Analysis of the relationship between transmission type and mileage

Ion Scerbatiuc

May 21, 2016

Summary

By using the `mtcars` dataset from the `Motor Trend` magazine, we're trying to understand what is the relationship between the transmission type and the mileage of cars and how can we quantify it.

We start with an exploratory analysis and a simple linear model that shows and average increase of fuel consumption for automatic cars of `7.25 mpg` compared to manual cars. After using a stewise model selection algorithm we find a better model that explains about `84%` of the mileage variation. Using that model, we conclude that the difference in consumption between automatic and manual cars (`1.81 mpg`) is not statistically significant.

Exploratory Analysis

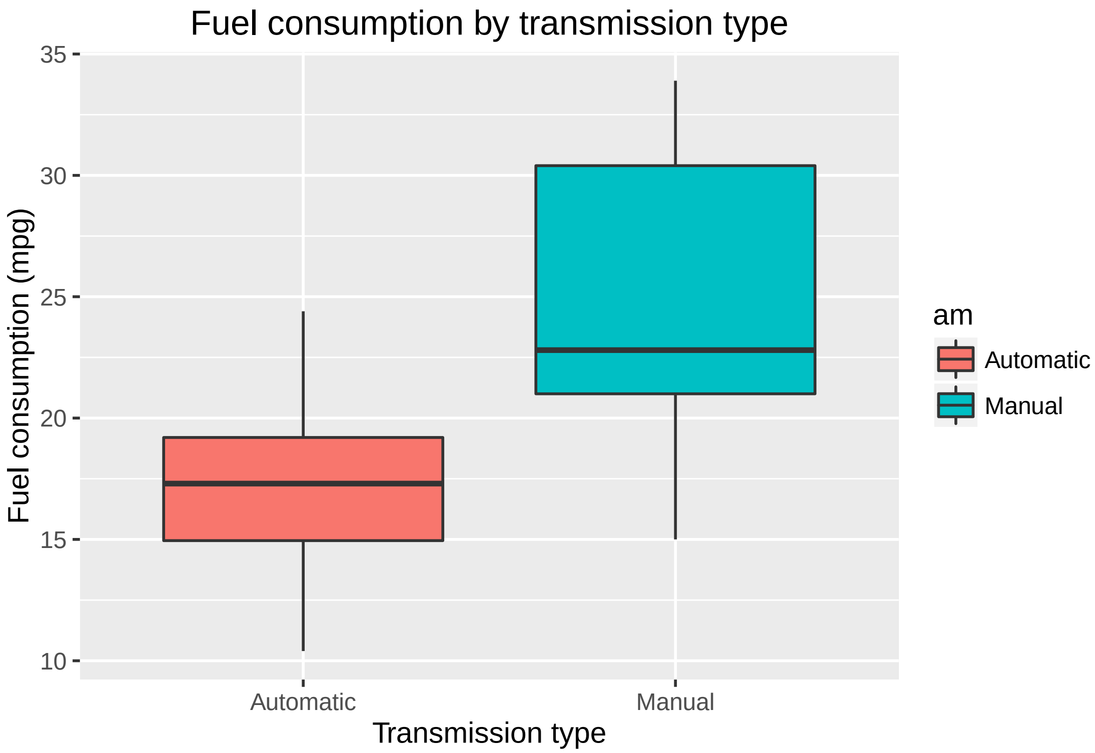
```r
library(datasets)
library(ggplot2)
library(dplyr)
data(mtcars)
```

The first step is to identify the variables we have and what values they take (Appendix 1). We can see that we have a few variables that are non-contiguous so we need to make them as factor variables, to prevent `lm` to interpret them on a contiguous scale, and thus distort their meaning.

```r
mtcars = mutate(
    mtcars,
    am = as.factor(am), cyl = as.factor(cyl), vs = as.factor(vs),
    gear = as.factor(gear), carb = as.factor(carb)
)
levels(mtcars$am) <- c('Automatic', 'Manual')
```

Next, we want to explore what is the measured fuel consumption by transmission type, to see if there is indeed some level of corelation between the two.

```r
ggplot(mtcars, aes(x = am, y = mpg, fill = am)) + geom_boxplot() +
    xlab("Transmission type") +
    ylab("Fuel consumption (mpg)") +
    ggtitle('Fuel consumption by transmission type')
```



From the plot above we can see that automatic cars are consuming on average more fuel than manual transmission cars. But is that result statistically significant? To answer that question, let's fit a linear model whit `mpg` as the outcome and the transmission type as the predictor.

```
fit <- lm(mpg ~ am, mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ 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 ***
## amManual       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
```

By inspecting the coefficients table we can conclude that manual cars have an expected `7.25 mpg` higher mileage than the expected mileage of automatic cars. The result is also statistically significant because the p-value for the associated t-test is lower than `0.05`.

However, if we look at the value of R-squared, we see that this model only explains about `33.85%` of the variation. There has to be a better model that explains more of the variation in mileage, so let's dig deeper using a multivariate linear model.

Model Selection

To find the best model from the variables in the dataset, let's use a stepwise model fit algorithm.

```
bestFit <- step(lm(data = mtcars, mpg ~ .), trace=0, steps=1000)
summary(bestFit)
```

```
##
## 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
```

From the results above it seems that by adding the number of cylinders, the gross horsepower and the wieght of the vehicle to the initial model, we obtain a better model that explains about `84%` of the mileage variation. Howerver, we can see that the coefficient for manual transmission is not statistically significant (`p-value > 0.05`), meaning that

To validate the result, we performed an analysis of variance and found a statistically significant difference between the two models (Appendix 2). We also performed residual diagnostics and found that the residuals are normally distributed and mostly patternless (Appendix 3).

