Regression Models Course Project

Koji

2018/7/21

Overview

Our 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:

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

Exploratory data analysis

```
library("ggplot2")
library("GGally")
library("gridExtra")
library("dplyr")
# Load data
data(mtcars)
```

Compute summary statistics of data subsets:

```
aggregate(mpg ~ factor(am, labels = c("AT", "MT")), mtcars, mean)
```

```
##   factor(am, labels = c("AT", "MT"))      mpg
## 1                                 AT 17.14737
## 2                                 MT 24.39231
```

Compute correlation:

```
round(cor(mtcars), 2)[1, ]
```

```
##   mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##  1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
```

Fit multiple models

```
fit1 <- lm(mpg ~ am, mtcars)
fit2 <- lm(mpg ~ am + wt, mtcars)
fit3 <- lm(mpg ~ am + wt + hp, mtcars)
fit4 <- lm(mpg ~ am + wt + hp + disp, mtcars)
fit5 <- lm(mpg ~ am + wt + hp + disp + cyl, mtcars)
```

```r
anova(fit1, fit2, fit3, fit4, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + hp
## Model 4: mpg ~ am + wt + hp + disp
## Model 5: mpg ~ am + wt + hp + disp + cyl
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 70.5432 7.017e-09 ***
## 3     28 180.29  1     98.03 15.6250 0.0005286 ***
## 4     27 179.91  1      0.38  0.0611 0.8066730
## 5     26 163.12  1     16.79  2.6758 0.1139322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix

Fig. 1

```r
# Factorize
mtcars$am <- factor(mtcars$am, labels = c("AT", "MT"))
ggpairs(mtcars[, c(1, 9, 6, 4)], aes(color = am, alpha = .4))
```

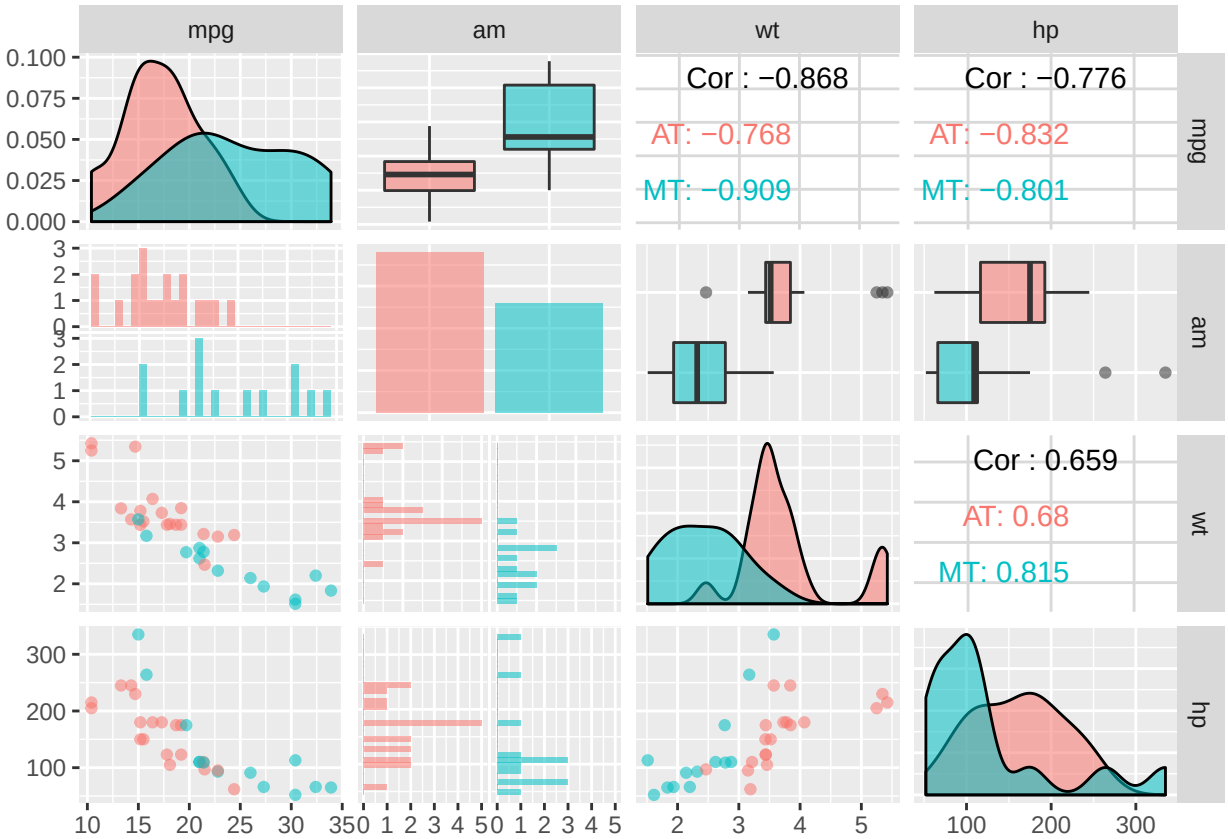

Fig. 2

```
# Residuals vs Fitted
plot1 <- ggplot(fit3, aes(.fitted, .resid)) +
        geom_point() +
        geom_hline(yintercept = 0) +
        geom_smooth(se = FALSE) +
        ggtitle("Residuals vs Fitted")
# Normal Q-Q
plot2 <- ggplot(fit3) +
        stat_qq(aes(sample = .stdresid)) +
        geom_abline() +
        ggtitle("Normal Q-Q")
# Scale-Location
plot3 <- ggplot(fit3, aes(.fitted, sqrt(abs(.stdresid)))) +
        geom_point() +
        geom_smooth(se = FALSE) +
        ggtitle("Scale-Location")
# Standardized Residuals vs Leverage
plot4 <- ggplot(fit3, aes(.hat, .stdresid)) +
        geom_point(aes(size = .cooksd)) +
        geom_smooth(se = FALSE) +
        ggtitle("Residuals vs Leverage")

grid.arrange(plot1, plot2, plot3, plot4, ncol = 2)
```