Regression Models - Course Project

Executive Summary

Analysis to answer the following two questions based on a motor vehicle data set.
1) "Is an automatic or manual transmission better for MPG"
2) "How different is the MPG between automatic manual transmission?"
Several regression models were evaluated, compared for accuracy and used to answer the above questions.

Results

Detailed analysis with the dataset revealed we cannot find a statistically significant effect of transmission type on gas mileage(mpg).

The slope coefficient of am is 2.08 with a **p value of 0.14** in a model using wt,hp and am. This p value is not enough to to reject the null hypothesis that transmission has no effect on mpg. If the p-value had been found to be lower than our significance level this slope of 2.08 would represent improvement in gas mileage of manual transmission over automatic transmission vehicles, answering question 2. However this slope currently does not mean anything significant.

Weight and horsepower were found to be the two most significant factors in estimating mpg with p values close to zero.

Description of analysis done

Initially a multivariable regression was done with all available data to serve as a baseline. This analysis is not ideal as it does not address multicollinearity, and includes several variables that ought not to affect mpg.

Model Selection

From studying several semesters of automobile engineering during B.Tech(Mechanical) I can classify the factors as YES/NO/MAYBE [Expert Opinion]

NO: Axle ratios, Gear/Carburetor/Cylinder count, V/Straight engine layout, will not affect mpg.
YES: Weight and horsepower(hp) definitely affect mpg. To produce more power more gas is burnt. Energy expended to move a Weight is directly proportional to its Weight.
MAYBE: Quarter mile time can be considered but would be expected to show collinearity with hp. Engine displacement(size) only has an indirect effect on power and not a very linear one. However it was evaluated.

Exploratory Analysis A pairs plot of the factors considered possibly relevant is shown in appendix
This shows a high negative correlation between mpg vs weight, displacement & hp. Displacement is also highly correlated with weight[0.89] and hp[0.79]

Four additional models were constructed. The model names reflect the factors they include
wthp | wthpam | wthpqsecam | wthpqsecamdspl

For instance wthp uses wt and hp, wthpam uses wt,hp and am and so on.

Adjusted R Square of different models

```
##           all          wthp         wthpam     wthpqsecam wthpqsecamdspl
##     0.8066423     0.8148396     0.8227357      0.8367919      0.8375334
```

Adjusted r square values showed improvement in using wt&hp over using all variables. Adding in am and qsec also improved adj.r.square but adding displ barely improved it.

p value for the F statistic of the fitted models to examine model significance

```
##            all           wthp         wthpam     wthpqsecam wthpqsecamdspl
##   3.793152e-07   9.109054e-12   2.907872e-11   4.589395e-11   1.843717e-10
```

The model with just wt and hp seems most significant from its lowest p value. This implies that Weight and Horsepower seem to explain most of the variation in mpg.

