Regression Analysis: MotorTrend_MPG

EHarris

7/19/2017

Executive Summary

Motor Trend has provided a data set, mtcars, with 32 observations from the 1974 Motor Trend magazine. They wish to understand the relationship between a number of variables, predictors, and miles-per-gallon (MPG), outcome, Specifically, were are asked to answer the following two questions:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions"

Under a simple linear regression model, it would appear that a manual transmission adds more than 7 mpg over an automatic. Using additional variables to our regression model, we conclude that switching from an automatic to manual adds 1.8 mpg, all else equal.

Data Processing

Given that this is an R dataset we can simply use data(mtcars) to load the dataset into the current enviromnent.

```
data(mtcars)
mtcars2 <- mtcars
mtcars2$am <- as.factor(mtcars2$am)
mtcars2$cyl <- as.factor(mtcars2$cyl)
```

Exploratory Data Analysis

To obtain a high-level insight to the relationship of the different variables in the data set, we created a panel plot (see Appendix: Figure 1) of these relationships. A correlation matrix (see Appendix: Figure 2) is also provided to provide a numeric value to the relationship. We find that the variables cyl, disp, hp, and wt have the strongest relationship with mpg. The variables drat, vs, and am are slightly less connected. We use this information to help assess our regression analysis results.

Regression Analysis

The focus of our analysis is miles per gallon (MPG). Starting with simple linear regression (mpg ~am), we add variables to build a more robust, and best fit, model. We use several metrics to evaluate our model fit, including: adjusted r-square, residual squared error (sigma), and p-values. We also use ANOVA and an analysis of the residuals to evaluate the model.

Regression Models

The initial model, ‘fit1’, looks at the effect of transmission type (automatic vs. manual) on the MPG of a car. We also create a ‘fit10’ model that includes all 10 independent variables. We perform a stepwise processs, adding & removing variables, to select the best variables. This can be done manually and/or using a step() function which uses algorithm to evaluate whether variables contribute significantly to predicting mpg.

```
fit1 <- lm(mpg ~ am, data = mtcars2)
coef1 <- round(summary(fit1)$coef, 4)
glance1 <- round(glance(fit1)[1:6], 4)

fit10 <- lm(mpg ~ ., data = mtcars2)
coef10 <- round(summary(fit10)$coef, 4)
glance10 <- round(glance(fit10)[1:6], 4)

bestfit <- step(fit10, direction = "both")
coef.best <- round(summary(bestfit)$coef, 4)
glance.best <- round(glance(bestfit)[1:6], 4)

## Obtained by manually adding / removing variables (using correlations as guide)
fit3 <- lm(mpg ~ am + wt + cyl, data = mtcars2)
coef3 <- round(summary(fit3)$coef, 4)
glance3 <- round(glance(fit3)[1:6], 4)

fit4 <- lm(mpg ~ am + wt + cyl + hp, data = mtcars2)
coef4 <- round(summary(fit4)$coef, 4)
glance4 <- round(glance(fit4)[1:6], 4)
```

Comparison of Regression Models

Two comparisons performed to evaluate the best model. The first is a comparison of the r.square and p-value for each model. Then, we perform ANOVA to evaluate whether model is significantly better.

Exhbit 1: R.Square and p-value Comparison

```
model <- c("fit1", "fit3", "fit4", "bestfit", "fit10")
model.compare <- rbind(glance1,glance3, glance4, glance.best, glance10)
model.compare <- cbind(model, model.compare)
model.compare
```

```
##     model r.squared adj.r.squared  sigma statistic p.value df
## 1    fit1    0.3598        0.3385 4.9020   16.8603   3e-04  2
## 2    fit3    0.8375        0.8134 2.6032   34.7917   0e+00  5
## 3    fit4    0.8659        0.8401 2.4101   33.5712   0e+00  6
## 4 bestfit    0.8497        0.8336 2.4588   52.7496   0e+00  4
## 5   fit10    0.8816        0.8165 2.5819   13.5381   0e+00 12
```

Exhibit 2: ANOVA

The findings provided below for Model 4: mpg ~ factor(am) + wt + factor(cyl) + hp produce a p-value of 0.026935, which indicates that this model is significant. These variables further contribute to the accuracy of the model.

