Regression_Analysis_Project

Anurag Ladage

July 24, 2014

EXECUTIVE SUMMARY

This project makes in depth analysis on the variable relationships in the 'mtcars' dataset from the datasets library by performing a multivariate regression analysis. The first part of the report consists of an exhaustive model selection procedure. Then we move on to statistically prove that the miles per gallon (mpg) performance parameter is different for automatic transmission and manual transmission. Additionally, we also go on to quantify the difference in mpg variable for both the transmission modes.

Note: All the supporting figures and tables are in the appendix

IMPORTING DATA AND PERFORMING EXPLORATORY ANALYSIS:

```
mydata <- mtcars;#pairs(mydata) - Refer to Fig(1)
```

The correlation matrix was also analysed but to avoid redundancy it is not included in the report.

MODEL SELECTION

First we just fit a linear model using all the input variables to get an overview of the relationships.

```
fit <- lm(mpg~.,data=mydata)
```

In the above data there is no significant regressor except may be weight that seems to be affecting the mpg variable. This makes us question the validity of the model and creates a need to filter out the extra variables we might have added in. It might we safe to assume for now that the weight and the trasmission factor will be included in the filtered regression model.

Note: One general rule that can be followed during any regression analysis from my experience is that if the main effects of a variable is statistically insginificant then all its higher level interactions will also be insignificant. Thus, we examine and reduce the our model first using only main effects and get a good fit. Once this is done we can then include the interactions if necessary.

Initial Variable Elimination Strategy:

Transmission('am') and weight('wt') will definitely included in the model. Now, we eliminate 'cyl' variable from the model as it is very highly correlated to weights (wt) and seems to be redundant. Further 'hp' and 'disp' are also highly correlated. As the correlation of 'disp' with 'mpg'(-0.846) is more than that of 'hp' with 'mpg'(-0.776) for now we decide to include 'disp'. Additionally, the 'drat' variable also shows high correlation with 'disp'(-0.71). Thus, for now we go ahead with 'disp' and eliminate 'hp' and 'drat'. Note that at this point this is only a judgement call and later on we can compare the models with likelihood ratios test to prove the correctness of the factors we have excluded.Elimination of 'qsec' is unclear at this stage as the its correlation with 'disp'(-0.434) and 'wt'(-0.175) is low. Hence, we decide it to include in the model for now. Similarly we eliminate other factor variables keeping only 'am' and 'vs' in the model.

Stepwise model selection and regressor variable elimination:

```
fit1 <- lm(mpg~disp+wt+factor(vs)+qsec+factor(am),data=mydata)
fit2 <- update(fit1,mpg~hp+drat+wt+factor(vs)+qsec+factor(am))
#anova(fit1,fit2) - Refer to Fig(2)
```

The first model was analysed and the regressors with the their co-efficents having the p-values greater than 0.05 were eliminated. The reason for this being that if the p-values are greater than alpha we fail to reject the null hypothesis that the regression co-efficient is equal to zero. Further, likelihood ratio test is performed to confirm the exclusion of both the disp and drat variable. The p-value for the anova is again greater than 0.05 which proves our initial asusmptions right. Our model now has the following regressors.

```
fit3 <- lm(mpg~wt+qsec+factor(am),data=mydata);#summary(fit)$coefficients[,4] - Refer to fig(3)
```

From the above displayed p-values we can see that all regression co-efficients significant. Further we perform an likelihood ratio test on the model without interactions and with interactions which proves that indeed the interaction effects need to be accounted for.

```
fit4 <- update(fit3,mpg~wt+qsec+factor(am)+wt*qsec*factor(am))
#anova(fit3,fit4) - Refer to Fig(4)
```

After analysing further and eliminating the insignificant 3 way interactions and some 2 way interactions our final model looks as below:

```
final_fit <- lm(mpg~wt+qsec+factor(am)+wt*factor(am),data=mydata)
#summary(final_fit) - Refer to Fig(5)
```

CONFIDENCE INTERVALS AND RESIDUAL ANALYSIS

To answer both the questions asked by the Motor Trend we further produce confidence intervals for the regression co-efficients at both factor levels to account for the uncertainty.

