Regression analysis of "Motor Trends Car Road Tests" data for best MPG outcome

Author: Srinivasan Sastry Date: 26-Dec-2015

Executive Summary

In this paper, the car data from "Motor Trend Car Road Tests" is analyzed for best outcome of Miles per Gallon based on other variables in the data set. A model fit is done by regressing through all the variables in the data set and choose one that can provide an accurate linearity for the best MPG outcome. The analysis then should lead us to answer the following questions

1. Is an automatic or manual transmission better for MPG
2. Quantify the MPG difference between automatic and manual transmissions

Summary of the data

The "Motor Trend Car Road Tests" (mtcars) dataset is loaded into R and observed. There are 32 observations and 11 variables in the data set including the mpg.Transmission variable 'am', V/S variable 'vs', 'gear' and 'carb' seemed to have ordinal data rather than continous data.

Exploratory analysis of the Mtcars data

In order to understand the data better, a pairwise plots of the correlation is done to indicate if all the variables are independant regressors or there is mutual relationships between them. The pairwise plot is given in the appendix. From the figure, It appears that with the **exception of carb - Number of carburetors, qsec - accelaration for 1/4 mile time and gear - Number of forward gears** are all correlated to mpg. However there are correlations within the other independant variables also. For instance **cyl - Number of cylinders and disp - Displacement (cu.in.) are stongly correlated . Also wt - Weight (lb/1000) is also strongly correlated to displacement**. So we can expect to see a model that will have a few variables removed due to the collinearity.

A box plot is plotted (given in appendix) to understand the relationship between Miles per gallon (mpg) vs the transmission (am). From the plot, it appears that the manual transmission is better for mpg than automatic transmission. In order to understand the coefficients better a model needs to be constructed for the same.

Building a Linear Model

A simple most parsimonious (best) model from candidate sub-models can be be chosen based on one of the following criteria (not exhaustive list).

1.Choose Maximum R-Square from candidate sub-models

2.Choose Maximum Adjusted-R-Square from candidate sub-models

3.Choose Minimum Mallows' Cp from candidate sub-models

4.Choose Minimum AICp from candidate sub-models

5.Choose Minimum BICp from candidate sub-models

6.Choose Minimum PRESSp from candidate sub-models

The model also can be achived with step-wise regression of variables through automated procedures like step or leaps or manually through "Forward" or "Backward" addition or elimination of variables by minimizing p-values or Maximizing the F-Values.

For the mtcars data set removing unnecessary regressors through `manual backward elimination by choosing residual p-values greater than a threshold (0.05) yielded a good model than by minimizing the "VIF". The function`constr_model_func‘ does this job. The progressive elimination is given in the table below from the full model.

Full <- lm(mpg ~ cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb, data = mtcars)

```
Full <- lm(mpg ~ cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb, data = mtcars)
mdl <- constr_model_func(Full)
```

```
## Removed cyl With Pr(>|t|) value 0.9161
## Removed vs With Pr(>|t|) value 0.8433
## Removed carb With Pr(>|t|) value 0.747
## Removed gear With Pr(>|t|) value 0.6196
## Removed drat With Pr(>|t|) value 0.4624
## Removed disp With Pr(>|t|) value 0.299
## Removed hp With Pr(>|t|) value 0.2231
```

```
summary(mdl)$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
wt	-3.916504	0.7112016	-5.506882	0.0000070
qsec	1.225886	0.2886696	4.246676	0.0002162
amManual	2.935837	1.4109045	2.080819	0.0467155

Validating or testing the Model

The model can be tested either building a new model through an automated process and the same model can be compared with the manual model based on p-statistics, F-statistic etc as described in the previous section. Alternatively regressors can be added to the model that was manually built and statistics checked using `anova`.The following section does automated step wise linear regression to see if any coefficients were missed or added incorrectly

```
stpMdl <- step(Full,direction= "both",trace=0)
summary(stpMdl)$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
wt	-3.916504	0.7112016	-5.506882	0.0000070
qsec	1.225886	0.2886696	4.246676	0.0002162
amManual	2.935837	1.4109045	2.080819	0.0467155

The automated procedure also arrived at the same cooeficients as the manual model, indicating that the

model manually obtained is one of the best possible models for this data set with 'mpg' as the outcome

The second test would be to iteratively add coefficients and see if the model obtained is more significant. Three models built are shown below.

