Effect of transmission type on MPG in cars

Marc T. Henry de Frahan

1 Executive summary

In this report, we analyze the relationship between transmission type (automatic vs manual) and miles per gallon (outcome). Through a simple regression modeling looking just at transmission type, we find that:

1. manual transmission is better for MPG;
2. manual transmission is about 7 MPG better and this difference is statistically significant (p-value < 0.05).

However, we also saw that there is a relationship between the transmission type and the weight of the car. When confounding the model with weight and transmission, we see that the transmission type does not have a significant effect on the mpg.

2 Exploratory Data Analysis

Let's first load the data and format the information we want from the mtcars data set. We can plot the data using a violin plot to get an idea of the differences between the transmission types. A first glance at the data seems to indicate that the automatic transmission leads to fewer miles per gallon. On the other hand, using a manual transmission seems to lead to a better outcome though the spread is more. We will perform linear regression to explore quantitatively the relationship between transmission type and MPG.

```
## 'data.frame':    32 obs. of  12 variables:
##  $ mpg         : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl         : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp        : num  160 160 108 258 360 ...
##  $ hp          : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat        : num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt          : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec        : num  16.5 17 18.6 19.4 17 ...
##  $ vs          : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am          : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear        : num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb        : num  4 4 1 1 2 1 4 2 2 4 ...
##  $ transmission: Factor w/ 2 levels "automatic","manual": 2 2 2 1 1 1 1 1 1 1 ...
```

3 Regression models

3.1 Simple linear regression model based on transmission

Let's perform the regression model to explore the relationship between mpg and transmission type.

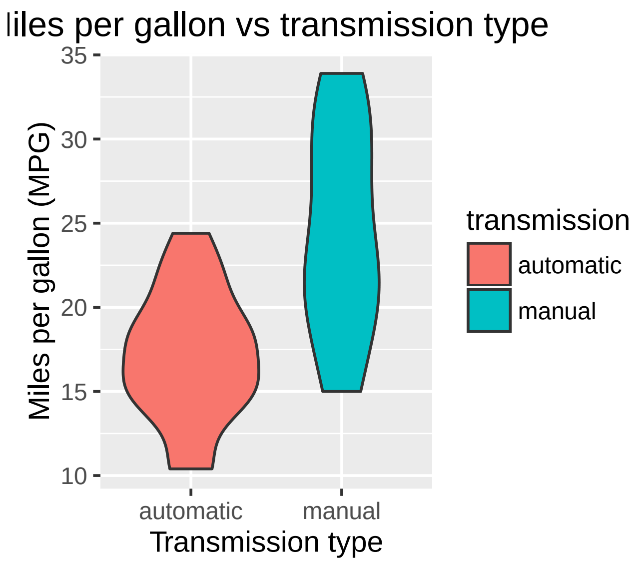

Figure 1: MPG vs transmission

```
##
## Call:
## lm(formula = mpg ~ transmission, data = df)
##
## 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 ***
## transmissionmanual    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
```

Since we are looking at categorical data, the output coefficients should be interpreted as intercepts. For automatic transmission the intercept is 17.1473684. For the manual transmission the intercept is 24.3923077 (the sum of the automatic transmission intercept and the manual transmission coefficient). The manual transmission coefficient indicates the change with respect to the automatic transmission intercept. In this case it is positive, which means that the intercept for manual transmission is greater than the one for the automatic transmission. The difference is also statistically significant since the p-value associated to this coefficient (2.8502074×10^{-4}) is less than 0.05. This implies that we can safely reject the null hypothesis, which is that the means of the different transmission types are equal.

The implications of this model are that manual transmission is better for MPG by about 7.2449393 MPG and this difference is statistically significant.

3.2 Residual and diagnostics

It is important to look at the residuals and other diagnostics to ensure that the model is ok. In the residual plot figure, we show that the residual does not follow any clear pattern, indicating that we are not missing a non-linearity in the model. In the residual diagnostics figures, we can see that the residuals closely follow a normal distribution. We can also see in the residual diagnostic figures that none of the data points have too much influence on the model. This can also be seen by looking at the dfbetas and hatvalues and noticing that none of the points are significantly different from each other. We conclude that the residual diagnostics do not invalidate our model.

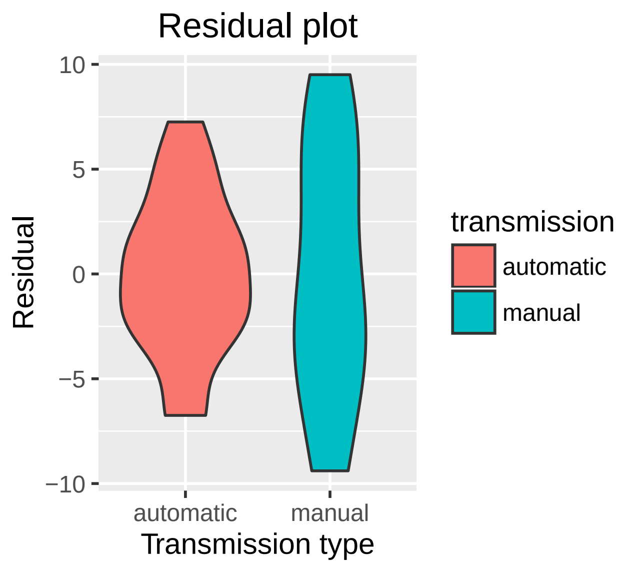

Figure 2: Residual

3.3 Confounding with the weight of the car

```
##
## Call:
## lm(formula = wt ~ transmission, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3039 -0.3694 -0.2049  0.3156  1.6551
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.7689     0.1646  22.895  < 2e-16 ***
## transmissionmanual  -1.3579     0.2583  -5.258 1.13e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7175 on 30 degrees of freedom
## Multiple R-squared:  0.4795, Adjusted R-squared:  0.4622
## F-statistic: 27.64 on 1 and 30 DF,  p-value: 1.125e-05
```

