What type of transmission is better for the MPG

Alexander Alexandrov

Overview

Work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- "Is an automatic or manual transmission better for MPG?"
- "Quantify the MPG difference between automatic and manual transmissions"

Its evident from my experience that *transimission type* is not the only variable that has impact on *mpg*. So several models should be analized and compared to each over. Following parameters should be considered: transmission type, weight, gross horsepower, number of cylinders, number of forward gears.

Exploratory Analysis

Load Motor Trend cars data. And adjust some variables.

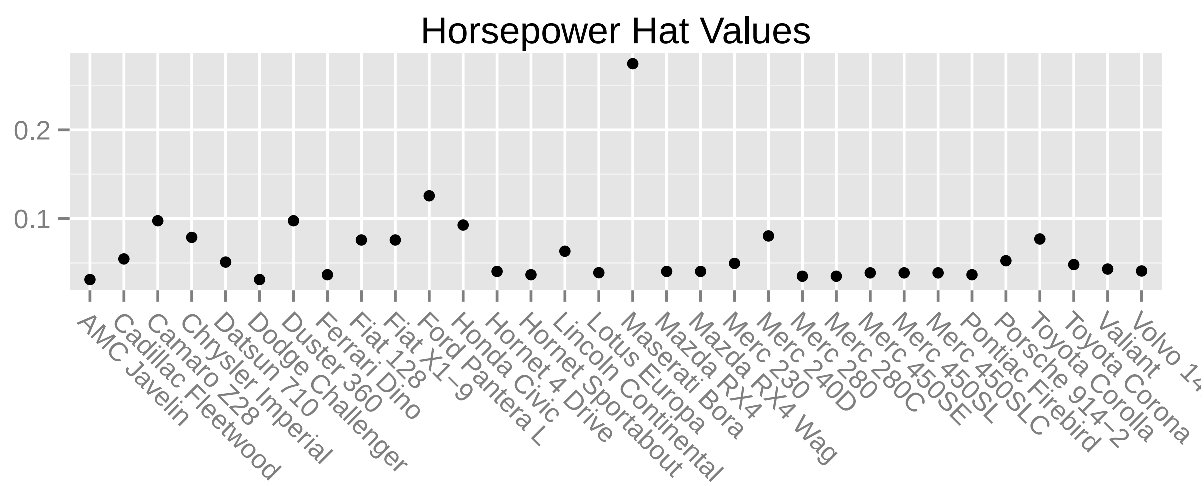
```
## 'data.frame':    32 obs. of  12 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
##  $ id  : chr  "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4 Drive" ...
```

Variable	Type	Domain	Description
mpg	numeric	positive real number	Miles/(US) gallon
cyl	factor	4, 6, 8	Number of cylinders
disp	numeric	positive real number	Displacement (cu.in.)
hp	numeric	positive real number	Gross horsepower
drat	numeric	positive real number	Rear axle ratio
wt	numeric	positive real number	Weight (lb/1000)
qsec	numeric	positive real number	1/4 mile time
vs	factor	0, 1	V/S
am	factor	0, 1	Transmission (0 = automatic, 1 = manual)
gear	factor	3, 4, 5	Number of forward gears
carb	numeric	positive integer number	Number of carburetors

Data Cleaning

Outliers of the *weight* variable

Outliers of the *horsepower* variable



The only evident outliers are: "Cadillac Fleetwood", "Chrysler Imperial", "Lincoln Continental", "Maserati Bora". This is error or extraordinary cases. And they should be removed from the data before the regression analysis to avoid injurious effect.

```
mtcars <- mtcars[!(mtcars$id %in% c("Cadillac Fleetwood", "Chrysler Imperial", "Lincoln Continental", "
```

Multivariate Regression Analysis

Following *nested models* will be treated against each over:

```
fit <- lm(mpg ~ am, data = mtcars)
fit.wt <- update(fit, mpg ~ am*wt)
fit.wt.hp <- update(fit, mpg ~ am*wt + am*hp)
anova(fit, fit.wt, fit.wt.hp)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + am:wt
## Model 3: mpg ~ am + wt + hp + am:wt + am:hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     26 512.40
## 2     24 165.49  2    346.92 36.0442 1.142e-07 ***
## 3     22 105.87  2     59.61  6.1937  0.007349 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to p-values (they are less than 0.05), the best choi e is the last and more complicated model. Lets test two additional models against the best one.

