# Regression Models Course Project

*Marcelo Tibau*

*30 de setembro de 2016*

# Executive Summary

This paper explores the relationship between miles-per-gallon (MPG) and other variables in the mtcars dataset.The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

Dataset source provided by: Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391-411.

The performed analysis attempts to determine whether an automatic or manual transmission is better for MPG and to quantifies the MPG difference. I transcript bellow the received instructions:

*You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:*

*1) "Is an automatic or manual transmission better for MPG"*

*2) "Quantify the MPG difference between automatic and manual transmissions"*

This document contains two sections: an Analysis section, where I intend to determine if there is a signficant difference between the mean MPG for automatic and manual transmission cars through a linear regression analysis and an Appendix section where I provide some exploratory analysis and visualizations.

# Analysis

Codes to perform the data processing, loading the dataset and transforming certain variables into factors.

```
data("mtcars")
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$vs   <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am   <- factor(mtcars$am,labels=c("Automatic","Manual"))
```

Codes to model some Linear Regressions. I will start with a simple Linear Regression to explore miles per gallon ~ transmission and set a null hypothesis that there is no significant difference in mpg between the two groups at alpha = .05.

```
n <- length(mtcars$mpg)
alpha <- .05
model1 <- lm(mpg ~ am, data = mtcars)
summary(coef(model1))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.245   9.721  12.200  12.200  14.670  17.150
```

```
coef(summary(model1))
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
```

```
## amManual     7.244939   1.764422  4.106127 2.850207e-04
```

I´m using the equation beta0 + beta1 to calculate the mpg mean for cars with manual transmissions.I consider the following: beta0/intercept as the mpg mean for cars with automatic transmissions; beta1 is the mean increase in mpg for cars with manual transmissions (am = 1); beta1/am is the mean increase in mpg for cars with manual transmissions.

Code to perform the 95% confidence interval for beta1:

```
tran_est <- coef(summary(model1))["amManual", "Estimate"]
std_err <- coef(summary(model1))["amManual", "Std. Error"]
```

Code to calculate the stat, using n - 2 to model with intercept and slope:

```
stat <- qt(1 - alpha/2, n-2)
tran_est + c(-1,1)*(std_err*stat)
```

```
## [1]  3.64151 10.84837
```

The p-value (2.850207e-04) is small and the confidence interval does not include zero. Therefore I can reject the null hypothesis in favor of the alternative hypothesis that there is a significant difference in mpg between the two groups at alpha = .05.