Let's take into account one more variable, the weight of the car. We can see in the figure of weight versus transmission type that automatic cars tend to be heavier. A linear model of weight (outcome) on transmission

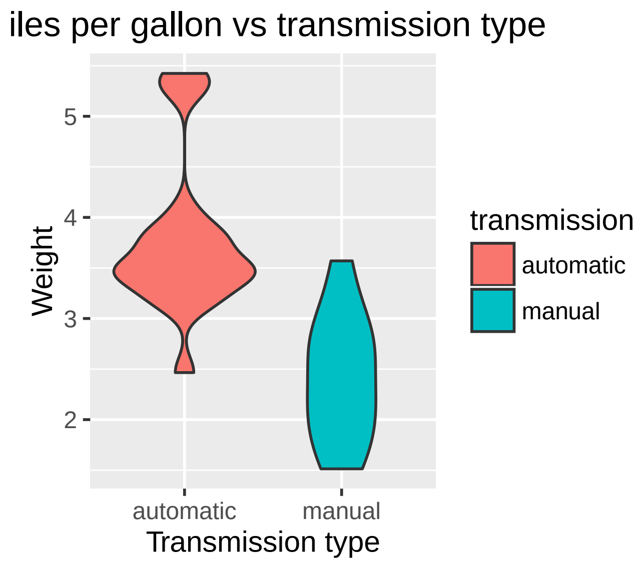

Figure 3: Weight vs transmission

type indicates that this is indeed the case. A manual car is predicted to be -1.3578947 lighter than a manual car. The difference is statistically significant (p-value = $1.1254396 \times 10^{-5} < 0.05$). This implies that we should also look at the weight of the car as a predicting variable.

```
##
## Call:
## lm(formula = mpg ~ transmission + wt, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        37.32155    3.05464  12.218 5.84e-13 ***
## transmissionmanual -0.02362    1.54565  -0.015    0.988
## wt                 -5.35281    0.78824  -6.791 1.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 29 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7358
## F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

This model confounding with the weight and transmission type indicates that there is a linear relationship of mpg on the weight (slope = `r summary(fit.wt.am)$coeff[3,1]`) and the transmission type does not have significant impact on the intercept (p-value = $0.9879146 > 0.05$).

4 Conclusions

For a simple linear model accounting only for transmission type , we can answer both our questions that explore the simple relationship between transmission type and mpg

1. The implications of our simple model are that manual transmission is better for MPG.
2. Manual transmission is about 7.2449393 MPG better and this difference is statistically significant (p-value $= 2.8502074 \times 10^{-4} < 0.05$).

However, when confounding the model with weight and transmission, we see that the transmission type does not have a significant effect on the mpg.

5 Appendix

The version history of this document can be found at the GitHub repository page. Here is the full code used in this document.

```
## ----echo = FALSE, message=FALSE, warning=FALSE--------------------------
library(dplyr)
library(ggplot2)
df <- mtcars %>%
    mutate(transmission = ifelse(am==0,"automatic","manual")) %>%
    mutate(am = as.factor(am)) %>%
    mutate(transmission = as.factor(transmission))
str(df)

## ----echo = FALSE, fig.cap="MPG vs transmission"-------------------------
ggplot(data=df, aes(x=transmission,y=mpg)) +
    geom_violin(aes(fill=transmission)) +
    labs(x = "Transmission type",
         y="Miles per gallon (MPG)",
         title="Miles per gallon vs transmission type");

## ----echo = FALSE, message=FALSE, warning=FALSE--------------------------
fit <- lm(mpg ~ transmission, data = df)
summary(fit)

## ----echo = FALSE, fig.cap="Residual"------------------------------------
df <- df %>%
    mutate(residual = resid(fit))
ggplot(data=df, aes(x=transmission,y=residual)) +
    geom_violin(aes(fill=transmission)) +
    labs(x = "Transmission type",
         y="Residual",
         title="Residual plot");

## ----echo = FALSE, fig.cap="Residual diagnostics"------------------------
## round(dfbetas(fit)[,2],3)
## round(hatvalues(fit),3)
## par(mfrow=c(2,2))
## plot(fit)

## ----echo = FALSE, message=FALSE, warning=FALSE--------------------------
fit.wt <- lm(wt ~ transmission, data = df)
summary(fit.wt)

## ----echo = FALSE, fig.cap="Weight vs transmission"----------------------
ggplot(data=df, aes(x=transmission,y=wt)) +
    geom_violin(aes(fill=transmission)) +
    labs(x = "Transmission type",
         y = "Weight",
         title="Miles per gallon vs transmission type");

## ----echo = FALSE, message=FALSE, warning=FALSE--------------------------
fit.wt.am <- lm(mpg ~ transmission+wt, data = df)
summary(fit.wt.am)
```

```
## ----code=readLines(knitr::purl('./linreg_cars.Rmd', documentation = 1)), eval = FALSE----
## NA
```