Note: The intercept shows the effect of automatic transmission and (intercept+factor(am1)) value shows the effect of manual transmission.

```
conf_automatic = coef(final_fit)[1] + c(1,-1)*qt(0.975,df=summary(final_fit)$df[2])
conf_manual = coef(final_fit)[1]+coef(final_fit)[4] + c(1,-1)*qt(0.975,df=summary(final_fit)$df[2])
#par(mfrow=c(2,2));plot(final_fit) - Refer to Fig(6)
```

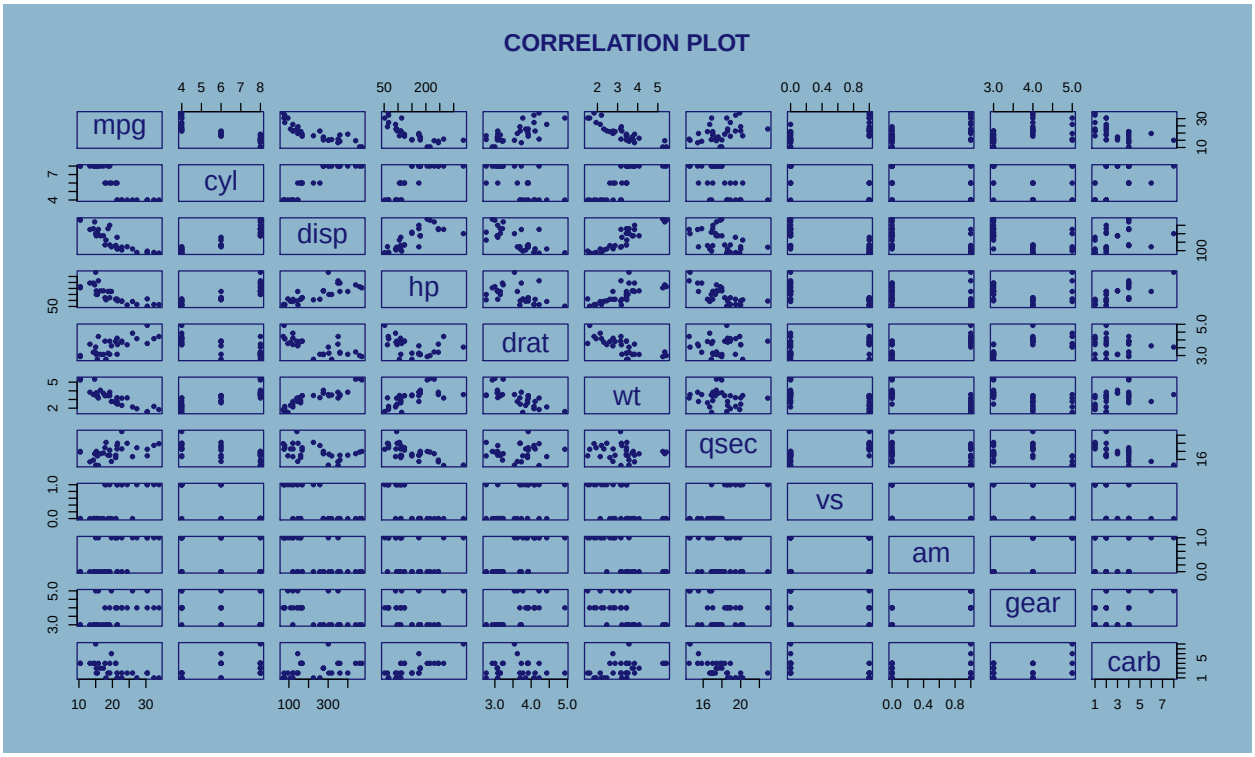
On observing the residual plots there is no noticeable pattern in residuals vs fitted plots which might suggest non-linearity. Additionally, the normal Q-Q plot shows the residuals to be normally distributed with a few outliers. The residuals vs leverage plot does show some outliers which is acceptable. For a more robust model the rlm function from the library MASS can be used.

CONCLUSION:

It can statistically be proven that manual transmission is better than automatic transmission for MPG as the confidence intervals for both the co-efficients donot overlap. For a confidence interval of 95% the range of MPG for manual transmission is [25.85431, 21.75065] and that for automatic transmission is [11.774883, 7.671222].

APPENDIX:

Fig(1)



Fig(2)

```
## Analysis of Variance Table
##
## Model 1: mpg ~ disp + wt + factor(vs) + qsec + factor(am)
## Model 2: mpg ~ hp + drat + wt + factor(vs) + qsec + factor(am)
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     26 166
## 2     25 159  1      7.38 1.16   0.29
```

Fig(3)

```
## (Intercept)         cyl        disp          hp        drat          wt
##     0.51812     0.91609     0.46349     0.33496     0.63528     0.06325
##        qsec          vs          am        gear        carb
##     0.27394     0.88142     0.23399     0.66521     0.81218
```

Fig(4)