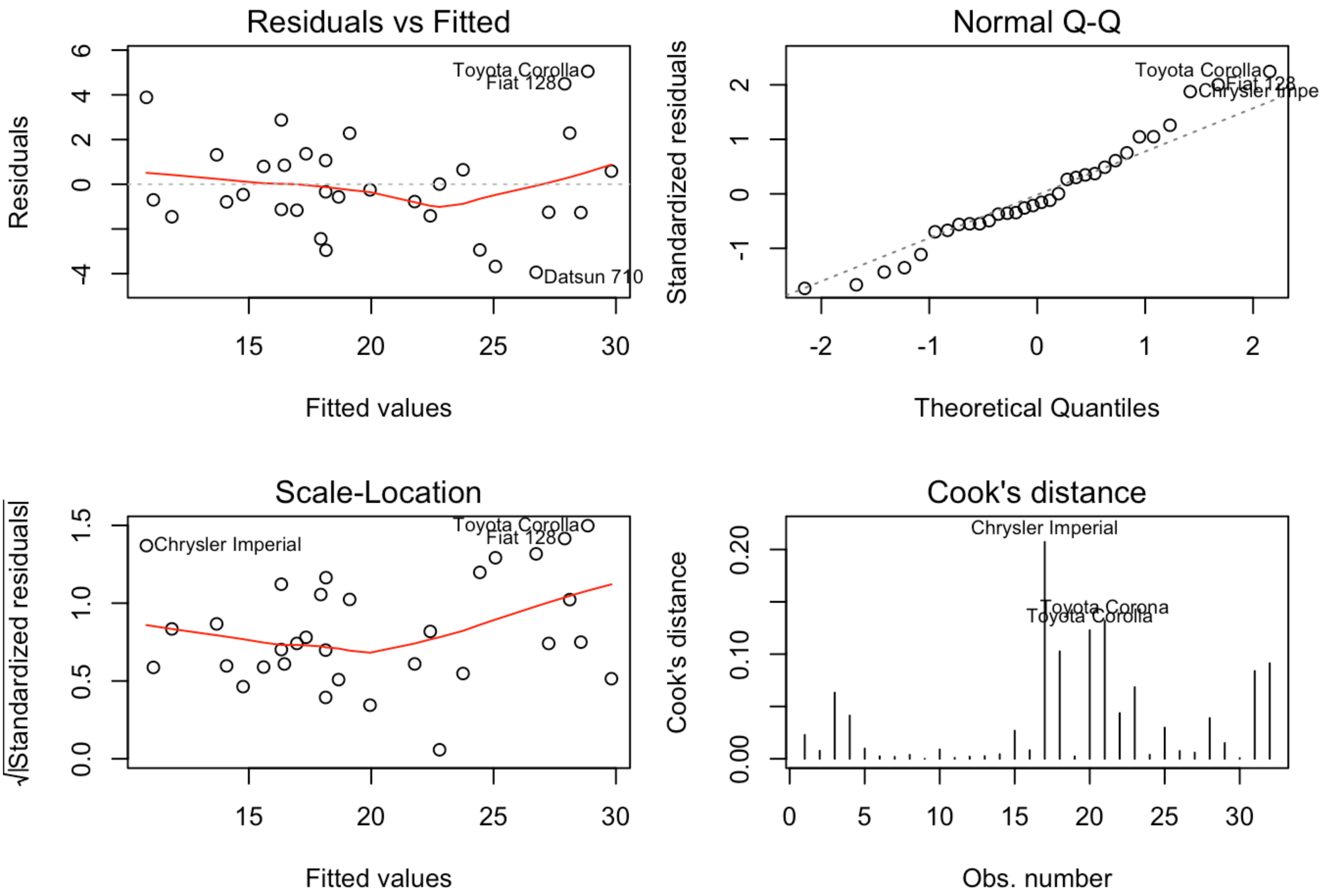
```
mpg.anova <- anova(fit1, fit3, fit4)
mpg.anova
```

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + cyl
## Model 3: mpg ~ am + wt + cyl + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1     30 720.90                                   
## 2     27 182.97  3    537.93 30.8692 1.008e-08 ***
## 3     26 151.03  1     31.94  5.4991   0.02693 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residuals & Diagnostics

Below is a panel plot of residuals. Understanding residuals is critical to understanding how well the regression model fits the data, goodness of fit. The different plots provide some insight to how closely the regression line fits the data, explaining changes in mpg.

```
par(mfrow = c(2,2), mar = c(5,4,2,1))
plot(fit4, which = 1:4)
```



** Interpreting the Plots**

1. Residuals vs Fitted: This plot shows if residuals have non-linear patterns. Want to see that the residuals are fairly well-distributed around the fitted line, and no particular pattern exists. 2. Normal Q-Q: This plot shows if residuals are normally distributed. It's good fit if residuals are lined closely the straight dashed line. 3. Scale-Location: Use to check the assumption of equal variance (homoscedasticity). Ideally, we would see points that are equally spread around a horizontal line along the entire range of predictors. 4. Cooks distance: The purpose of this plot is identify may be considered outliers. The "Cooks distance" is considerably different from others. These points may be influential against the regression line, changing intercept and/or slope.

