Project - Behnaz Refahi

2023-12-17

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.

Load the dataset and convert categorical variables to factors.

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

Exploratory Data Analysis:

```
# Boxplot for MPG by Transmission Type
ggplot(mtcars, aes(x = factor(am), y = mpg)) +
  geom_boxplot() +
  labs(x = "Transmission Type", y = "Miles per Gallon", title = "MPG by Transmission Type") +
  scale_x_discrete(labels = c("0" = "Automatic", "1" = "Manual"))
```

The median MPG of manual cars, not only higher but also outside the box range of automatics, leads us to believe that there is a significant increase in MPG when for vehicles with a manual transmission vs automatic. Moreover, manual cars show a positive skew, in contrast to the normal distribution seen in automatics.

T-Test transmission type and MPG

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

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

```
testResults$estimate
```

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

Since outcome is not binary, count, or rate, we use the usual linear regression model.

Fit the full model of the data

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

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

Backward selection to determine which variables are most statistically significant

```r
stepFit <- step(fullModelFit)
summary(stepFit) # results hidden
```

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.

Check the model with only transmission as the predictor

```r
# Modeling: simple linear regression with am as the predictor
model <- lm(mpg ~ am, data = mtcars)
summary(model)
```

Intercept (17.147): This represents the estimated mean value of the automatic cars' MPG. It means that if a car has an automatic transmission (am = 0), its MPG rating is expected to be approximately 17.147, on average, when other factors are held constant. Coefficient for am (7.245): This coefficient represents the average difference in mpg between manual and automatic transmissions. The positive value (7.245) indicates that, on average, cars with manual transmission (am = 1) have a higher MPG rating compared to automatic ones by about 7.245 miles per gallon.

The analysis reveals that, on average, cars with automatic transmission exhibit 17.147 mpg, while those with manual transmission experience an increase of 7.245 mpg. An Adjusted R-squared value of 0.3385, indicating that the model explains approximately 34% of the variance in the MPG variable, the relatively low Adjusted R-squared value suggests a need for the incorporation of additional variables into the model.

Residual Analysis and Diagnostics

```r
# Plotting residuals
par(mfrow = c(2, 2))
plot(stepFit)
```

Upon examination of the residual plots, we can validate the following underlying assumptions:

• The Residuals vs. Fitted plot reveals no discernible pattern, affirming the accuracy of the independence assumption. • The Normal Q-Q plot portrays residuals conforming to a normal distribution, with points closely aligning along the line. • The Scale-Location plot attests to the constancy of variance, as the points are randomly dispersed. • The Residuals vs. Leverage plot suggests the absence of outliers, as all values comfortably fall within the 0.5 bands.

We can check the outlier in another way too:

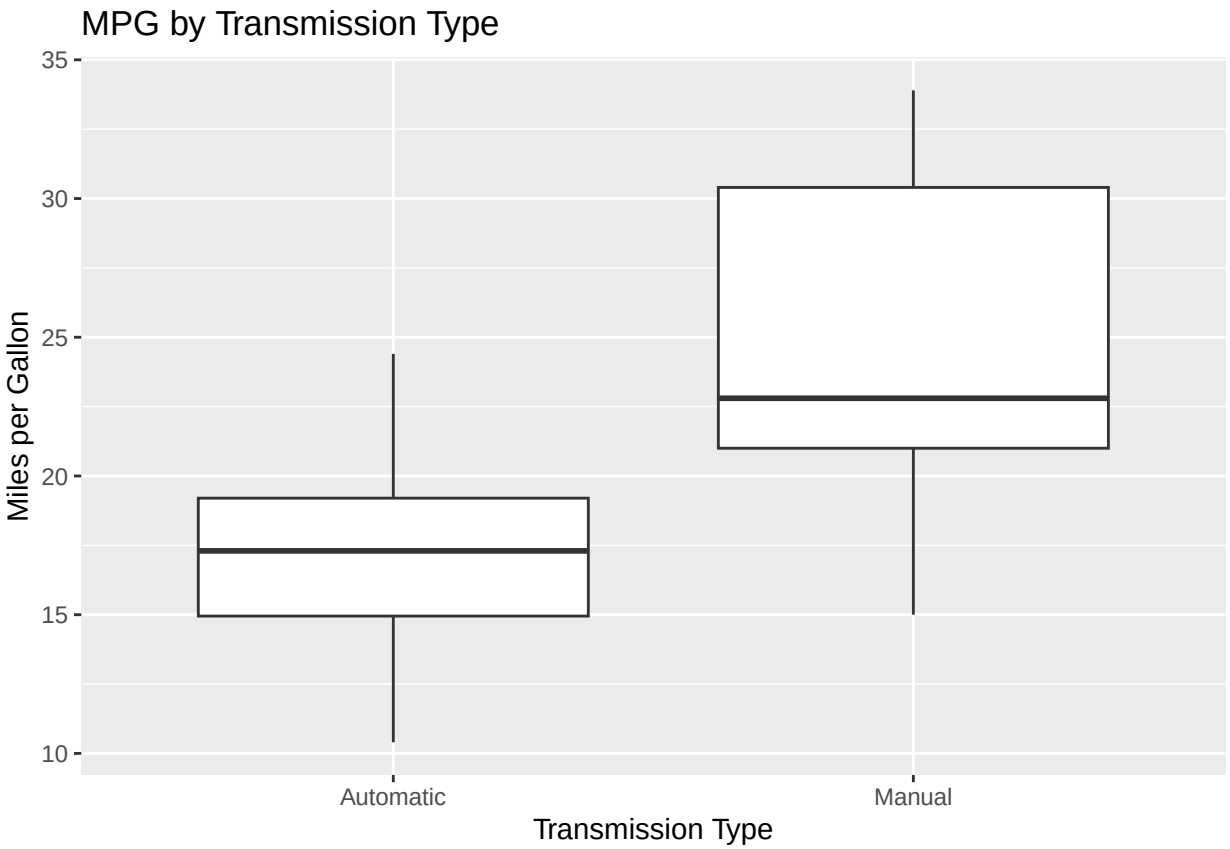
```r
round(dfbetas(stepFit)[1 : 32, 2], 3)
```

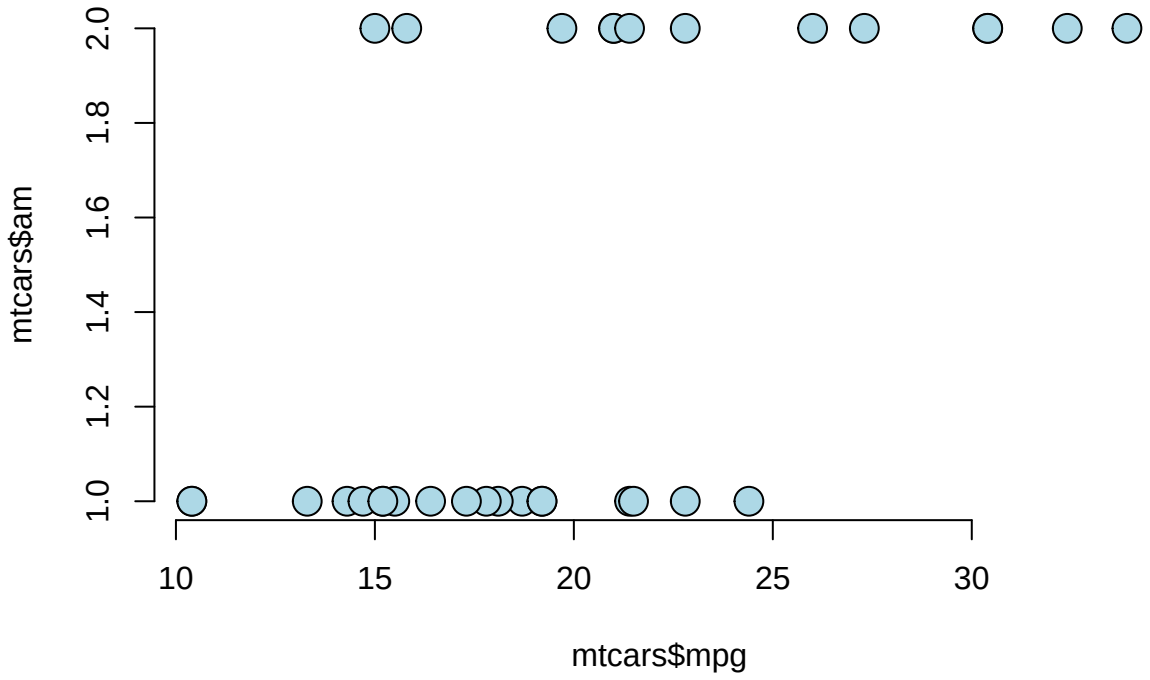
Dfbeta values for some cars show minor deviations from others, yet not significantly enough to label them as outliers. Further diagnostic tests are warranted to confirm these findings.