Model 1: mpg ~ wt + qsec + am

Model 2: mpg ~ wt + qsec + am + cyl

Model 3: mpg ~ wt + qsec + am + cyl + disp

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
28	169.2859	NA	NA	NA	NA
27	167.7849	1	1.501058	0.2417852	0.6270456
26	161.4139	1	6.370929	1.0262072	0.3203861

Thus addition of coefficients bring the p-value higher thus we would `accept the NULL hypothesis that the coefficients of the variables added are zero showing the Model 1 is best so far considered`.Hence it can be ascertained that mpg is dependent on wt,qsec and am(Manual).

Relationship between Manual and Automatic Transmission

In the model building the best model obtained was in relation to Manual transmission. However in order to understand the relation between the Manual and automated transmission, following model is built

mpg ~ wt+qsec+I(1 * (am == 'Automatic')) + I(1 * (am == 'Manual'))

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.553618	6.0573391	2.072464	0.0475434
wt	-3.916504	0.7112016	-5.506882	0.0000070
qsec	1.225886	0.2886696	4.246676	0.0002162
I(1 * (am == "Automatic"))	-2.935837	1.4109045	-2.080819	0.0467155

It is evident from the coefficients of the model, that holding other variables constant, automatic transmission offers `-2.935837` mpg less than the Manual transmission.

Diagnostic plots

In order to understand the influence of the outliers or residuals,residual plots are generated. These are shown in the appendix. In the residual plot residuals are seem equally distributed and the mean close to zero.(Value shown in Appendix).

There are 3 outliers point 17, 18 and 20 and one big influencer point 9. Once an subject area understanding of the data is obtained,(i.e is this outliers real or due some spurious processes), the treatment of this data can be established.

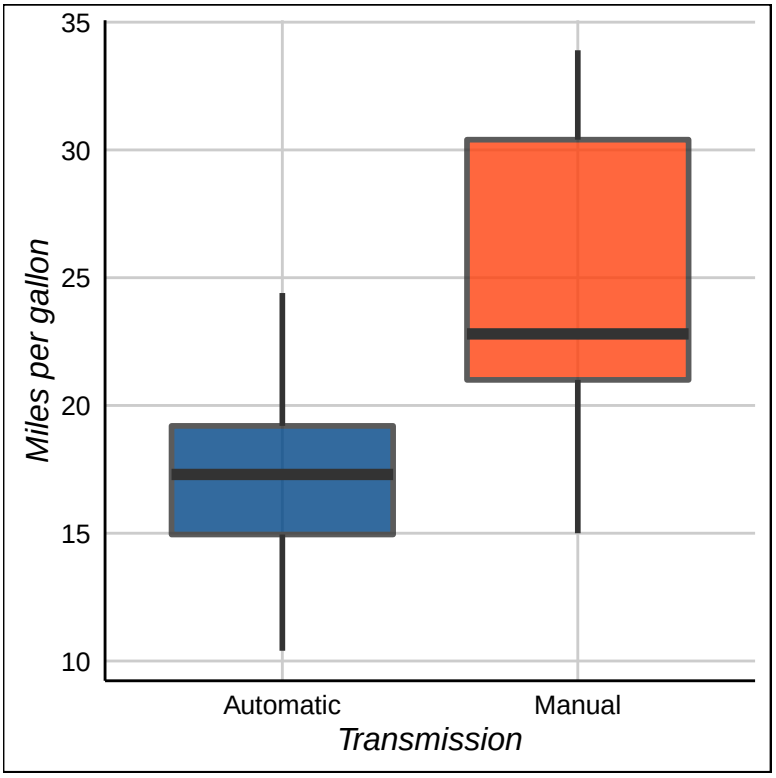
Final Inferences based on the model

1. The model fitted is as follows: `mpg = 12.553618 -3.916504 wt + 1.225886 qsec -2.935837 am (Automatic)`.The confidence interval for the coefficients of the model are given in Appendix.
2. We see that approximately for every 3000 lb there is a decrese of one mpg in the mileage for automatic transmission

