Regression model - Motor trend

Goga

2023-06-13

Executive Summary

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:

First, let's check the average.

```
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
```

The MT car seems to have a higher MPG.

```
par(mar=c(5,4,3,4))
boxplot(mpg~am,data=mtcars,xlab="Fig 3. Transmission Type : 0-Auto, 1-
Manual",ylab="Mileage mpg",
        main="MPG Vs. Transmission Design")
```

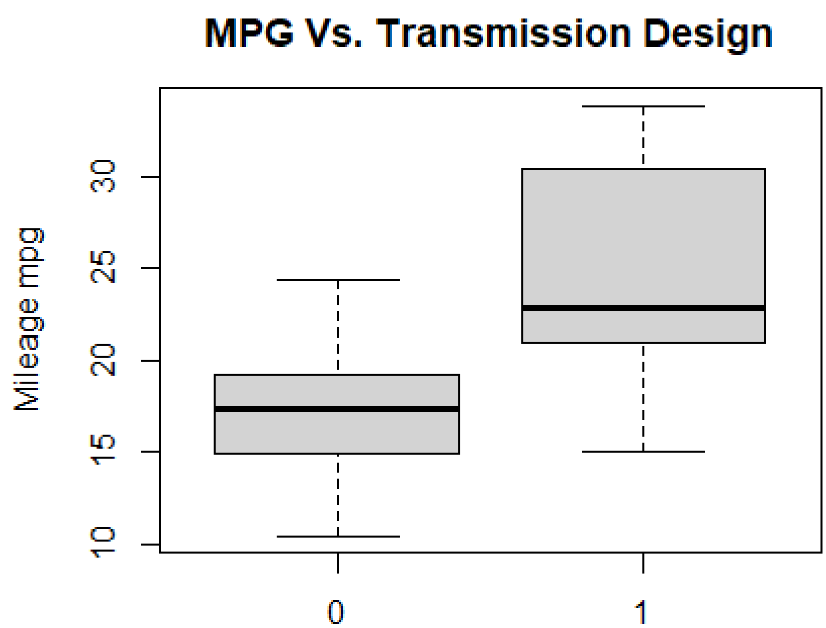


Fig 3. Transmission Type : 0-Auto, 1-Manual

Here we see the boxplot that proves the average calculations for automatic and manual MPG.

Calculate correlation:

Calculate the correlation to see the relationship with other elements.

```
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
```

wt, disp, cyl and hp show high correlation.

Fit Multiple Regression Models

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

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

## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + disp
## Model 4: mpg ~ am + wt + disp + cyl
```

```
## Model 5: mpg ~ am + wt + disp + cyl + hp
##   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     28 246.56  1     31.76  5.0628  0.033130 *
## 4     27 188.43  1     58.13  9.2655  0.005289 **
## 5     26 163.12  1     25.31  4.0336  0.055097 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The **Model 4** p-value is near 0.005, so we will not reject the hypothesis. Model 4 (`fit4`) will fit better.

```
betterFit <- fit4
summary(betterFit)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + disp + cyl, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.318 -1.362 -0.479  1.354  6.059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.898313   3.601540  11.356 8.68e-12 ***
## am           0.129066   1.321512   0.098  0.92292
## wt          -3.583425   1.186504  -3.020  0.00547 **
## disp         0.007404   0.012081   0.613  0.54509
## cyl         -1.784173   0.618192  -2.886  0.00758 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.642 on 27 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8079
## F-statistic: 33.59 on 4 and 27 DF,  p-value: 4.038e-10
```

This Multivariable Regression test now gives us an R-squared value of over .83, suggesting that 83% or more of variance can be explained by the multivariable model. P-values for `cyl` and `wt` are below 5%, suggesting that these are confounding variables in the relation between car Transmission and MPG.

Residual and Diagnostics

Now, we examine residual plots of our regression model and compute some of the regression diagnostics of our model to uncover outliers in the data set.

Appendix

Figure 1

```
# Factorize
mtcars$am <- factor(mtcars$am, labels = c("AT", "MT"))
ggpairs(mtcars[, c(1, 9, 2, 6)], aes(color = am, alpha = .4))
```

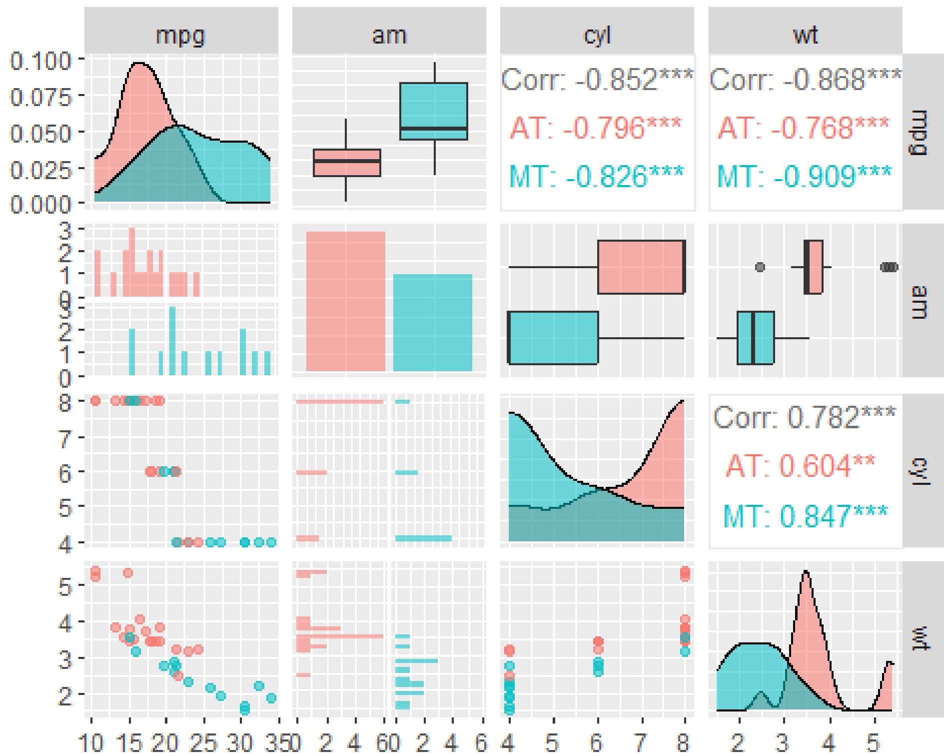
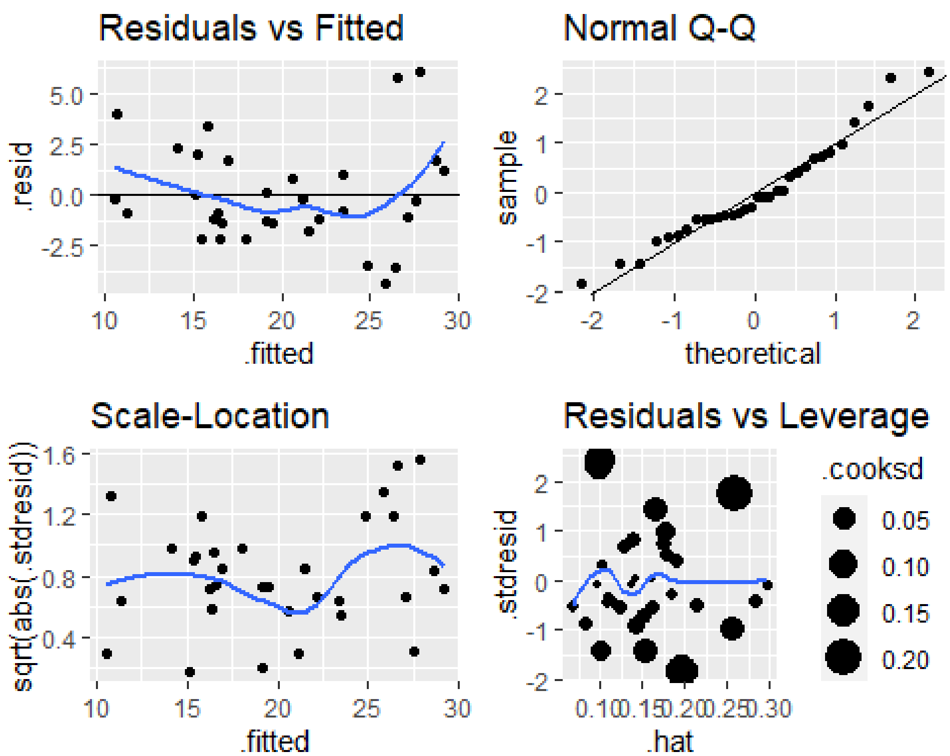


Figure 2

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

```
        geom_point(aes(size = .cooksd)) +
        geom_smooth(se = FALSE) +
        ggtitle("Residuals vs Leverage")

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



From Appendix Figure 2, we can make the following observations,

- The points in the Residuals vs Fitted plot seem to be randomly scattered on the plot and verify the independence condition.
- 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.