

Regression Models Course Project

echf

Sep-2021

Executive Summary

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:

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

Exploring the dataset

```
echo = TRUE
options(width=80)
knitr::opts_chunk$set(out.width="400px", dpi=120)
library(ggplot2) #for plots
```

```
head(mtcars)
```

```
##           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
```

```
data(mtcars)
summary(mtcars)
```

```
##           mpg           cyl           disp           hp
##  Min.      :10.40   Min.      :4.000   Min.      : 71.1   Min.      : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean     :20.09   Mean     :6.188   Mean     :230.7   Mean     :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.     :33.90   Max.     :8.000   Max.     :472.0   Max.     :335.0
##           drat           wt           qsec           vs
```

```
## Min. :2.760 Min. :1.513 Min. :14.50 Min. :0.0000
## 1st Qu.:3.080 1st Qu.:2.581 1st Qu.:16.89 1st Qu.:0.0000
## Median :3.695 Median :3.325 Median :17.71 Median :0.0000
## Mean :3.597 Mean :3.217 Mean :17.85 Mean :0.4375
## 3rd Qu.:3.920 3rd Qu.:3.610 3rd Qu.:18.90 3rd Qu.:1.0000
## Max. :4.930 Max. :5.424 Max. :22.90 Max. :1.0000
##      am      gear      carb
## Min. :0.0000 Min. :3.000 Min. :1.000
## 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:2.000
## Median :0.0000 Median :4.000 Median :2.000
## Mean :0.4062 Mean :3.688 Mean :2.812
## 3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:4.000
## Max. :1.0000 Max. :5.000 Max. :8.000
```

```
# Transform certain variables into factors
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"))
### view From appendix fig. 1 MPG Distribution
```

Regression Analysis

```
aggregate(mpg~am, data = mtcars, mean)
```

We've visually seen that automatic is better for MPG, but we will now quantify his difference. From appendix fig. 3

```
##      am      mpg
## 1 Automatic 17.14737
## 2   Manual 24.39231
```

```
fit_simple <- lm(mpg ~ factor(am), data=mtcars)
summary(fit_simple)
```

We will use mpg as the dependent variable and am as the independent variable to fit a linear regression, where Beta1 is the group mean for automatic and Beta0 is the intercept.

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    17.147      1.125  15.247 1.13e-15 ***
## factor(am)Manual  7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

It shows that on average, a car has 17.147 mpg with automatic transmission, and if it is manual transmission, 7.245 mpg is increased. This model has the Residual standard error as 4.902 on 30 degrees of freedom. And the Adjusted R-squared value is 0.3385, which means that the model can explain about 34% of the variance of the MPG variable. The low Adjusted R-squared value also indicates that other variables should be added to the model.

Anova test and Residuals

```
init <- lm(mpg ~ am, data = mtcars)
###summary(init)
betterFit <- lm(mpg~am + cyl + disp + hp + wt, data = mtcars)
####betterFit <- lm(mpg ~ am + wt + qsec, data = mtcars)
anova(init, betterFit)
```

Finally, the final model is selected.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp + hp + wt
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      30 720.90
## 2      25 150.41  5    570.49 18.965 8.637e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This results in a p-value of 8.637e-08, and we can claim the betterFit model is significantly better than our init simple model. We double-check the residuals for non-normality and can see they are all normally distributed and homoskedastic.

Residual Analysis and Diagnostics

According to the residual plots, the following underlying assumptions can be varified: From appendix fig. 3

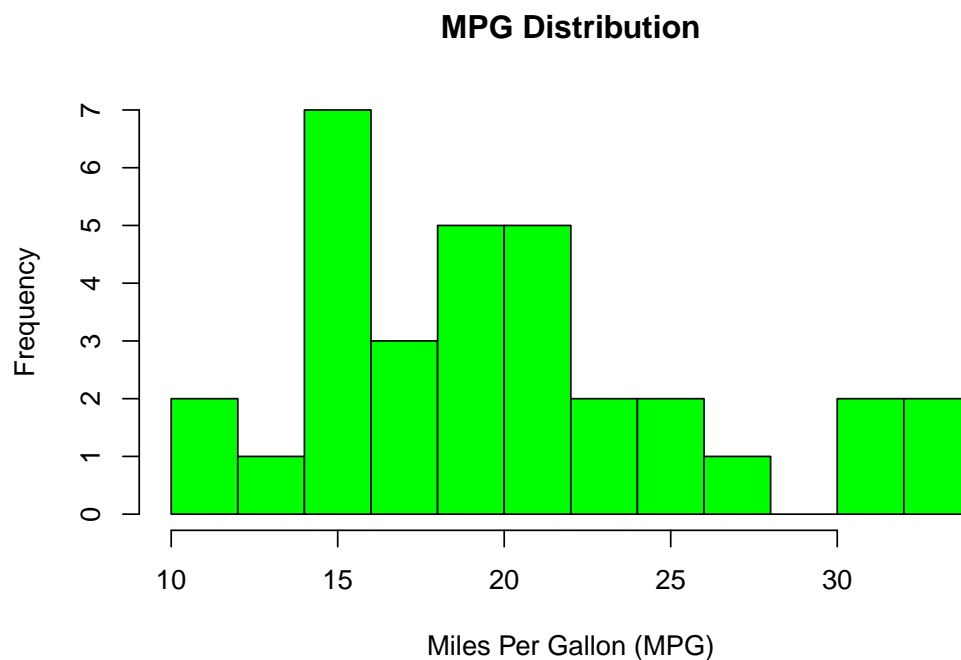
1. The Residuals vs. Fitted plot shows no consistent pattern, supporting the accuracy of the independence assumption.

2. The Normal Q-Q plot indicates that the residuals are normally distributed because the points lie closely to the line.
3. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed.
4. The Residuals vs. Leverage argues that no outliers are present, as all values fall well within the 0.5 bands.

Appendix: Figures

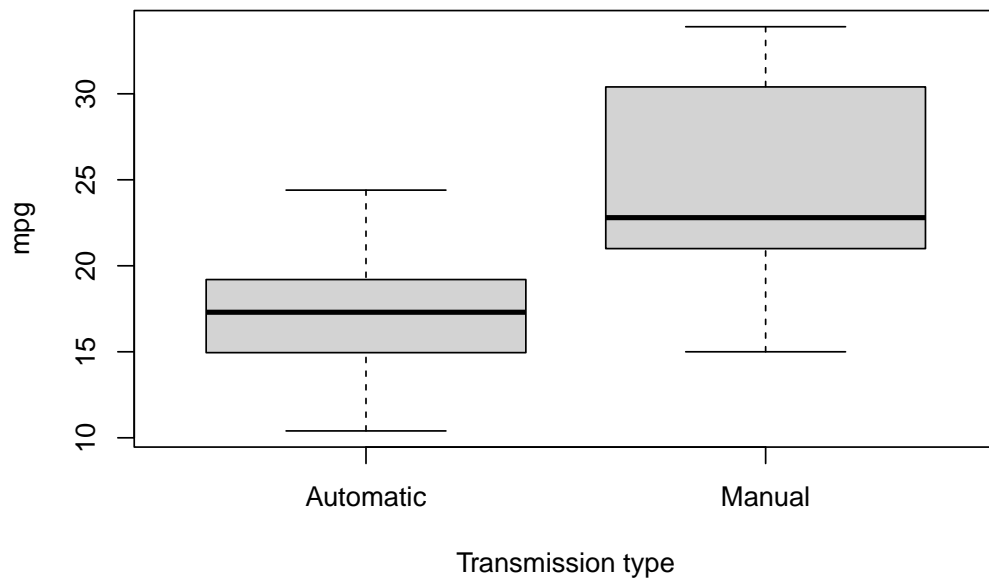
MPG Distribution fig. 1

```
## Histogram of MPG
hist(mtcars$mpg, breaks=12, xlab="Miles Per Gallon (MPG)", main="MPG Distribution",
     col="green")
```



Transmission type fig. 2

```
boxplot(mpg ~ am, data = mtcars, xlab = "Transmission type")
```



Residual plots fig. 3

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