Next, I intend to explore a Multiple Regression that includes all variables as predictors of mpg.Then, I will perform a stepwise model selection to pick significant predictors for the final model.

```
model2 <- lm(mpg ~ ., data = mtcars)
best_model <- step(model2, direction = "both")
```

```
## Start:  AIC=76.4
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  5   13.5989 134.00 69.828
## - gear  2    3.9729 124.38 73.442
## - am    1    1.1420 121.55 74.705
## - qsec  1    1.2413 121.64 74.732
## - drat  1    1.8208 122.22 74.884
## - cyl   2   10.9314 131.33 75.184
## - vs    1    3.6299 124.03 75.354
## <none>              120.40 76.403
## - disp  1    9.9672 130.37 76.948
## - wt    1   25.5541 145.96 80.562
## - hp    1   25.6715 146.07 80.588
##
## Step:  AIC=69.83
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  2    5.0215 139.02 67.005
## - disp  1    0.9934 135.00 68.064
## - drat  1    1.1854 135.19 68.110
## - vs    1    3.6763 137.68 68.694
## - cyl   2   12.5642 146.57 68.696
## - qsec  1    5.2634 139.26 69.061
## <none>              134.00 69.828
## - am    1   11.9255 145.93 70.556
```

```
## - wt    1   19.7963 153.80 72.237
## - hp    1   22.7935 156.79 72.855
## + carb  5   13.5989 120.40 76.403
##
## Step:  AIC=67
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1    0.9672 139.99 65.227
## - cyl   2   10.4247 149.45 65.319
## - disp  1    1.5483 140.57 65.359
## - vs    1    2.1829 141.21 65.503
## - qsec  1    3.6324 142.66 65.830
## <none>              139.02 67.005
## - am    1   16.5665 155.59 68.608
## - hp    1   18.1768 157.20 68.937
## + gear  2    5.0215 134.00 69.828
## - wt    1   31.1896 170.21 71.482
## + carb  5   14.6475 124.38 73.442
##
## Step:  AIC=65.23
## mpg ~ cyl + disp + hp + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1    1.2474 141.24 63.511
## - vs    1    2.3403 142.33 63.757
## - cyl   2   12.3267 152.32 63.927
## - qsec  1    3.1000 143.09 63.928
## <none>              139.99 65.227
## + drat  1    0.9672 139.02 67.005
## - hp    1   17.7382 157.73 67.044
## - am    1   19.4660 159.46 67.393
## + gear  2    4.8033 135.19 68.110
## - wt    1   30.7151 170.71 69.574
## + carb  5   13.0509 126.94 72.095
##
## Step:  AIC=63.51
## mpg ~ cyl + hp + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - qsec  1     2.442 143.68 62.059
## - vs    1     2.744 143.98 62.126
## - cyl   2    18.580 159.82 63.466
## <none>              141.24 63.511
## + disp  1     