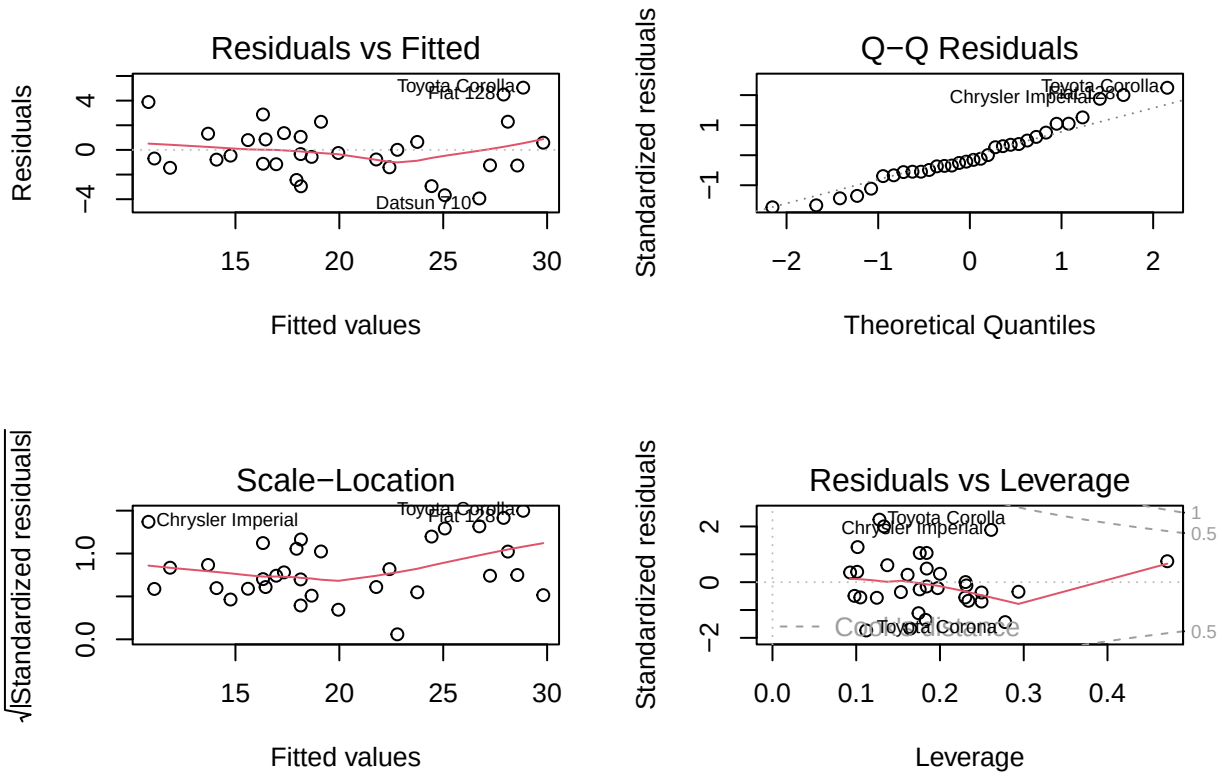
```r
round(hatvalues(stepFit)[1 : 32], 3)
```

Since all the hat values of the cars which had different Dfbeta values are almost the same, we do not opt out any point.

Appendix:


MPG by Transmission Type
35
30
25
20
15
10
Miles per Gallon
Automatic
Manual
Transmission Type


mtcars$am
1.0
1.2
1.4
1.6
1.8
2.0
10
15
20
25
30
mtcars$mpg


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
Toyota Corona