Analysis on the Mtcars Data Set

December 23, 2015

SUMMARY

The company is interested in whether the automatic or manual transmission is better for MPG. To address this, the Motor Trend Car Road Data Set is used. The final result of the study shows that we cannot actually tell whether the automatic or the manual transmission is better for mileage as the mileage is also dependent on other factors such as the number of cylinders and the displacement of the automobile.

EXPLORATORY ANALYSIS

The data set was from the 1974 Motor Trend US Magazine comprises of fuel consumption (mpg) and 10 aspects of automobile performance for 32 automobiles. These aspects are: number of cylinders(cyl), displacement(disp), gross horsepower(hp), rear axle ratio(drat), weight(wt), 1/4 mile time(qsec), V/S(vs), transmission(am), number of forward gears(gear) and number of carburetors (card). For more information, this data set is also accessible in R built-in datasets.

The data set is loaded in R using the following commands. Based on the summary of the data set, `cyl`, `vs`, `am`, `gear` and `carb` columns are numeric/continuous despite that they should be factored variables. This should be noted so that the correct analysis will be given.

```
data("mtcars");summary(mtcars);cars <- mtcars;cor(cars)
```

To find possible confounders, the `cor` function is used in the data set. Ignoring all factored columns as they affect the correlation result, `hp` and `wt` seems to have a strong correlation with `disp` (see Appendix, Figure 1). Also, `wt` and `qsec` have strong correlation with `hp` (see Appendix, Figure 2). And `drat` and `wt` have strong correlation with each other (see Appendix, Figure 3). Hence, some of these variables should be eliminated from the model.

Furthermore, the `cor` function result shows that `cyl`, `disp`, and `wt` have the strongest correlations to `mpg`.

MODEL SELECTION

The company needs to determine whether MPG is better in manual than automatic. The following are different models that represent this relationship: the simplest of which is that mpg is dependent on transmission. The next models after it are included to see which variables significantly impact mpg.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp
## Model 3: mpg ~ am + cyl + disp + wt + hp
## Model 4: mpg ~ am + cyl + disp + wt + hp + drat + qsec
## Model 5: mpg ~ am + cyl + disp + wt + hp + drat + qsec + vs + gear
## Model 6: mpg ~ am + cyl + disp + wt + hp + drat + qsec + vs + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 252.08  2    468.82 33.3746 3.015e-07 ***
## 3     26 163.12  2     88.96  6.3331  0.007043 **
```

```
## 4     24 149.09  2     14.03  0.9988  0.385162
## 5     22 147.90  2      1.19  0.0846  0.919167
## 6     21 147.49  1      0.41  0.0579  0.812179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result shows that adding `cyl` and `disp` on the model is highly significantly. It is also parsimonious. Hence, `fit1` is selected as the model for the data.

RESIDUALS

The model selected is that of `mpg` is dependent on `am`, `cyl` and `disp`. The residual of this model is plotted on Appendix figure 4 and it shows that there is no certain pattern in the plot. Thus, it can be concluded that residuals were independently and almost identically distributed with mean zero. Furthermore, using the `hatvalues` function and `dfbetas` function, it seems that there is no outlier in the data. The residuals are also tested for their normality. Using `shapiro.test` function, it can be said that the residuals are normally distributed.

```
##
##  Shapiro-Wilk normality test
##
## data:  fit1$residuals
## W = 0.95392, p-value = 0.1861
```

SELECTED MODEL

```
mdl <- lm(mpg~am+cyl+disp,cars)
summary(mdl)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 27.83036408 1.96997395 14.127275 5.441362e-14
## am1          1.63580341 1.31740800  1.241683 2.250322e-01
## cyl6        -4.67997453 1.63631225 -2.860074 8.073209e-03
## cyl8        -5.05008177 2.86682804 -1.761557 8.946573e-02
## disp        -0.02241283 0.01122392 -1.996881 5.600984e-02
```

The following is the model for the data:

$$mpg = 27.83 + 1.64am1 - 4.68cyl6 - 5.05cyl8 - 0.02disp$$

where `am1` is 0 when the automobile is automatic and 1 if it is manual. `cyl6` is 1 if there is 6 cylinders in the automobile and 0 if otherwise. `cyl8` is 1 if there is 8 cylinders in the automobile and 0 if otherwise. Furthermore, for every increase in the displacement there is a decrease in `mpg` by 0.02 when other variables are held fixed.

CONCLUSION

The resulting model has three independent variables: the transmission, the number of cylinders in the automobile and the displacement. With this, it cannot be actually concluded whether the mpg of the automatic is better than the manual or vice versa, as the mileage of the automobile is dependent on other factors and they are changing.

