Regression Models

Maverix13

October 19, 2015

Introduction

The report presents an analysis in the relationship automobile transmission and miles per gallon (MPG) as output. The dataset(mtcars) used in the study was extracted from the 1974 Motor Trend US magazine. This report also quantifies the MPG difference between automatic and manual transmission.

Analysis

Exploratory Analysis

The dataset mtcars is loaded for current analysis.

```
## Loading required package: gridExtra
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

Table above shows an intercept estimate 17.12 interpreted as mean MPG for manual transmission and slope of 7.24 interpreted as difference between the means of manual and automatic transmission with a *p-value* of 2.85e-04 which is significant. Hence, we can reject the **null hypothesis** and further investigate the effect of other variables. Figure 1 shows a graphical depiction of above analysis.

Further, pair analysis of Figure 2 shows the correlation of variables other than am may have effect on MPG.

Model Selection

Model selection requires a combination of predictors to best determine overall fuel efficiency. Including all the predictors will result in high standard error. Following steps will evaluate models to make up best formula for prediction.

Collinearity

To diagnose collinearity in multiple variables in our model, variance inflation factor(VIF) is used as a diagnostic tool. Since this model contains factor variables, VIF values for factor variables will be very high depending on the number of factor values measured as Degrees of freedom. Hence to provide for comparison, we use GVIF^(1/(2*Df)) (the square root of the VIF/GVIF value as DF=1) which is the proportional change of the standard error and confidence interval of their coefficients due to the level of collinearity.

```r
model <- lm(mpg ~ factor(cyl) + disp + hp + drat + wt + qsec + factor(vs) + factor(am) + factor(gear) +
vif(model)[,3]
```

```
##  factor(cyl)         disp           hp         drat           wt
##     3.364380     7.769536     5.312210     2.609533     4.881683
##         qsec   factor(vs)   factor(am) factor(gear) factor(carb)
##     3.284842     2.843970     3.151269     2.670408     1.862838
```

We notice that disp has unusually high value. Also, referring to Figure 2, we can see that cyl and disp has a correlation of 0.902 signifying that disp is a redundat variable can be dropped from the model.

Stepwise Selection

[Reference: http://www.biostat.jhsph.edu/~iruczins/teaching/jf/ch10.pdf, Section: 10.2/10.3]

We start with a model including all the variables. Stepwise model selection uses the Akaike information criterion that implements both forward selection and backward elimination. This ensures that we have included useful variables while omitting ones that do not contribute significantly to predicting mpg.

```r
model <- lm(mpg ~ factor(cyl) + disp + hp + drat + wt + qsec + factor(vs) + factor(am) + factor(gear) +
bestModel <- stepAIC(model, direction = "both")
```

```r
summary(bestModel)
```

```
##
## Call:
## lm(formula = mpg ~ factor(cyl) + hp + wt + factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.70832    2.60489  12.940 7.73e-13 ***
## factor(cyl)6 -3.03134    1.40728  -2.154  0.04068 *
## factor(cyl)8 -2.16368    2.28425  -0.947  0.35225
## hp           -0.03211    0.01369  -2.345  0.02693 *
## wt           -2.49683    0.88559  -2.819  0.00908 **
## factor(am)1   1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

The best model is based on cyl, hp, wt and am as predictors with R-squared of 86.6%, meaning 86.6% of the variability is captured by this model.