3. Automatic transmission offers -2.935837 mpg less than the Manual transmission.

Appendix

Fig 1: Mpg vs Transmission




```
constr_model_func<-function(Full_Mod,thresh=0.05,trace=T,...){
        Mod_init<-NULL
        thresh_not_reached <- TRUE
        New_Mod <- Full_Mod
        while (thresh_not_reached) {
                #get the max  p-value for  regressors
                j <- summary(New_Mod)$coeff[,'Pr(>|t|)']
                #remove intercept
                j <- j[names(j) != '(Intercept)']
                k <- max(j)

                if (k < thresh) {break}

                # else remove the regressor from the model
                remove_regressor <- names(which.max(summary(New_Mod)$coeff[,'Pr(>|t|)']))
                cat(paste("Removed",remove_regressor,"With Pr(>|t|) value",round(k,4)),'\n')
                flush.console()
                # construct new model
                form <- NULL
                form <- paste('. ~ . -', remove_regressor)
             New_Mod <- update(New_Mod,  form)
        }
```

```
        return (New_Mod)
}
```

Fig 2: Mtcars- Correlation between variables

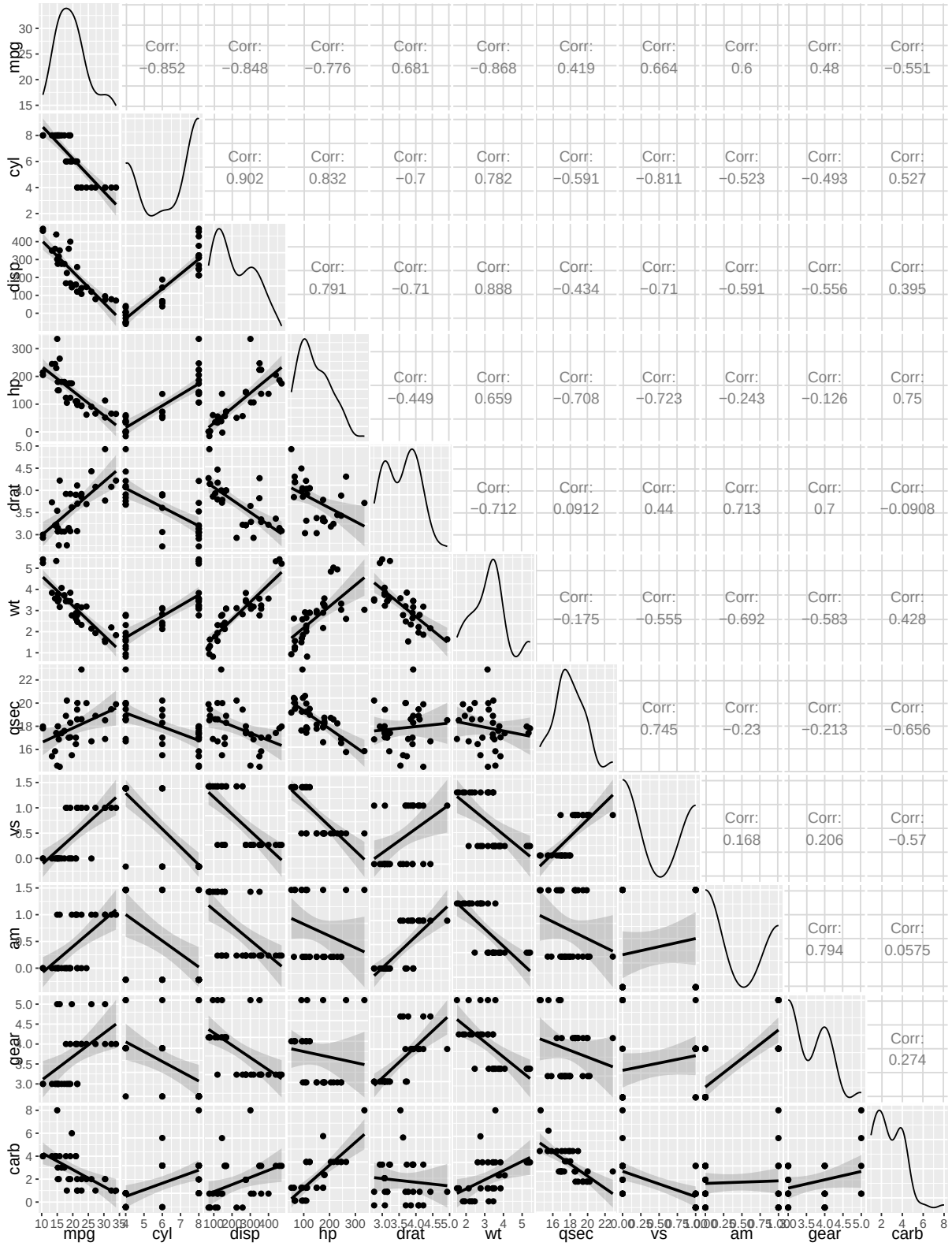


Fig 3: Model diagnostic plots

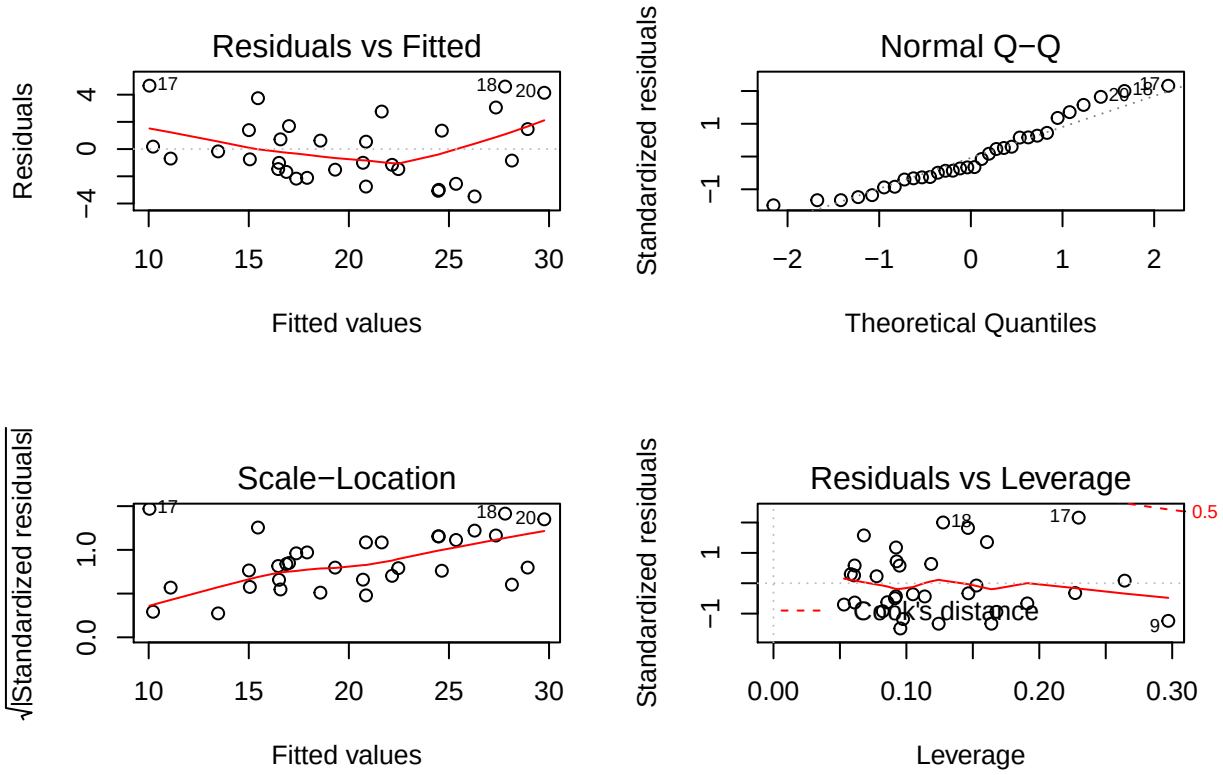


Fig 4: Confidence interval for the model

	2.5 %	97.5 %
(Intercept)	0.1457211	24.9615143
wt	-5.3733342	-2.4596732
qsec	0.6345732	1.8171987
I(1 * (am == “Automatic”))	-5.8259441	-0.0457303
I(1 * (am == “Manual”))	NA	NA

Fig 5: Mean of residuals

```
## [1] "Mean of the residuals from the model 3.7470027081099e-15"
```