1.247 139.99 65.227
## + drat  1     0.666 140.57 65.359
## - hp    1    18.184 159.42 65.386
## - am    1    18.885 160.12 65.527
## + gear  2     4.684 136.55 66.431
## - wt    1    39.645 180.88 69.428
## + carb  5     2.331 138.91 72.978
##
## Step:  AIC=62.06
## mpg ~ cyl + hp + wt + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - vs    1     7.346 151.03 61.655
## <none>              143.68 62.059
```

```
## - cyl   2    25.284 168.96 63.246
## + qsec  1     2.442 141.24 63.511
## - am    1    16.443 160.12 63.527
## + disp  1     0.589 143.09 63.928
## + drat  1     0.330 143.35 63.986
## + gear  2     3.437 140.24 65.284
## - hp    1    36.344 180.02 67.275
## - wt    1    41.088 184.77 68.108
## + carb  5     3.480 140.20 71.275
##
## Step:  AIC=61.65
## mpg ~ cyl + hp + wt + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              151.03 61.655
## - am    1     9.752 160.78 61.657
## + vs    1     7.346 143.68 62.059
## + qsec  1     7.044 143.98 62.126
## - cyl   2    29.265 180.29 63.323
## + disp  1     0.617 150.41 63.524
## + drat  1     0.220 150.81 63.608
## + gear  2     1.361 149.66 65.365
## - hp    1    31.943 182.97 65.794
## - wt    1    46.173 197.20 68.191
## + carb  5     5.633 145.39 70.438
```

```
summary(best_model)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + 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 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -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 **
## amManual     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 adjusted R-squared value of 0.84 is the maximum obtained considering all combinations of variables. Thereafter I can conclude that more than 84% of the variability is explained by the best_model.

To be sure, I intend to compare this best_model with the model1 that brings only the transmission (am) as the predictor variable:

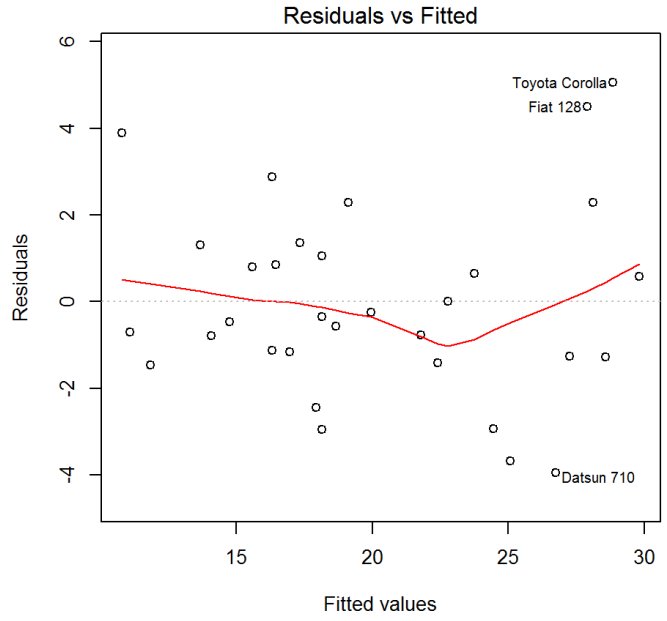
```
anova(model1, best_model)
```

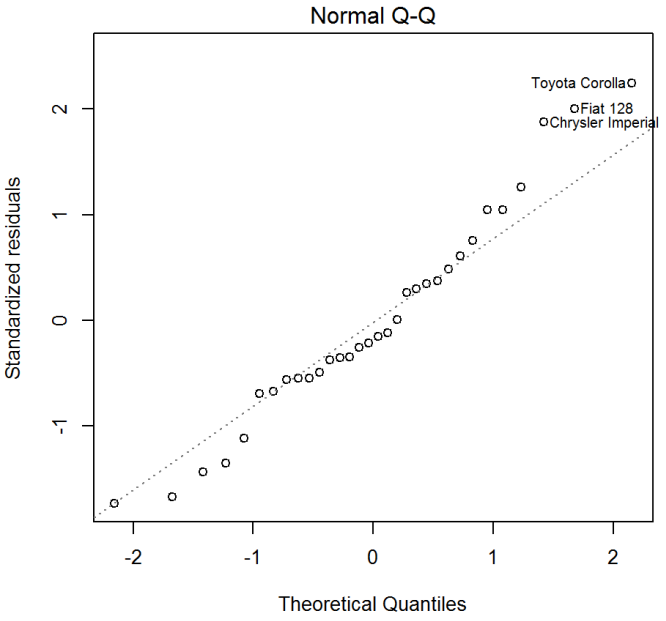
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

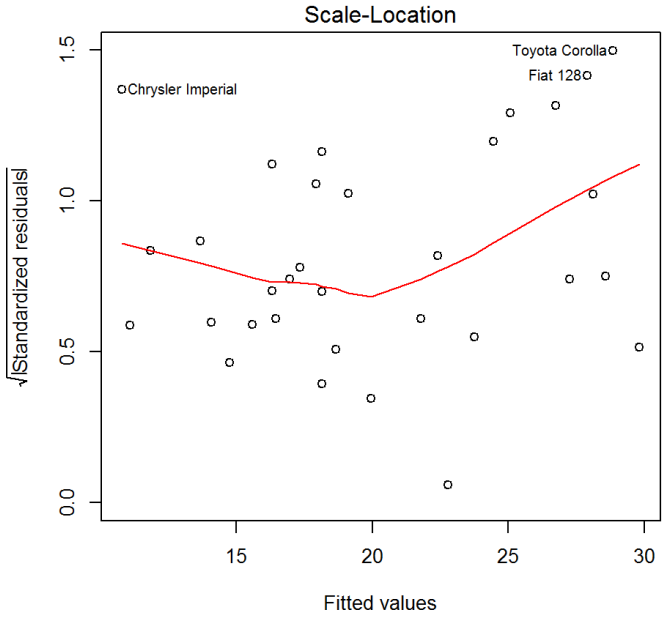
The p-value obtained for the best_model (1.688e-08) is highly significant. Therefore I can conclude that the confounder variables do contribute to the accuracy of the model.

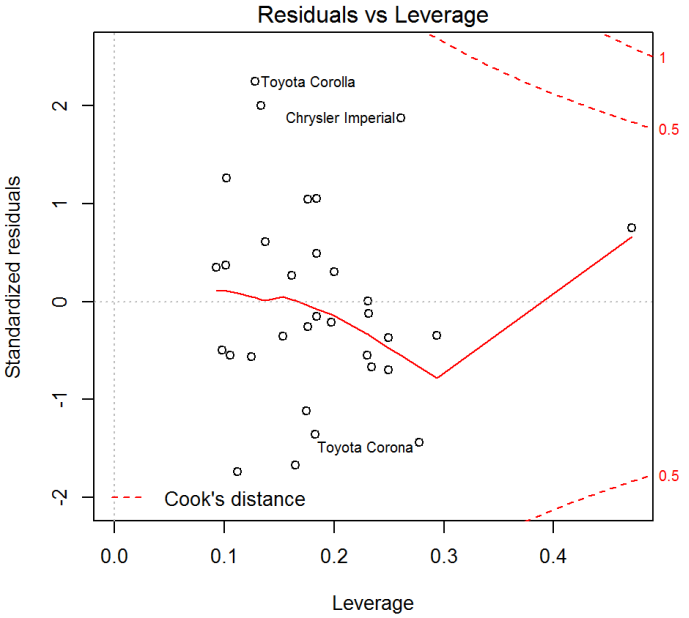
I intend to do some exploratory analysis to examine the residuals and finding leverage points that show any potential problems with the model.