Model Comparison

In this section, we compare the models using Nested Likelihood Ratio Test. The models we are using are simple model (mpg ~ factor(am)), stepwise selected model (mpg ~ factor(cyl) + hp + wt + factor(am)), collinearity model (mpg ~ factor(cyl) + hp + drat + wt + qsec + factor(vs) + factor(am) + factor(gear) + factor(carb)) and model containing all the variables (mpg ~ factor(cyl) + disp + hp + drat + wt + qsec + factor(vs) + factor(am) + factor(gear) + factor(carb)).

```
fit1 <- lm(mpg ~ factor(am), data = mtcars)
fit2 <- lm(mpg ~ factor(cyl) + hp + wt + factor(am), data = mtcars)
fit3 <- lm(mpg ~ factor(cyl) + hp + drat + wt + qsec + factor(vs) + factor(am) + factor(gear) + factor(c
fit4 <- lm(mpg ~ factor(cyl) + disp + hp + drat + wt + qsec + factor(vs) + factor(am) + factor(gear) + f
anova(fit1, fit2, fit3, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(cyl) + hp + wt + factor(am)
## Model 3: mpg ~ factor(cyl) + hp + drat + wt + qsec + factor(vs) + factor(am) +
##     factor(gear) + factor(carb)
## Model 4: mpg ~ factor(cyl) + disp + hp + drat + wt + qsec + factor(vs) +
##     factor(am) + factor(gear) + factor(carb)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 17.7489 1.476e-05 ***
## 3     16 130.37 10     20.66  0.2573    0.9822
## 4     15 120.40  1      9.97  1.2417    0.2827
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting the results above, we see that second model has a p-value which is significant and we can reject the null hypothesis that additional varaibles do not contribute to MPG. While model 3 and 4 have insignificat p value so null hypothesis holds.

Further analysis will be done on model 2 (mpg ~ factor(cyl) + hp + wt + factor(am)).

```
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ factor(cyl) + hp + wt + factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.70832    2.60489  12.940 7.73e-13 ***
## factor(cyl)6 -3.03134    1.40728  -2.154  0.04068 *
## factor(cyl)8 -2.16368    2.28425  -0.947  0.35225
## hp           -0.03211    0.01369  -2.345  0.02693 *
## wt           -2.49683    0.88559  -2.819  0.00908 **
## factor(am)1   1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

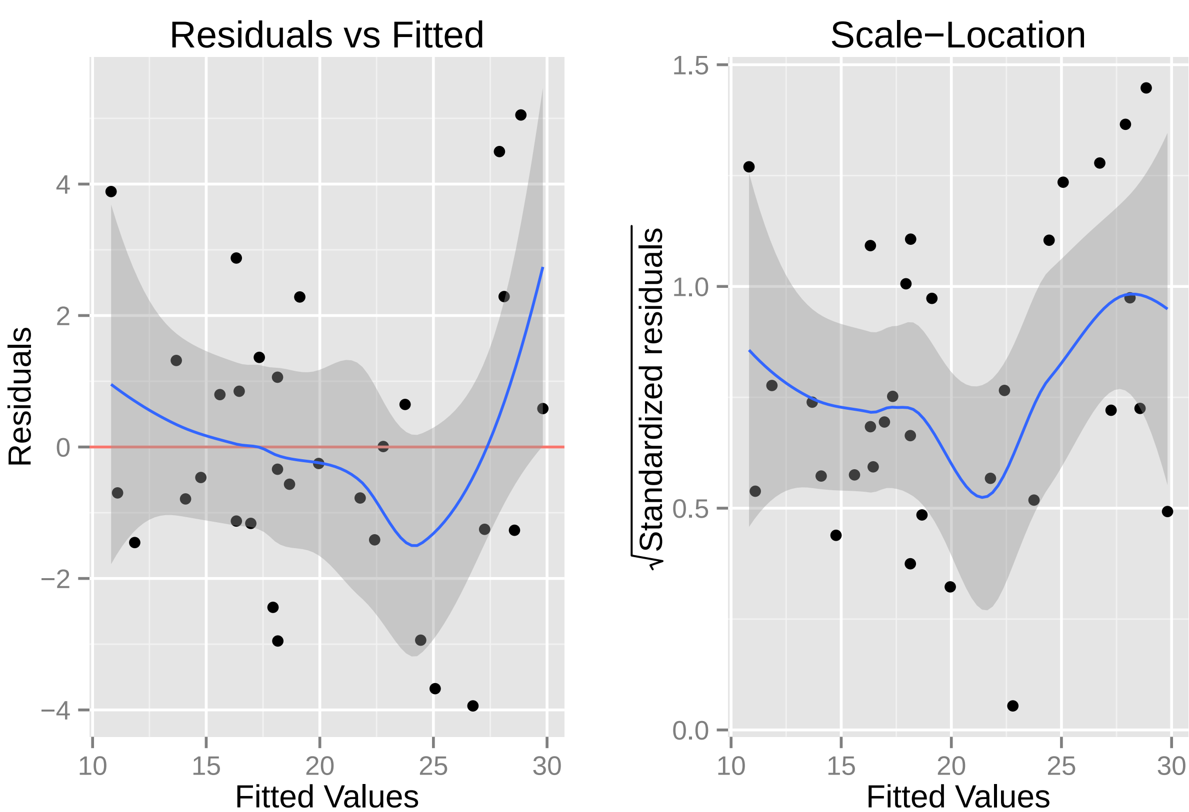
The model above shows a R-squared of 0.8659 explaining 86.59% of variation. Also, model has a very low p value and we can confidently reject the null hypothesis.