APPENDIX

Figure 1: Correlation of Horsepower and weight to displacement

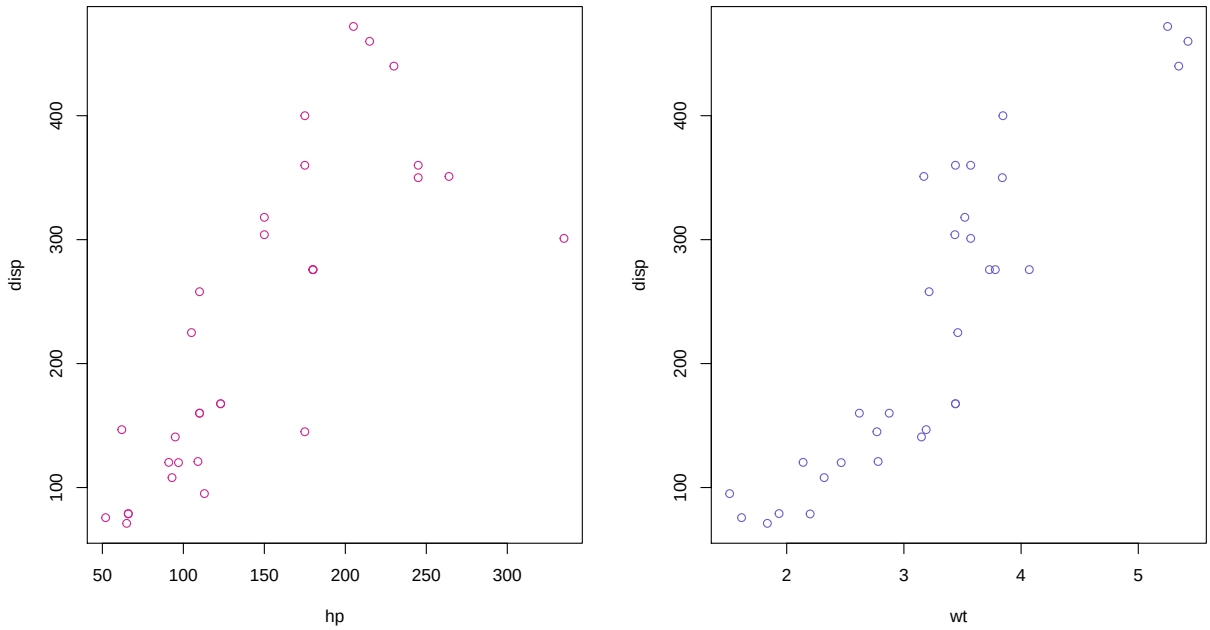

Figure 2: Correlation of Mile Time and Weight to Horsepower

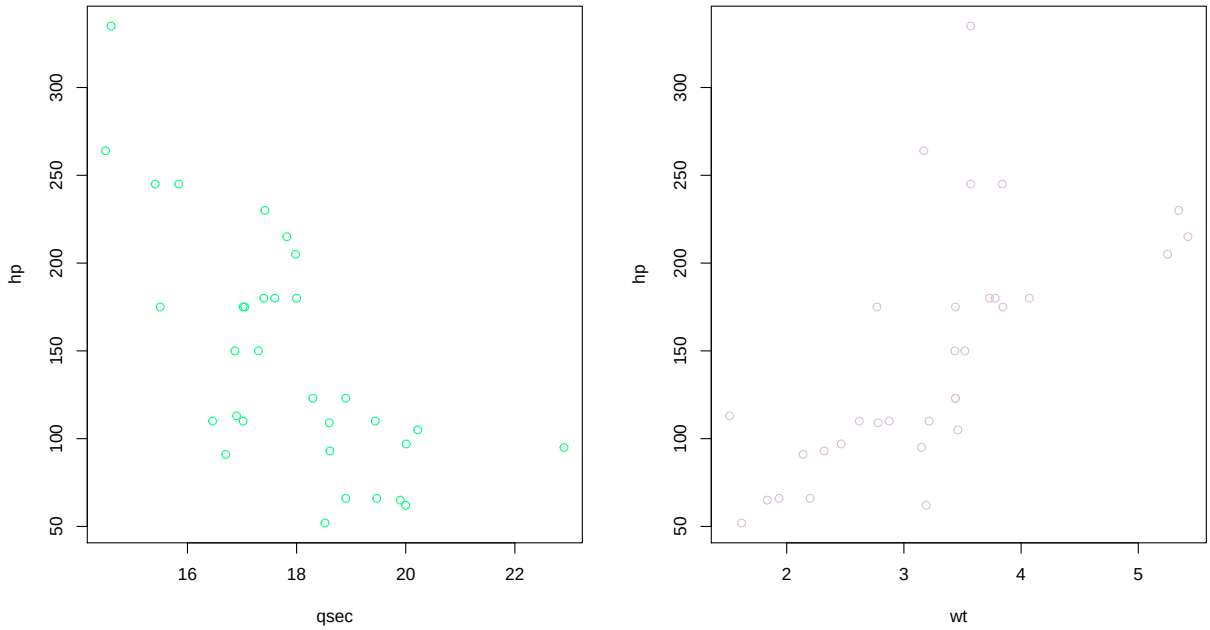