Conclusions

By interpreting the coefficients of the `bestFit` model we can conclude that manual transmission cars consume about `1.81 mpg` less fuel than automatic transmission cars, holding all the other caracteristics constant. However the result is not statistically significant, so the difference in fuel consumption between the two transmission types, might be due to noise in the data or some other hidden variables we're missing from the model. It could also be the case that the 32 observations that we have are not enough to reach significance.

Appendixes

Appendix 1 - Dataset definition

```
A data frame with 32 observations on 11 variables.

[, 1]   mpg Miles/(US) gallon
[, 2]   cyl Number of cylinders
[, 3]   disp    Displacement (cu.in.)
[, 4]   hp  Gross horsepower
[, 5]   drat    Rear axle ratio
[, 6]   wt  Weight (1000 lbs)
[, 7]   qsec    1/4 mile time
[, 8]   vs  V/S
[, 9]   am  Transmission (0 = automatic, 1 = manual)
[,10]   gear    Number of forward gears
[,11]   carb    Number of carburetors
```

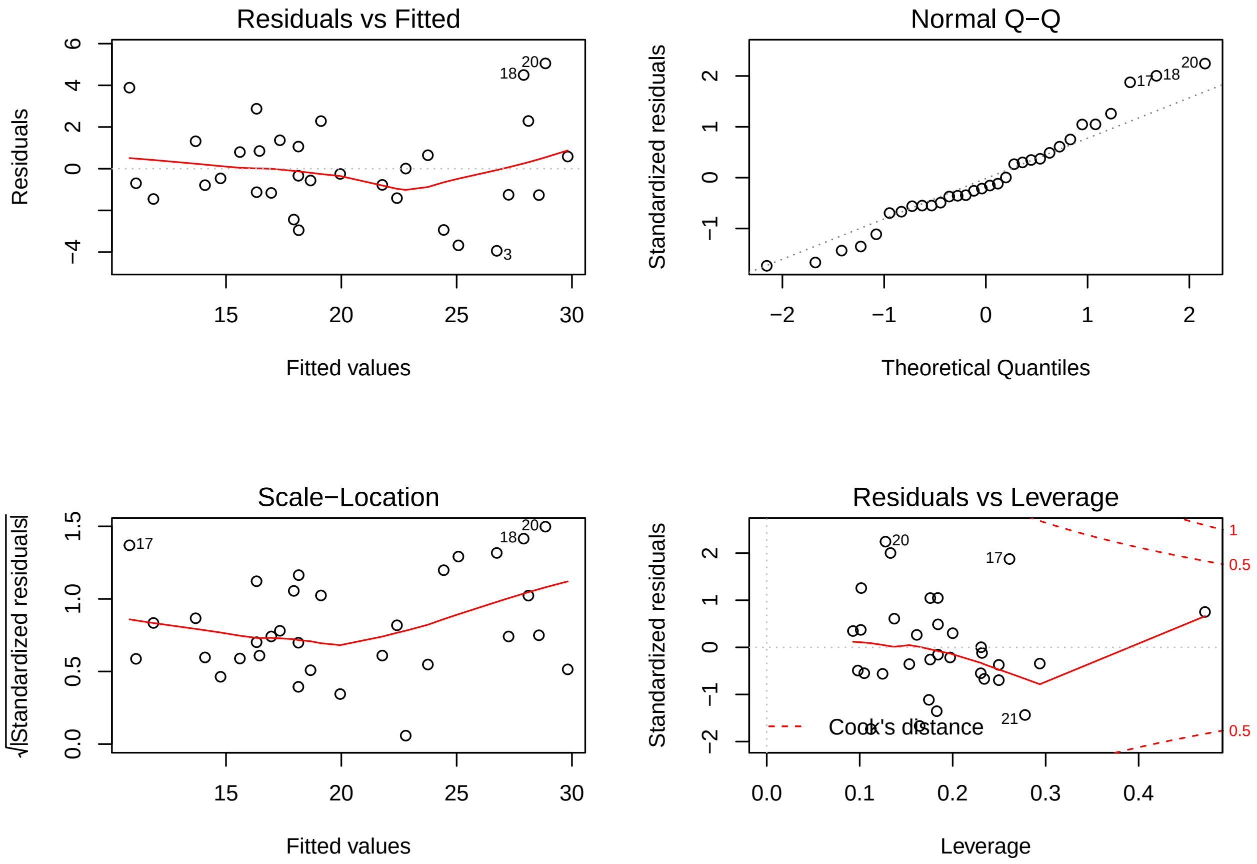
Appendix 2 - Analysis of variance for the fitted model

```r
anova(fit, bestFit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 3 - Residual diagnostics of the fitted model

```r
par(mfrow = c(2,2))
plot(bestFit)
```


Residuals vs Fitted
Residuals
Fitted values
−4
0
2
4
6
15
20
25
30
18
20
3
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
0
1
2
−2
−1
0
1
2
17
18
20
Scale−Location
√|Standardized residuals|
Fitted values
0.0
0.5
1.0
1.5
15
20
25
30
17
18
20
Residuals vs Leverage
Standardized residuals
Leverage
−2
−1
0
1
2
0.0
0.1
0.2
0.3
0.4
20
17
21
Cook's distance
1
0.5
0.5