Additional Regression Diagnostics

In addition to plots provided above, we can look evaluate our model through some other tests. Below, we look at a couple and explain the findings.

High Leverage/Influential Data Points

There are two functions, dfbetas() and hatvalues() to identify points of interest. We could also use the function influence.measures() to obtain several measures about influence of data points.

DFBETAS

Measure that standardizes the absolute difference in parameter estimates between a (mixed effects) regression model based on a full set of data, and a model from which a (potentially influential) subset of data is removed

```
fit4.dfb <- dfbetas(fit4, parameters = 0, sort = TRUE)   ## ordered based on magnitud
e
influence.fit4.id <- which(fit4.dfb > (2/(length(mtcars$mpg)^.5))) ##  dfbetas > #2 /
sqrt(n)
rbind(influence.fit4.id, fit4.dfb[influence.fit4.id])   ## value relfects a car & fit
4 variable combination
```

```
##                         [,1]       [,2]       [,3]       [,4]        [,5]
## influence.fit4.id 50.0000000 53.0000000 81.0000000 85.0000000 117.0000000
##                    0.4292043  0.7305402  0.9389082  0.3643262   0.4771765
##                          [,6]        [,7]      [,8]        [,9]
## influence.fit4.id 128.0000000 149.0000000 160.00000 191.0000000
##                     0.4240429   0.5436814   0.35893   0.5250878
```

HATVALUES

A hat value measures the leverage of Xi variable, indicating that the point is sufficiently far from the mean of X. The further the distance from the mean, the greater the leverage. NOTE: Leverage does not mean it influences the regression coeffiencent.

```
fit4.hat <- hatvalues(fit4)
leverage.fit4.id <- which(fit4.hat > (2*(4+1)/length(mtcars2$mpg))) ##  hatvalue > #2
*(k+1)/n
fit4.hat[leverage.fit4.id]
```

```
## Maserati Bora
##     0.4713671
```

INFLUENCE MEAUSRES

The function influence.measures provides a summary of multiple measures to evaluate the influence of a point, or observation, on the coefficients.

```
summary(influence.measures(fit4))
```

```
## Potentially influential observations of
##   lm(formula = mpg ~ am + wt + cyl + hp, data = mtcars2) :
##
##                     dfb.1_ dfb.am1 dfb.wt dfb.cyl6 dfb.cyl8 dfb.hp dffit
## Lincoln Continental  0.16  -0.09   -0.19   0.06     0.04     0.04  -0.22
## Maserati Bora       -0.18   0.04   -0.09  -0.14    -0.25     0.53   0.70
##                     cov.r   cook.d hat
## Lincoln Continental  1.74_*  0.01   0.29
## Maserati Bora        2.10_*  0.08   0.47
```

Conclusion

Following our review of Exploratory Data Analysis, it was decided to test two assumptions. From the boxplot, it appeared that the supplement 'OJ' promoted greater tooth growth than the supplement 'VC'. It was also apparent that a higher dose promoted greater tooth growth, regardless of supplement introduced. We chose to test these assumptions by setting NULL hypothesis that there was no difference in the tooth growth for supplement 'OJ' versus 'VC' and a higher dose versus lower dose. The alternative hypothesis being that 'OJ' promotes greater tooth growth than 'VC' and a higher dose promotes greater tooth growth than a lower dose.

From our t.test, the p-values were rather small, with the exception of testing whether 'OJ' promotes greater tooth growth than 'VC' at a dose = 2.0, indicating that we reject the NULL hypothesis. In summary, the analysis suggests that 'OJ' promotes greater tooth growth than 'VC'. Additionally, regardless of the supplement selected, 'OJ' or 'VC', a higher dose of the supplement promotes greater tooth growth than lower dose.

Appendix

The information provided below is an outline of the request or question driving this assingment and analysis of data for different cars and related miles per gallon (MPG).

Review criteria

1. Did the student interpret the coefficients correctly?
2. Did the student do some exploratory data analyses?
3. Did the student fit multiple models and detail their strategy for model selection?
4. Did the student answer the questions of interest or detail why the question(s) is (are) not answerable?
5. Did the student do a residual plot and some diagnostics?
6. Did the student quantify the uncertainty in their conclusions and/or perform an inference correctly?
7. Was the report brief (about 2 pages long) for the main body of the report and no longer than 5 with supporting appendix of figures?
8. Did the report include an executive summary?
9. Was the report done in Rmd (knitr)?