Residual and Diagnostics

Residuals vs Fitted Values

The graphs below show residuals plotted against fitted values. First graphs show that there are no systematic patterns. Standardized residuals provide more comparable scale (making it a t like statistic). Again there is no systematic pattern visible.

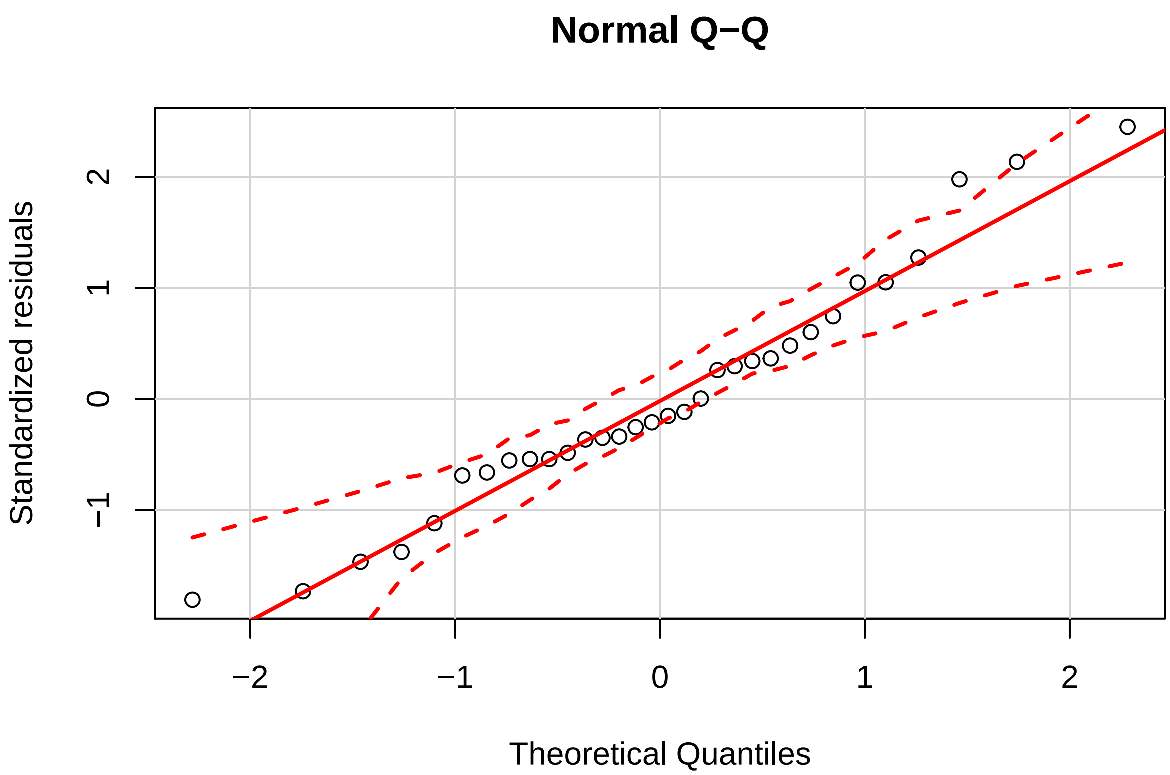
```
ei <- resid(fit2)
residFitPlot <- ggplot(data.frame(x = fit2$fitted.values, y = ei), aes(x = x, y = y)) +
  geom_point() + geom_hline(aes(yintercept=0, colour = "red")) + geom_smooth(method = "loess") +
  xlab("Fitted Values") + ylab("Residuals") + ggtitle("Residuals vs Fitted")

s <- sqrt(deviance(fit2)/df.residual(fit2))
rs <- ei/s
sqrt.rs <- sqrt(abs(rs))
scaleLocPlot <- ggplot(data.frame(x = fit2$fitted.values, y= sqrt.rs), aes(x = x, y = y)) +
  geom_point() + geom_smooth(method = "loess") +
  xlab("Fitted Values") + ylab(expression(sqrt("Standardized residuals"))) + ggtitle("Scale-Location")

grid.arrange(residFitPlot, scaleLocPlot, ncol = 2)
```