```
par(mfrow=c(2, 2))
plot(best_model)
```




As observed at the residual vs fitted plot, the residuals for the Chrysler Imperial, Fiat 128, and Toyota Corolla exert some influence on the shape of the curve.The curve is shaped slightly from normality.

The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed. The Scale-Location plot consists of points scattered in a constant band pattern, indicating constant variance.

Finally, there are some distinct points of interest (leverage points) in the top right of the plots that may indicate values of increased leverage of outliers.

I will perform some statistical inference, using the t-test on the two subsets of mpg data: manual and automatic transmission. I assume that the transmission data has a normal distribution and will test the null hypothesis that the mpg distributions for manual and automatic transmissions are the same. The t-test will the performed on (mpg ~ am).

```
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

Based on the t-test results, I reject the null hypothesis.

# Conclusions drawn from the Analysis:

1. “Is an automatic or manual transmission better for MPG”

- cars with Manual transmission get 1.8 more miles per gallon compared to cars with Automatic transmission. (1.8 adjusted for hp, cyl, and wt).

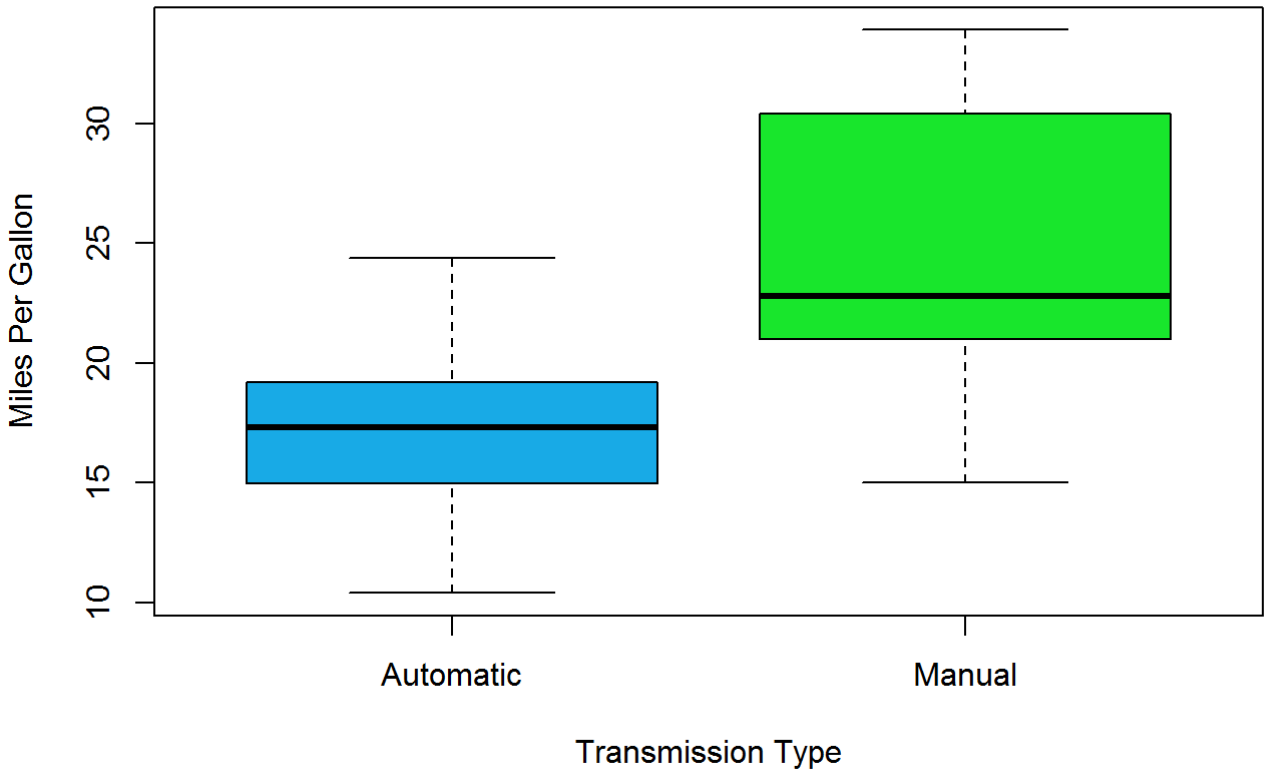
2. “Quantify the MPG difference between automatic and manual transmissions”

- mpg will decrease by 2.5 for every 1000 lb increase in wt.
- mpg decreases negligibly (only 0.32) with every increase of 10 in hp.
- if number of cylinders, cyl increases from 4 to 6 and 8, mpg will decrease by a factor of 3 and 2.2 respectively (adjusted by hp, wt, and am).