Evaluate collinearity using Variance Inflation Factors

```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

cyl, disp and weight showed the highest variance inflation factors indicating possible collinearity. This also makes intuitive sense since larger displacement engines often have more cylinders, are heavier and are typically installed in heavier vehicles. In the model with all variables cyl and displ coefficients have p values of 0.91 and 0.46, indicating they are not significant. Remove the models using cyl and dspl from consideration.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + hp
## Model 2: mpg ~ wt + hp + am
## Model 3: mpg ~ wt + hp + qsec + am
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     29 195.05
## 2     28 180.29  1    14.757 2.4892 0.12628
## 3     27 160.07  1    20.225 3.4115 0.07573 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the anova results, p value for adding am to wt+hp [0.12590] is higher than our significance level - adding am to a model having wt+hp does not improve the model significantly. Adding qsec to the wthpam model did not improve it either [p value 0.076].

Model coefficients

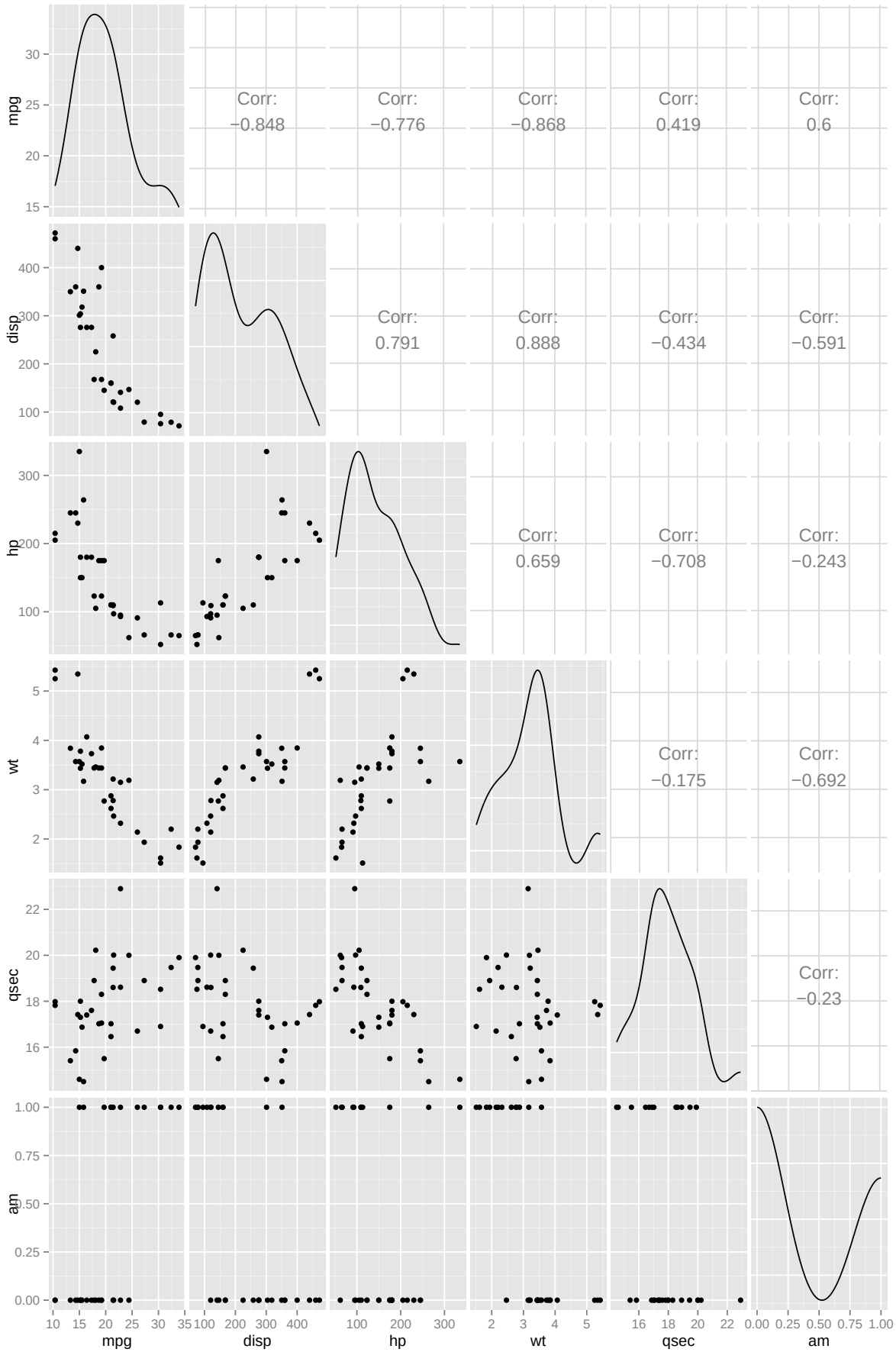
```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
## am           2.08371013 1.376420152  1.513862 1.412682e-01
```

The p value for am is 0.14 which is much much higher than permissible(0.01 or even 0.05). **Based on the p-value and anova results we conclude that the trasmission type has no significant effect on gas mileage.**

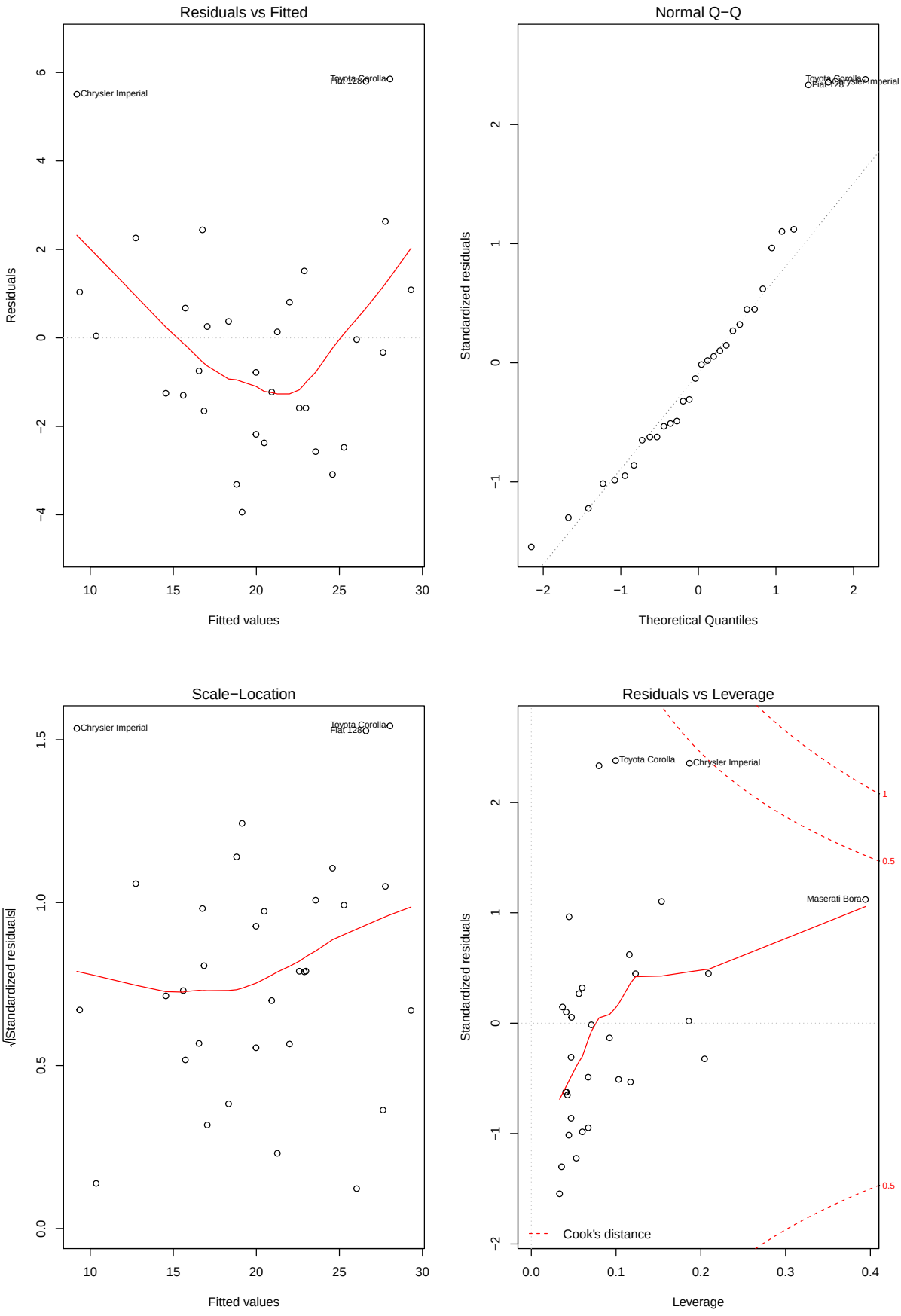
wt and hp have a significant effect on mileage a model having just wt and hp is finalized.

Model residuals do not show any pattern in their scatter which is good. The residuals do show some skew in their distribution but for such a small sample size we can ignore this.

Appendix


mpg
disp
hp
wt
qsec
am
Corr: −0.848
Corr: −0.776
Corr: −0.868
Corr: 0.419
Corr: 0.6
Corr: 0.791
Corr: 0.888
Corr: −0.434
Corr: −0.591
Corr: 0.659
Corr: −0.708
Corr: −0.243
Corr: −0.175
Corr: −0.692
Corr: −0.23

Residuals of model with weight and horsepower



Distribution of residuals