Normality of Residuals

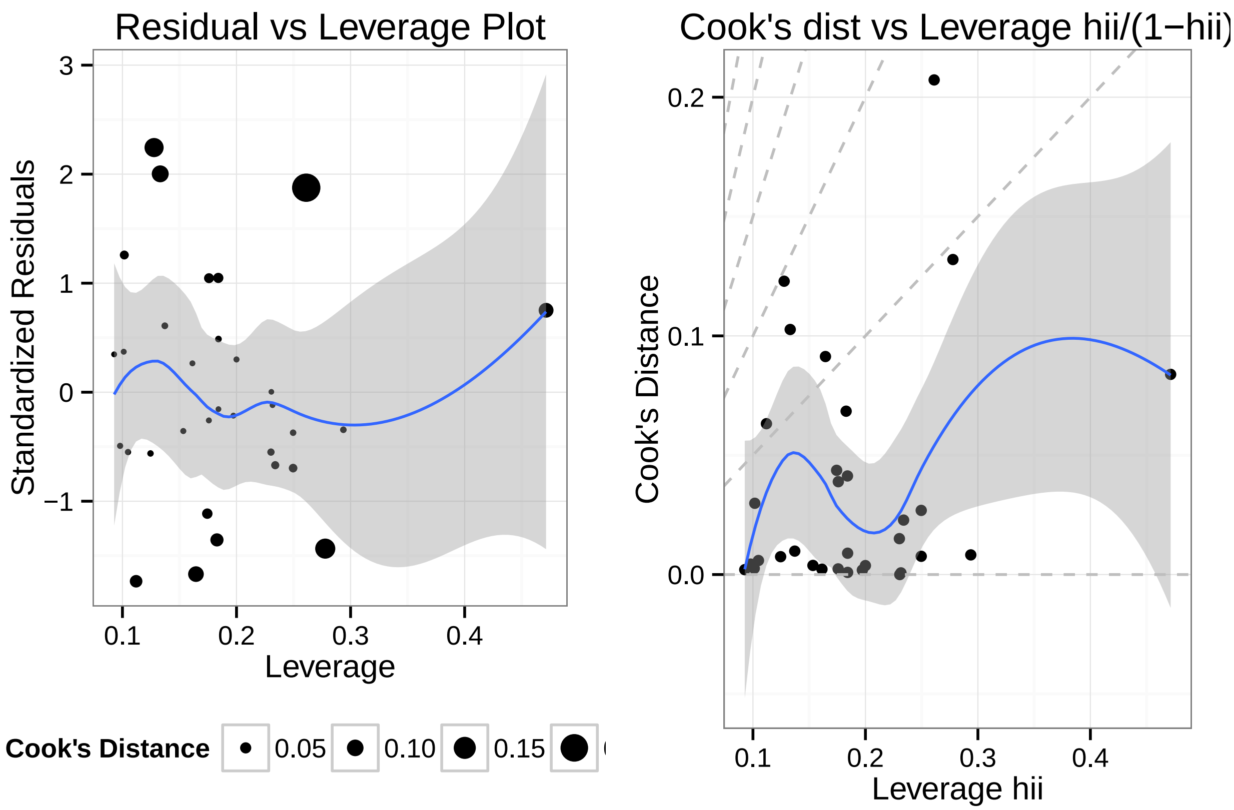
Normal Q-Q plot testing the normality of errors by plotting theoretical quantiles by standardized residuals.

```
qq <- qqPlot(fit2, ylab = "Standardized residuals", xlab = "Theoretical Quantiles", main = "Normal Q-Q")
```