# Appendix

## Boxplot of miles per gallon by transmission type:

```
boxplot(mpg ~ am, data = mtcars, col = (c("#18aae6","#18e62c")), ylab = "Miles Per Gallon", xla
b = "Transmission Type")
```




# Correlations:

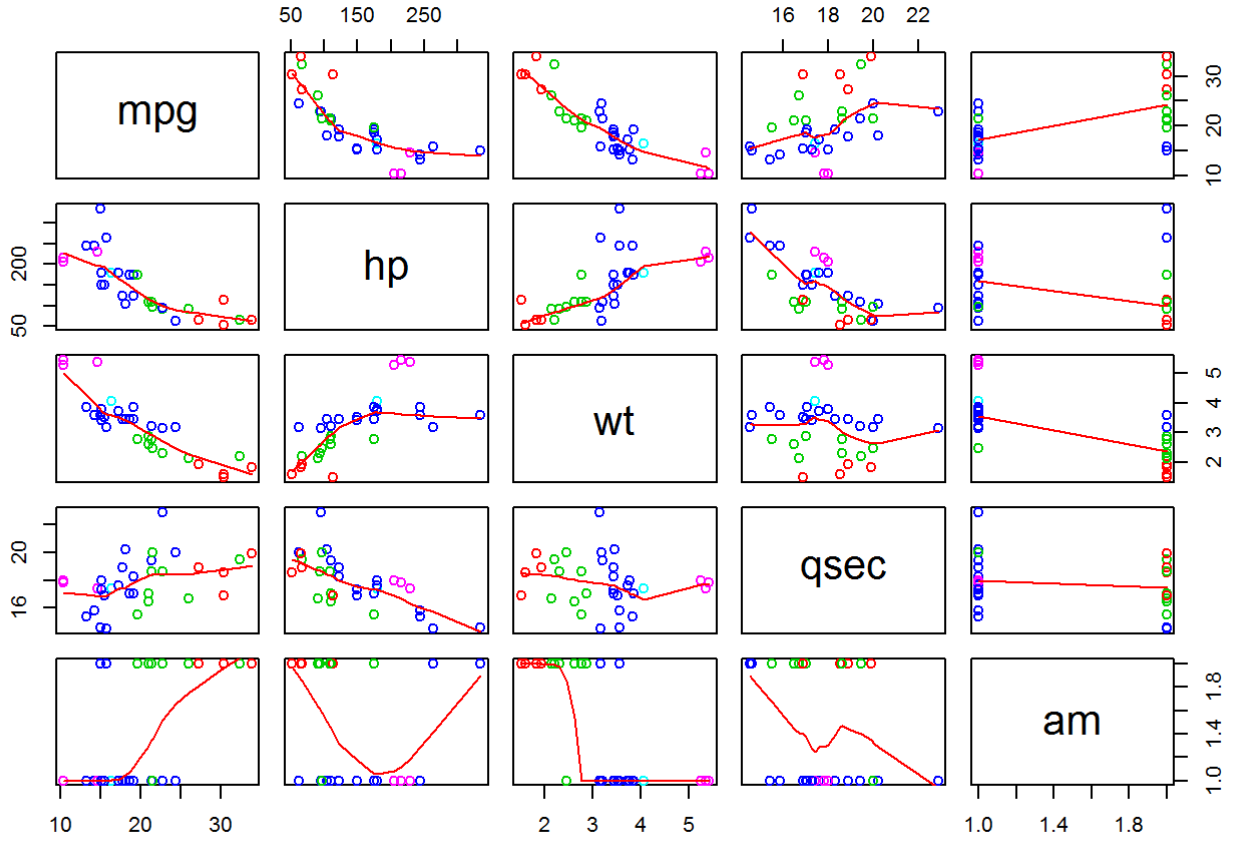
```
mtcars_vars <- mtcars[, c(1, 4, 6, 7, 9)]
```

Code to save the original values:

```
mar.orig <- par()$mar
```

code to set the new values:

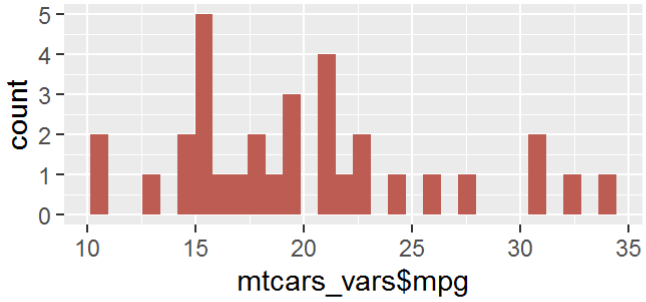
```
par(mar = c(1, 1, 1, 1))
pairs(mtcars_vars, panel = panel.smooth, col = 9 + mtcars$wt)
```

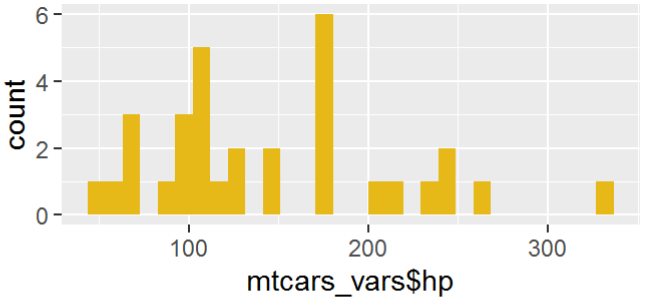


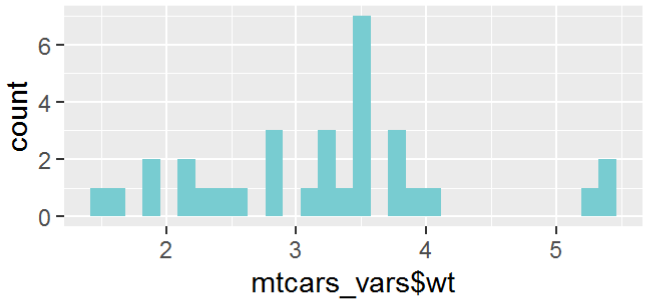

Histogram of the correlations:

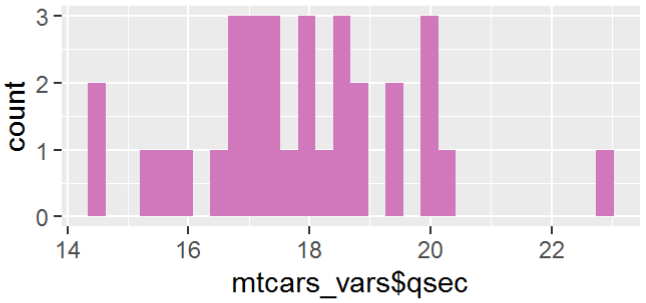
```
library(ggplot2)
library(gridExtra)
mpg_dist <- qplot(mtcars_vars$mpg, fill = I("#bc5c52"))
hp_dist <- qplot(mtcars_vars$hp, fill = I("#e6b818"))
wt_dist <- qplot(mtcars_vars$wt, fill = I("#78ccd1"))
qsec_dist <- qplot(mtcars_vars$qsec, fill = I("#d178bd"))
am_dist <- qplot(mtcars_vars$am, fill = I("#5ade0a"))
grid.arrange(mpg_dist, hp_dist, wt_dist, qsec_dist, am_dist, ncol = 2)
```

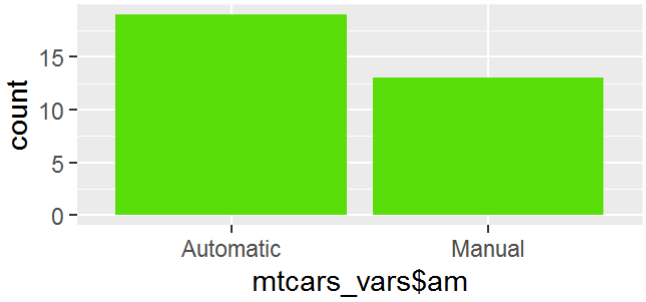
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


count
mtcars_vars$mpg


count
mtcars_vars$hp


count
mtcars_vars$wt


count
mtcars_vars$qsec


count
Automatic
Manual
mtcars_vars$am