Regression Models Course Project

Rodrigo Falcão

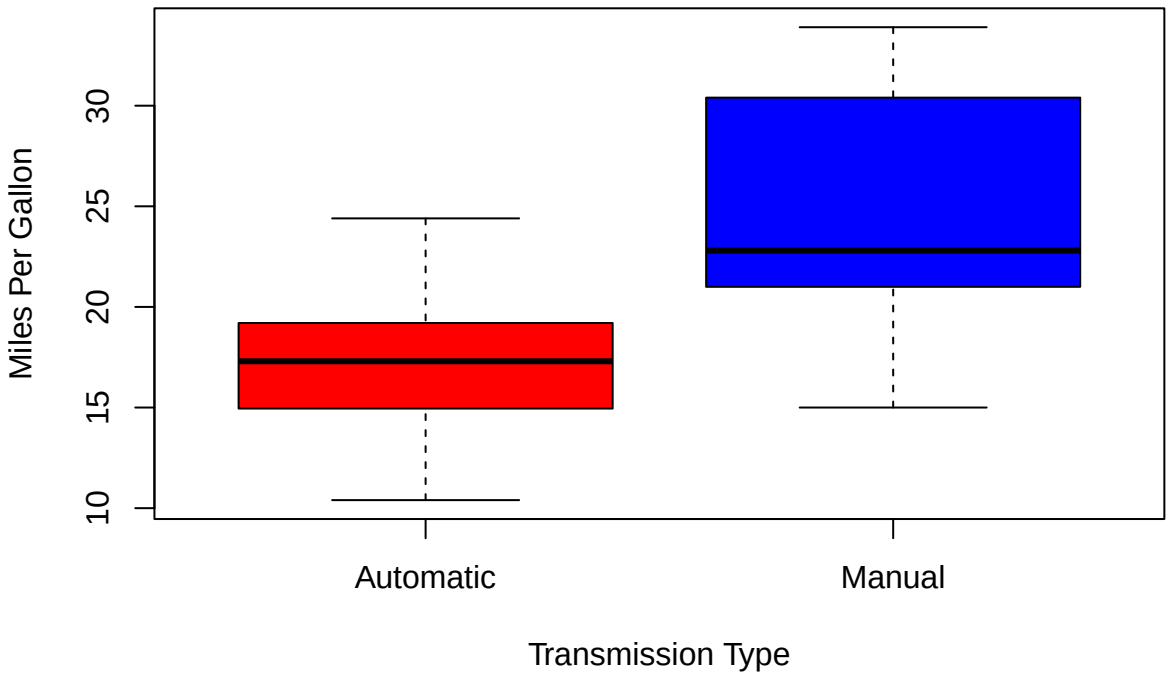
2023-03-23

Exploratory Data Analysis

First we have to cast categorical variables to factor and check the box plot between "mpg" and "am".

```
factor_cols <- c("vs", "gear", "carb")
mtcars[factor_cols] <- lapply(mtcars[factor_cols], factor)
mtcars$am <- factor(mtcars$am, labels=c("Automatic","Manual"))

boxplot(mpg ~ am, data = mtcars, col = (c("red","blue")), ylab = "Miles Per Gallon", xlab = "Transmissi
```



From the plot is it possible to see that the automatic transmissions tends to be worse than manual transmission when it comes to fuel consumption (mpg). # Statistical Inference

```
t.test(mpg ~ am, mtcars)
```

```
##
##  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 between group Automatic and group Manual is not equa
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

From the T test is it possible to see that manual transmission is better when it comes to "mpg". But since the mean value itself does not represent a concrete proof that the manual transmission is better it is necessary to fit the data to a regression model. Let's start fitting just with the factor variable of transmission.

Regression model

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

```
##
## 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
```

From the summary is it possible to see that the P-value is less than 0.0003, so the hypothesis is not rejected. Although the P-value is low the R-squared value for this test is approximately 0.35, indicating that around one-third of the variation in "mpg" can be explained by transmission type alone. As a next step, let's conduct an Analysis of Variance for the dataset.

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

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6533 -1.3325 -0.5166  0.7643  4.7284
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.31994   23.88164   1.060   0.3048
## cyl         -1.02343    1.48131  -0.691   0.4995
## disp         0.04377    0.03058   1.431   0.1716
## hp          -0.04881    0.03189  -1.531   0.1454
## drat         1.82084    2.38101   0.765   0.4556
## wt          -4.63540    2.52737  -1.834   0.0853 .
## qsec         0.26967    0.92631   0.291   0.7747
## vs1          1.04908    2.70495   0.388   0.7032
## amManual     0.96265    3.19138   0.302   0.7668
## gear4        1.75360    3.72534   0.471   0.6442
## gear5        1.87899    3.65935   0.513   0.6146
## carb2       -0.93427    2.30934  -0.405   0.6912
## carb3        3.42169    4.25513   0.804   0.4331
## carb4       -0.99364    3.84683  -0.258   0.7995
## carb6        1.94389    5.76983   0.337   0.7406
## carb8        4.36998    7.75434   0.564   0.5809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.823 on 16 degrees of freedom
## Multiple R-squared:  0.8867, Adjusted R-squared:  0.7806
## F-statistic: 8.352 on 15 and 16 DF,  p-value: 6.044e-05
```