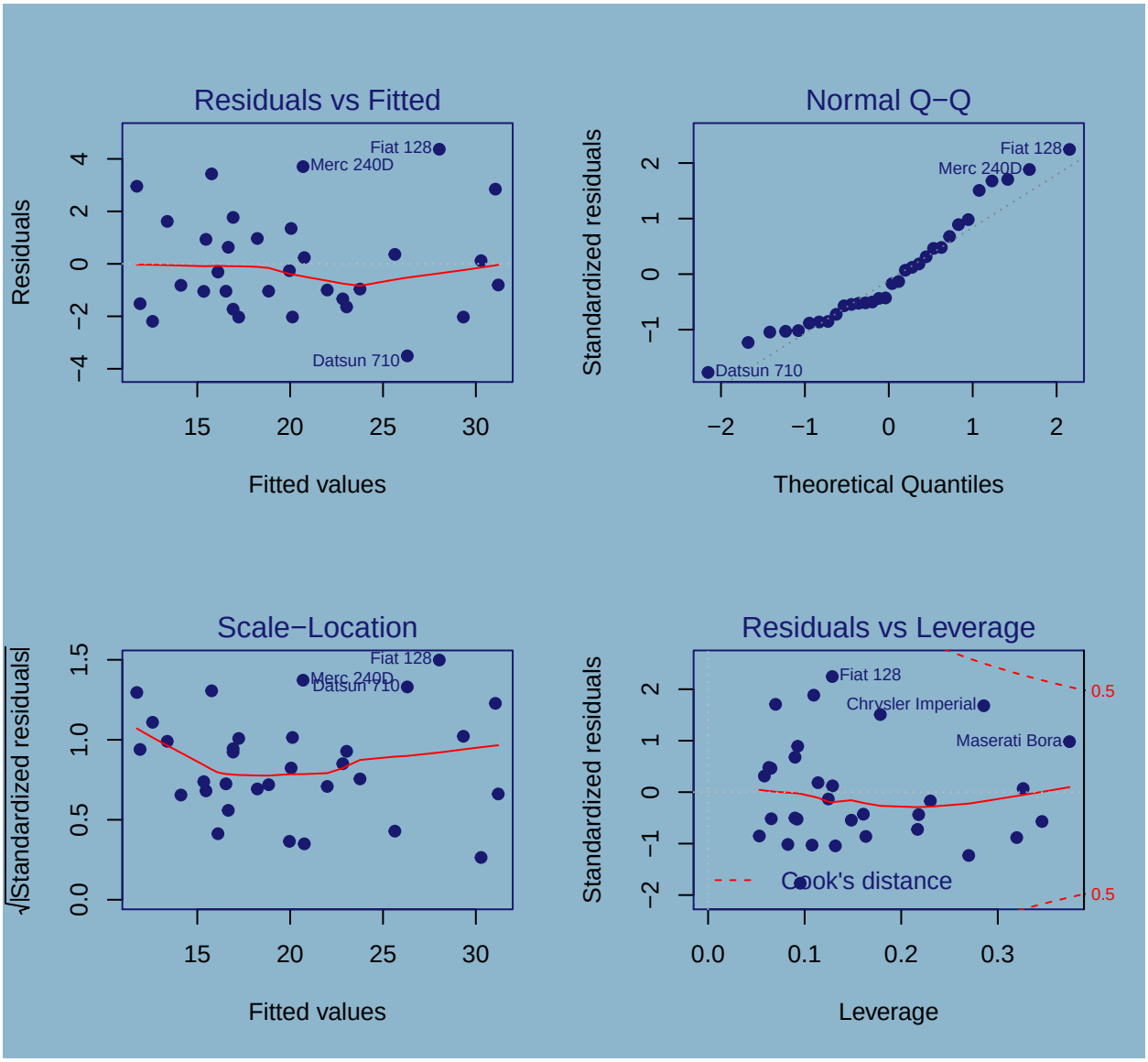
```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + qsec + factor(am)
## Model 2: mpg ~ wt + qsec + factor(am) + wt:qsec + wt:factor(am) + qsec:factor(am) +
##     wt:qsec:factor(am)
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
```

```
## 1     28 169
## 2     24 108  4      61.1 3.39  0.025 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig(5)

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + factor(am) + wt * factor(am),
##     data = mydata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.508 -1.380 -0.559  1.063  4.368
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.723      5.899    1.65  0.11089
## wt               -2.937      0.666   -4.41  0.00015 ***
## qsec              1.017      0.252    4.04  0.00040 ***
## factor(am)1      14.079      3.435    4.10  0.00034 ***
## wt:factor(am)1   -4.141      1.197   -3.46  0.00181 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.08 on 27 degrees of freedom
## Multiple R-squared:  0.896,  Adjusted R-squared:  0.88
## F-statistic: 58.1 on 4 and 27 DF,  p-value: 7.17e-13
```

Fig(6)


Residuals vs Fitted
Residuals
Fitted values
Fiat 128
Merc 240D
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Fiat 128
Merc 240D
Datsun 710
Scale−Location
√|Standardized residuals|
Fitted values
Fiat 128
Merc 240D
Datsun 710
Residuals vs Leverage
Standardized residuals
Leverage
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5