Figure 3: Correlation of Weight to Rear Axle Ratio

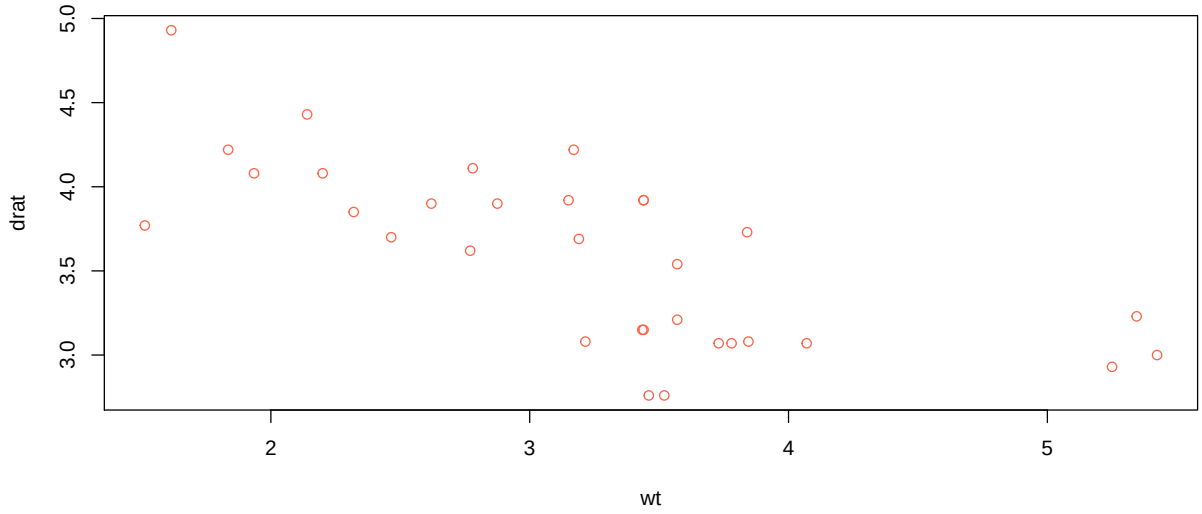

Figure 4: Residuals vs. Fitted Value

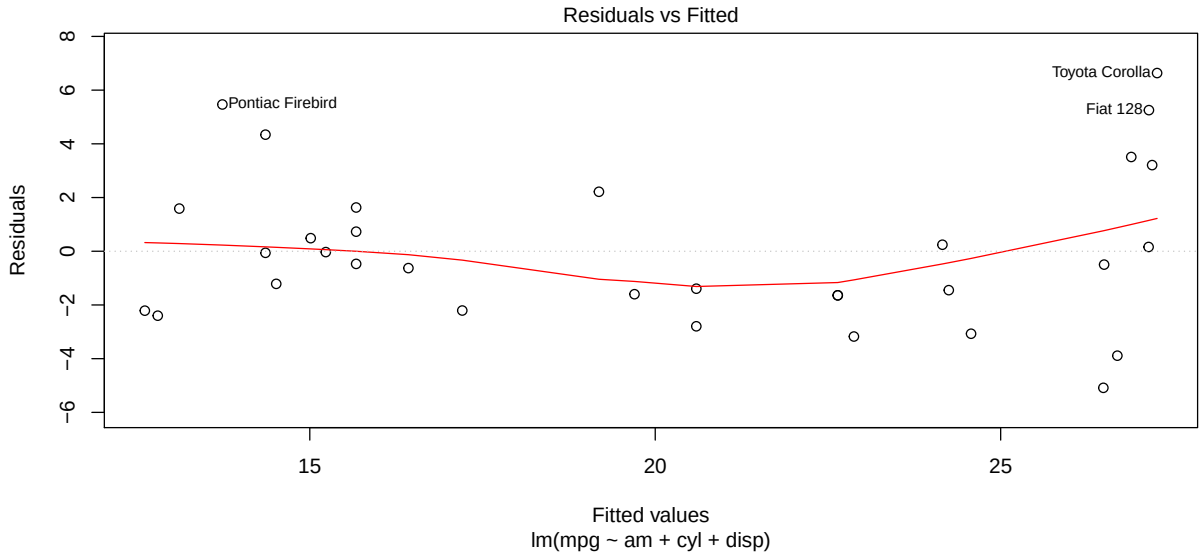

Figure 5: Normality of the Residuals