The full model has an R-squared value of 0.8867, as expected. However, based on the summary output, none of the coefficients are statistically significant at the 0.05 level.

Removing variables that are correlated with transmission type could result in biased coefficients, while adding unnecessary regressors can increase the model's variance. To determine the appropriate variables to include in our final model, we will use the step function in R.

```
fit_optimal <- step(fit_full, direction = "backward", trace = FALSE)
summary(fit_optimal)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
```

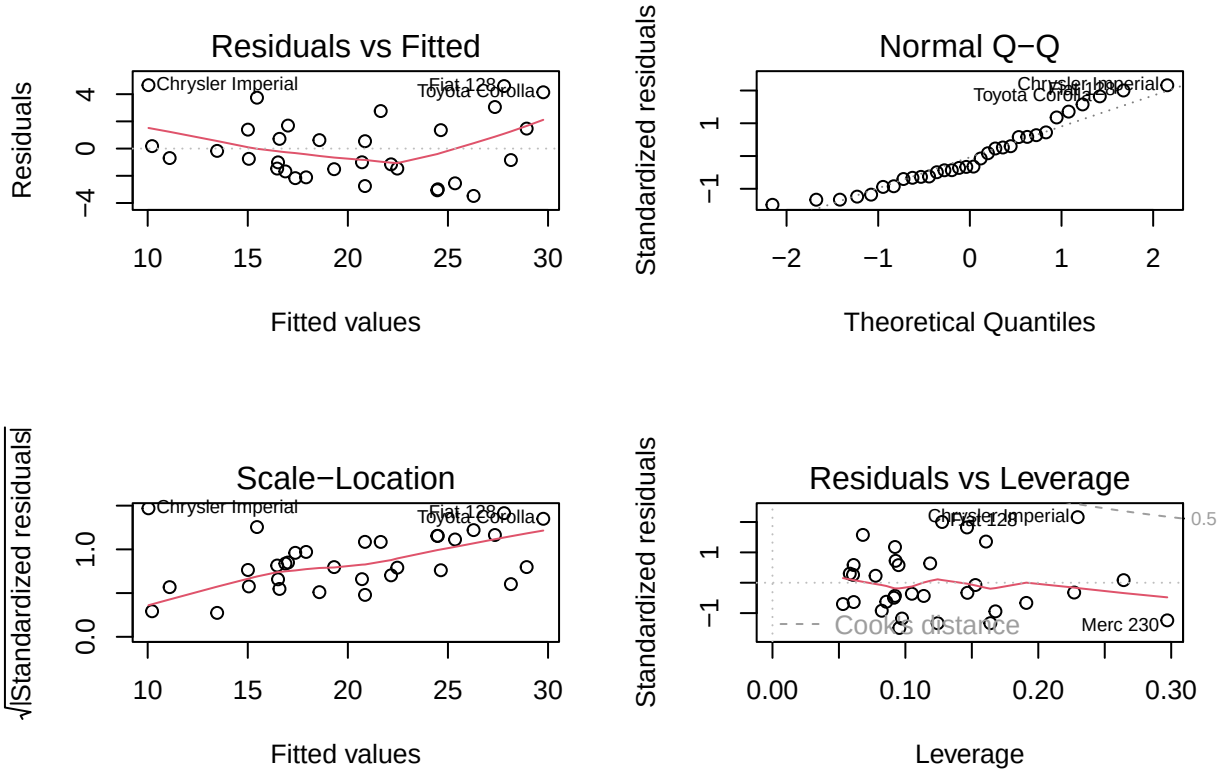
```
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

The step algorithm generated a model that includes the variables "wt", "qsec", and "am." All three variables have a significant impact at the 0.05 level, and the model accounts for around 85% of the variation.

Based on the coefficients of this model, we can infer that manual cars, on average, achieve 2.9 more miles per gallon than automatic cars, assuming "weight" and "qsec" remain constant.

The base graphics diagnostic plots demonstrate that there is no correlation between the residuals and the fitted values. Additionally, the quantile-quantile plot suggests that the residuals follow a normal distribution.

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



Appendix

```
ggpairs(mtcars, mapping = aes(colour = am))
```