Second result shows the dominance of **all covariates model** over singel covariate model.
Hence it is safe to go with model that includes all the covariates i.e. **fit1** model.

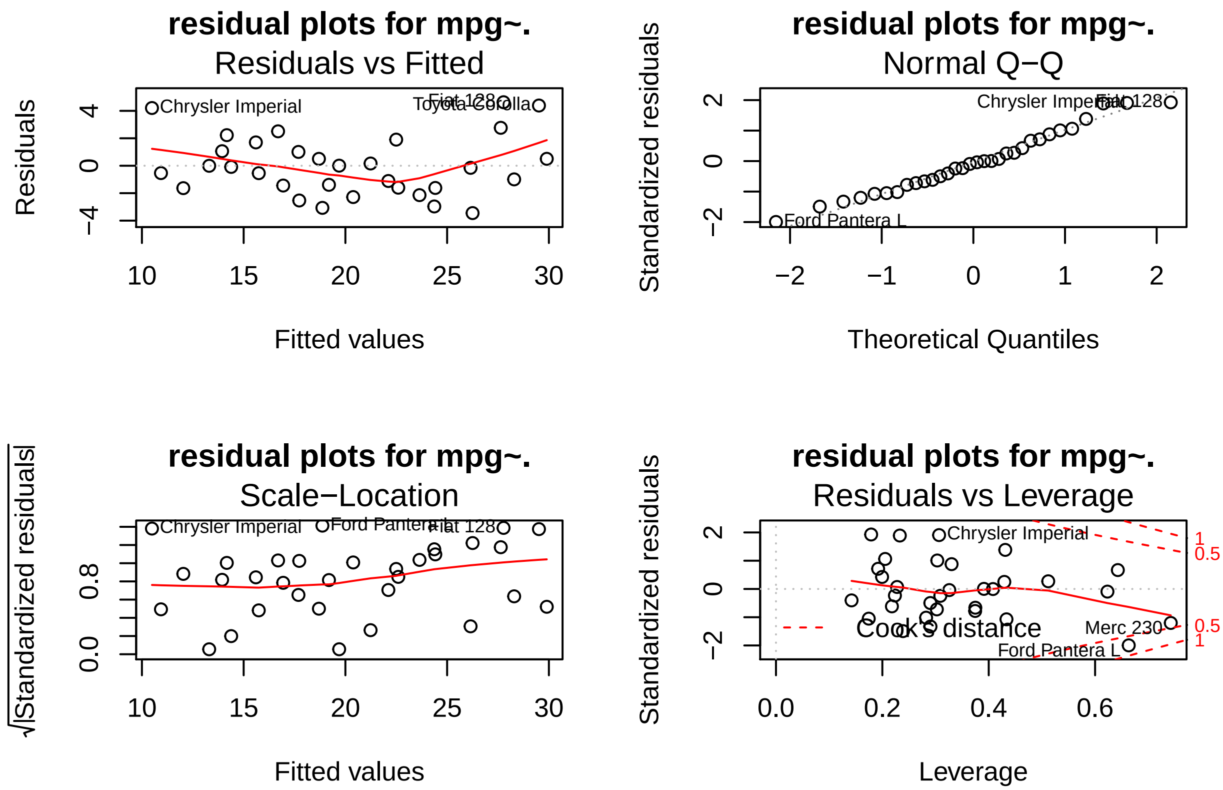
Result

- coefficients for "am" are: 2.5202269, 2.0566506, 1.2254035, 0.2339897
- Manual Transmission (am==1) is 2.5202269 points better than Automatic Transmission i.e. (am==0)

Appendix


Plot2: box plot for mpg~am
10
15
20
25
30
manual
automatic


residual plots for mpg~am
Residuals vs Fitted
Residuals
Fitted values
residual plots for mpg~am
Normal Q−Q
Standardized residuals
Theoretical Quantiles
residual plots for mpg~am
Scale−Location
√|Standardized residuals|
Fitted values
residual plots for mpg~am
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


residual plots for mpg~.
Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
residual plots for mpg~.
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Ford Pantera L
residual plots for mpg~.
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Ford Pantera L
Fiat 128
residual plots for mpg~.
Residuals vs Leverage
Standardized residuals
Leverage
Chrysler Imperial
Cook's distance
Merc 230
Ford Pantera L