Influence Measures

```
residLevPlot<-ggplot(fit2, aes(.hat, .stdresid))+geom_point(aes(size=.cooksd)) +
    stat_smooth(method="loess", na.rm=TRUE) +
    xlab("Leverage")+ylab("Standardized Residuals") +
    ggtitle("Residual vs Leverage Plot") +
    scale_size_continuous("Cook's Distance", range=c(1,5)) +
    theme_bw()+theme(legend.position="bottom")

cksdPlot <-ggplot(fit2, aes(.hat, .cooksd))+geom_point(na.rm=TRUE)+stat_smooth(method="loess") +
    xlab("Leverage hii")+ylab("Cook's Distance") +
    ggtitle("Cook's dist vs Leverage hii/(1-hii)") +
    geom_abline(slope=seq(0,3,0.5), color="gray", linetype="dashed") +
    theme_bw()

grid.arrange(residLevPlot, cksdPlot, ncol = 2)
```


Residual vs Leverage Plot
Standardized Residuals
Leverage
Cook's Distance
0.05
0.10
0.15
Cook's dist vs Leverage hii/(1−hii)
Cook's Distance
Leverage hii

Appendix

Exploratory Analysis

Figure 1

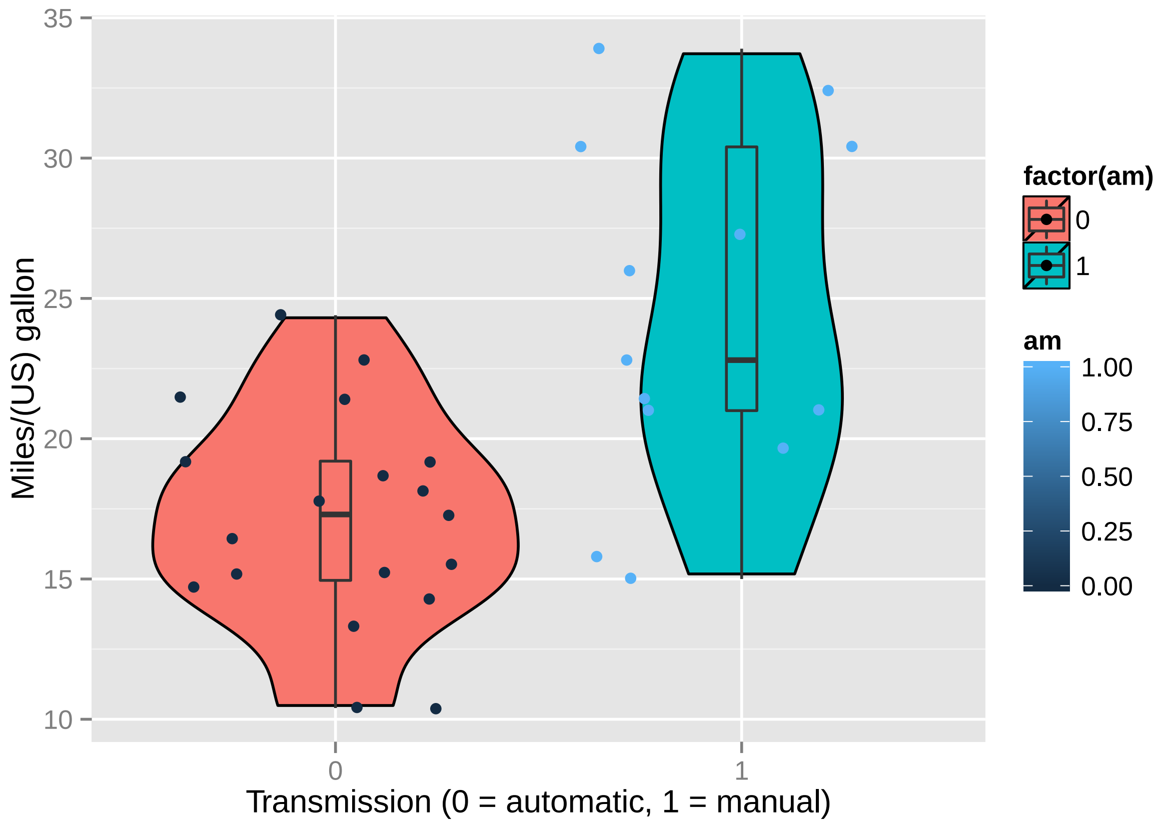
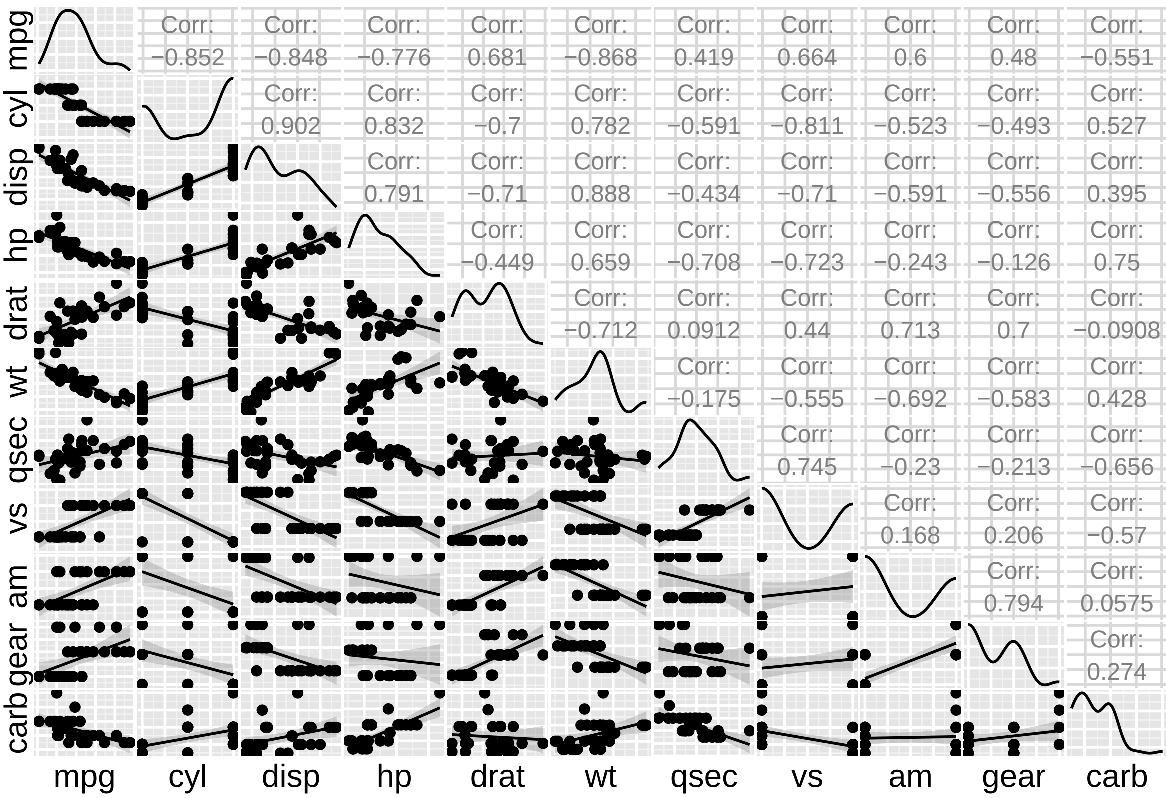


Figure 2


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb
Corr: −0.852
Corr: −0.848
Corr: −0.776
Corr: 0.681
Corr: −0.868
Corr: 0.419
Corr: 0.664
Corr: 0.6
Corr: 0.48
Corr: −0.551
Corr: 0.902
Corr: 0.832
Corr: −0.7
Corr: 0.782
Corr: −0.591
Corr: −0.811
Corr: −0.523
Corr: −0.493
Corr: 0.527
Corr: 0.791
Corr: −0.71
Corr: 0.888
Corr: −0.434
Corr: −0.71
Corr: −0.591
Corr: −0.556
Corr: 0.395
Corr: −0.449
Corr: 0.659
Corr: −0.708
Corr: −0.723
Corr: −0.243
Corr: −0.126
Corr: 0.75
Corr: −0.712
Corr: 0.0912
Corr: 0.44
Corr: 0.713
Corr: 0.7
Corr: −0.0908
Corr: −0.175
Corr: −0.555
Corr: −0.692
Corr: −0.583
Corr: 0.428
Corr: 0.745
Corr: −0.23
Corr: −0.213
Corr: −0.656
Corr: 0.168
Corr: 0.206
Corr: −0.57
Corr: 0.794
Corr: 0.0575
Corr: 0.274
mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb