Regression_models

adler_jai

01/11/2020

Executive Summary

This report analyzed the relationship between transmission type (manual or automatic) and miles per gallon (MPG). The report set out to determine which transmission type produces a higher MPG. The `mtcars` dataset was used for this analysis. A t-test between automatic and manual transmission vehicles shows that manual transmission vehicles have a 7.245 greater MPG than automatic transmission vehicles. After fitting multiple linear regressions, analysis showed that the manual transmission contributed less significantly to MPG, only an improvement of 1.81 MPG. Other variables, weight, horsepower, and number of cylinders contributed more significantly to the overall MPG of vehicles. ## Initial setup

```
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggplot2)
data(mtcars)
```

Let us look at the data.

```
head(mtcars)
```

Exploratory Analysis

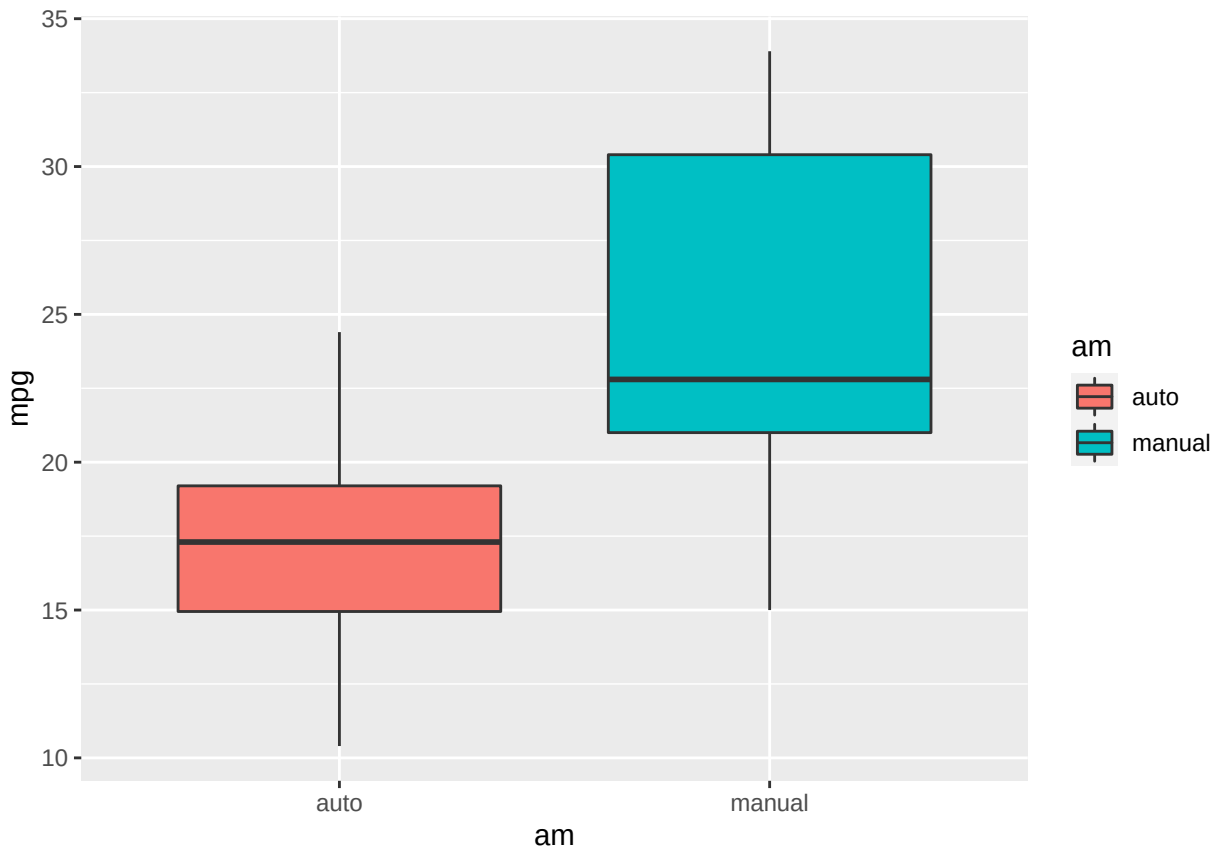
Load the dataset and convert categorical variables to factors.

```
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- gsub("0", "auto", mtcars$am)
mtcars$am <- gsub("1", "manual", mtcars$am)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
attach(mtcars)
```

See Plot I Exploratory Box graph that compares Automatic and Manual transmission MPG. The graph leads us to believe that there is a significant increase in MPG when for vehicles with a manual transmission vs automatic.

I

```
g <- ggplot(mtcars,aes(x = am, y = mpg, fill = am)) + geom_boxplot()
g
```



Statistical Inference

T-Test transmission type and MPG

```
testResults <- t.test(mpg ~ am)
testResults
```

```
##
##  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 auto mean in group manual
##             17.14737             24.39231
```

The T-Test rejects the null hypothesis that the difference between transmission types is 0.

```
testResults$estimate[2]-testResults$estimate[1]
```

```
## mean in group manual
##             7.244939
```

The difference estimate between the 2 transmissions is 7.24494 MPG in favor of manual.

Regression Analysis

Fit the full model of the data

```
fullModelFit <- lm(mpg ~ ., data = mtcars)
summary(fullModelFit)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## ammanual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

Since none of the coefficients have a p-value less than 0.05 we cannot conclude which variables are more statistically significant.

As we can see in above summary `cyl` has most effective on mpg than am. But doing analysis on only am gives slight affect to `mpg`.

Excluding variables that are correlated with transmission type will introduce bias in the coefficients. However, including unnecessary regressors will inflate the model's variance. We will use the step function in R to determine which variables to include in our final model.

```
step_model <- step(fullModelFit, direction = "backward", trace = FALSE)
summary(step_model)
```

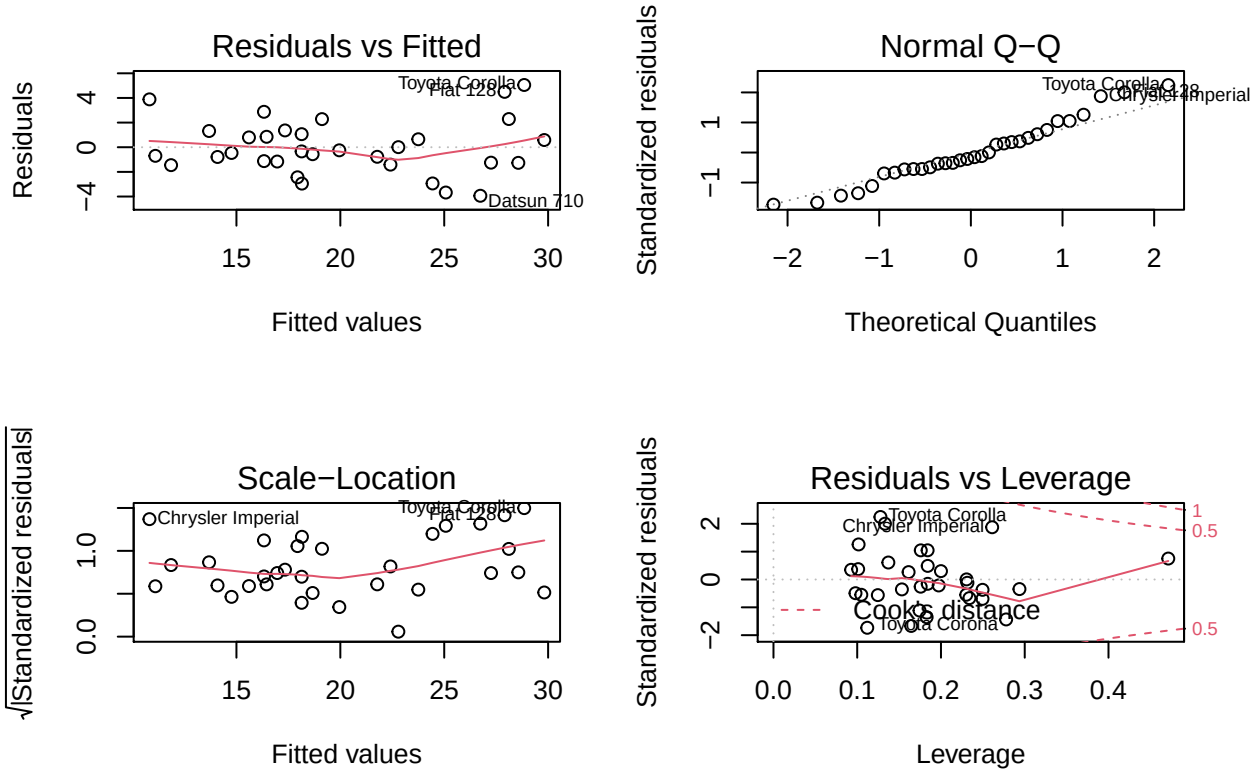
```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## 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 new model has 4 variables (cylinders, horsepower, weight, transmission). The R-squared value of 0.8659 confirms that this model explains about 87% of the variance in MPG. The p-values also are statistically significantly because they have a p-value less than 0.05. The coefficients conclude that increasing the number of cylinders from 4 to 6 with decrease the MPG by 3.03. Further increasing the cylinders to 8 with decrease the MPG by 2.16. Increasing the horsepower is decreases MPG 3.21 for every 100 horsepower. Weight decreases the MPG by 2.5 for each 1000 lbs increase. A Manual transmission improves the MPG by 1.81.

Residuals & Diagnostics

Residual Plot **See Plot II**

II

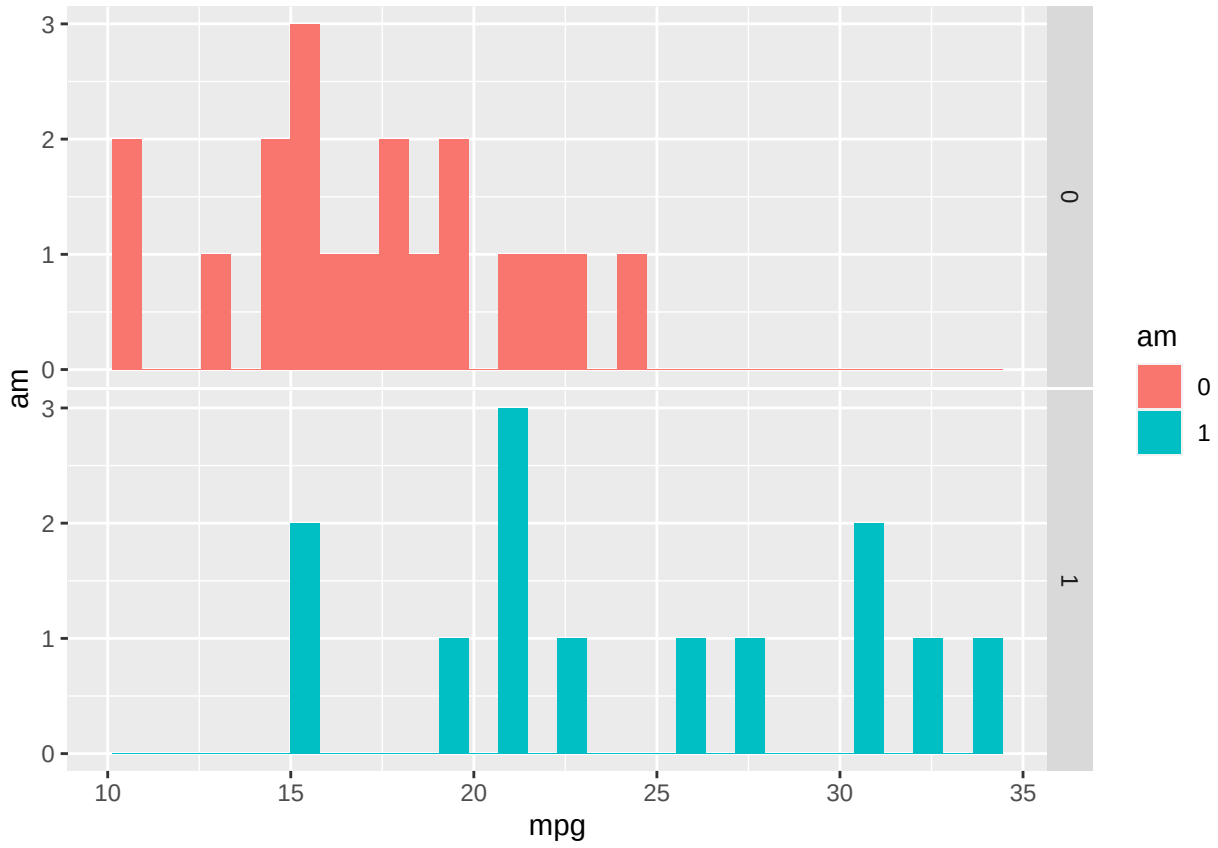


The plots conclude:

1. The randomness of the Residuals vs. Fitted plot supports the assumption of independence
2. The points of the Normal Q-Q plot following closely to the line conclude that the distribution of residuals is normal
3. The Scale-Location plot random distribution confirms the constant variance assumption
4. Since all points are within the 0.05 lines, the Residuals vs. Leverage concludes that there are no outliers

```
data(mtcars)
mtcars$am <- factor(mtcars$am)
ggpairs(mtcars, mapping = aes(colour = am))[9,1]
```

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



Conclusion

There is a difference in MPG based on transmission type. A manual transmission will have a slight MPG boost. However, it seems that weight, horsepower, & number of cylinders are more statistically significant when determining MPG.