Figure 1 (Panel Plot): Illustrate the relationship of each variable

This panel plot provides a quick, high-level, illustration of a variable within the dataset against each other variable in the dataset. The pu

```
pairs(mtcars, panel = panel.smooth, main = "MT Cars Data", col = 3)
```

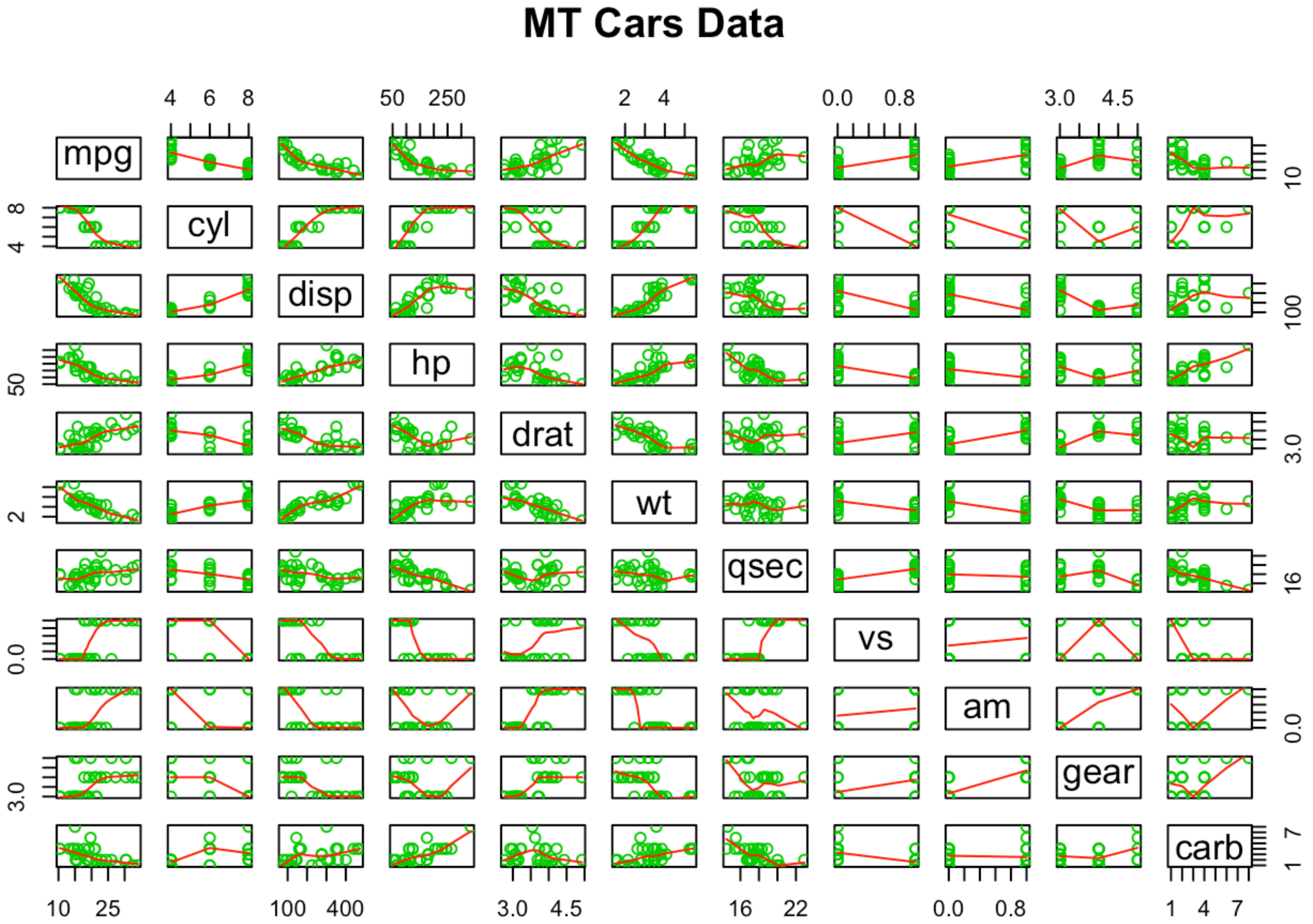
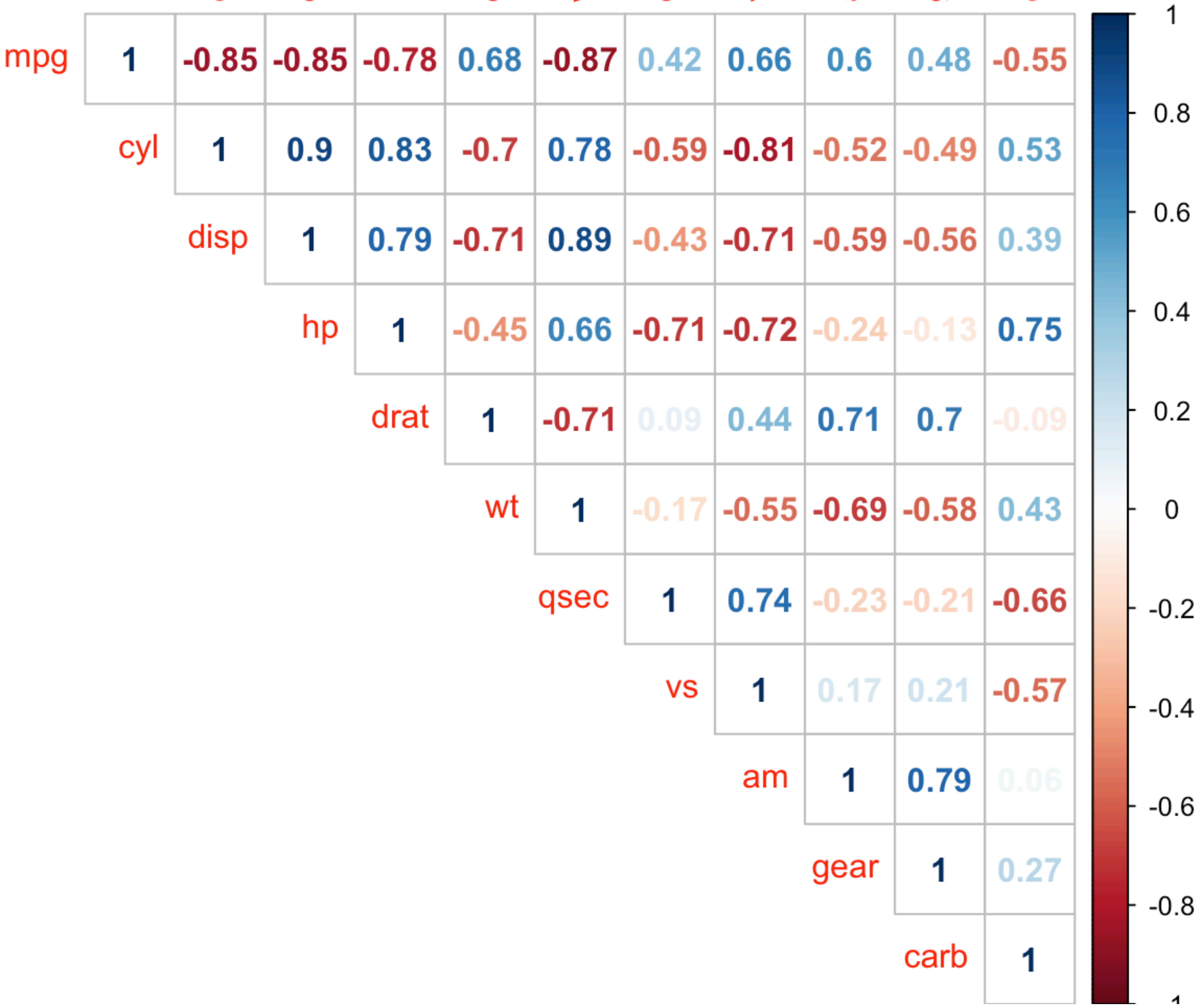


Figure 2: Correlation Factor Matrix

```
par(mar = c(5,4,2,1))
mtcars.cor <- cor(mtcars)
corrplot(mtcars.cor, method = "number", type = "upper", add = FALSE,
                   order = "original", is.corr = TRUE)
```



mpg	1	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.6	0.48	-0.55
	cyl	1	0.9	0.83	-0.7	0.78	-0.59	-0.81	-0.52	-0.49	0.53
		disp	1	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
			hp	1	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
				drat	1	-0.71	0.09	0.44	0.71	0.7	-0.09
					wt	1	-0.17	-0.55	-0.69	-0.58	0.43
						qsec	1	0.74	-0.23	-0.21	-0.66
							vs	1	0.17	0.21	-0.57
								am	1	0.79	0.06
									gear	1	0.27
										carb	1