```
best.fit <- lm(mpg ~ am*wt + am*hp, data = mtcars)
fit.wt.hp.cyl <- update(best.fit, mpg ~ am*cyl*wt + am*cyl*hp)
anova(best.fit, fit.wt.hp.cyl)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am * wt + am * hp
## Model 2: mpg ~ am + cyl + wt + hp + am:cyl + am:wt + cyl:wt + am:hp +
##     cyl:hp + am:cyl:wt + am:cyl:hp
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     22 105.872
## 2     12  74.133 10     31.74 0.5138 0.8501
```

```
fit.wt.hp.gear <- update(best.fit, mpg ~ am*gear*wt + am*gear*hp)
anova(best.fit, fit.wt.hp.gear)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am * wt + am * hp
## Model 2: mpg ~ am + gear + wt + hp + am:gear + am:wt + gear:wt + am:hp +
##     gear:hp + am:gear:wt + am:gear:hp
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     22 105.872
## 2     16  63.729  6    42.143 1.7634 0.1704
```

Adding of the extra predictors has no sensible impact according to p-values (gerater than 0.05). So the best model remains intact.

Chosen Model Quality Analysis

```r
summary(best.fit)
```

```
##
## Call:
## lm(formula = mpg ~ am * wt + am * hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9793 -1.5338 -0.2435  1.1612  5.1652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.28217    5.94151   4.928 6.27e-05 ***
## am1         15.92584    6.73612   2.364  0.02731 *
## wt          -1.26080    2.06659  -0.610  0.54805
## hp          -0.04515    0.01476  -3.059  0.00575 **
## am1:wt      -6.06647    2.72719  -2.224  0.03669 *
## am1:hp       0.01707    0.02187   0.781  0.44339
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.194 on 22 degrees of freedom
## Multiple R-squared:  0.8757, Adjusted R-squared:  0.8474
## F-statistic: 30.98 on 5 and 22 DF,  p-value: 2.955e-09
```

R-squared is very close to the 1, thus the most variance is explained by this model. Coefficients could be interpreted as follows:

1. Manual trans. intercept is greater than automatic trans. intercept by:

```
## [1] "15.926 MPG"
```

2. Change in MPG per one thousand lb automatic vs manual:

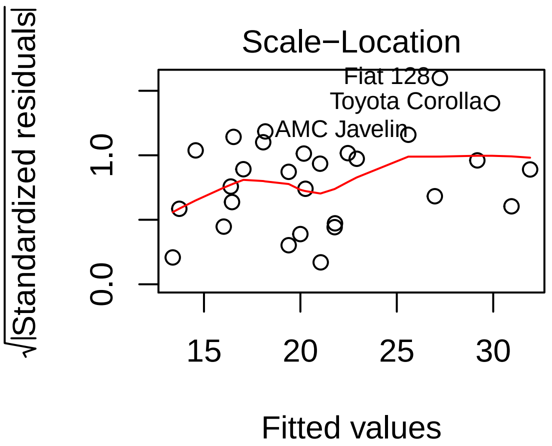
```
## [1] "-1.261 vs -7.327 MPG per 1000*lb"
```

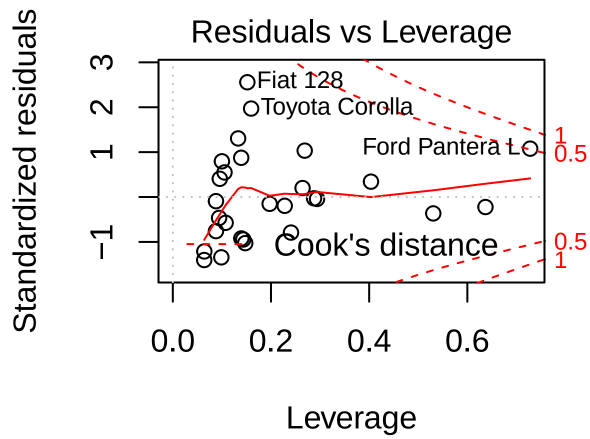
3. Change in MPG per one horsepower automatic vs manual:

```
## [1] "-0.045 vs -0.028 MPG per horsepower"
```

4. Also according to the p-values of the *wt* and *am1:hp* coefficients. Null hypothesis about equality to zero could *not* be rejected.

Residuals



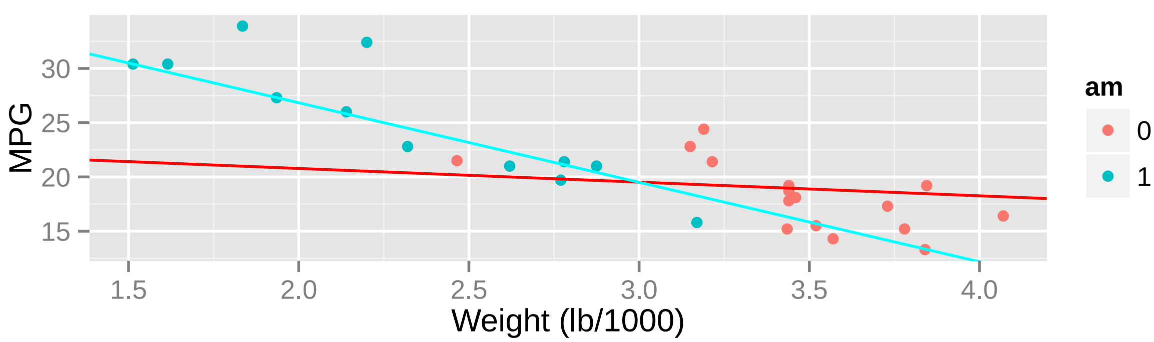


It seems that residuals do not form a pattern. So the conclusions could be made.

Conclusions

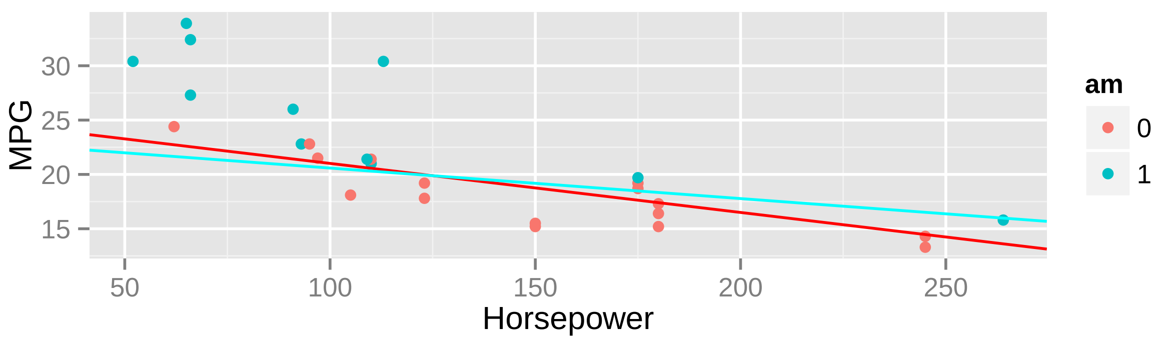
Greater MPG is better.

```r
ggplot(mtcars, aes(wt, mpg, color = am)) + geom_point(size = 2) +
    geom_abline(intercept = coeff[[1]] + coeff[["hp"]] * mean(mtcars$hp),
                slope = coeff[["wt"]], color = "red") +
    geom_abline(intercept = coeff[[1]] + coeff[["am1"]] +
                    (coeff[["hp"]] + coeff[["am1:hp"]]) * mean(mtcars$hp),
                slope=coeff[["wt"]] + coeff[["am1:wt"]], color="cyan") +
    labs(x = "Weight (lb/1000)", y = "MPG")
```



According to hypothetical relationship manual transmission is the best choice for lightwight cars, but automatic transmission is better for heavy cars. Thus nothing particular could be said. Also attention should be paid to the real points that form two clusters. Lightwieght cars with manual transmission and heavy cars with automatic transmission.

```r
ggplot(mtcars, aes(hp, mpg, color = am)) + geom_point(size = 2) +
    geom_abline(intercept = coeff[[1]] + coeff[["wt"]] * mean(mtcars$wt),
                slope = coeff[["hp"]], color = "red") +
    geom_abline(intercept = coeff[[1]] + coeff[["am1"]] +
                    (coeff[["wt"]] + coeff[["am1:wt"]]) * mean(mtcars$wt),
                slope=coeff[["hp"]] + coeff[["am1:hp"]], color="cyan") +
    labs(x = "Horsepower", y = "MPG")
```



Hypothetical relationships between MPG and horsepower in case of automatic transmission is very close to the case of manual transmission.

So it's very hard to quantify the difference between automatic and manual transmission. At average they are very close to each over according to this model.

Thus t-test could be useful to quantify the difference at average.

```r
test <- t.test(mpg ~ am, data = mtcars)
test
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.8331, df = 16.046, p-value = 0.00146
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.918977  -3.143523
## sample estimates:
## mean in group 0 mean in group 1
##        18.14375        25.17500
```

P-value is less than 0.05, so it's statistically significant difference between automatic and manual transmissions. Manual trasmission is better than autimatic. And change in MPG from manual to automatic could be estimated by the following confidence interval:

```
## [1] -10.918977  -3.143523
## attr(,"conf.level")
## [1] 0.95
```

centered at:

```
## [1] "-7.03125 Miles/(US) gallon"
```