Exploring the relationship between type of transmission and miles per gallon

Elena Tikalenko

Summary

The purpose of this dociment is to show that there is relationship between type of transsmission and miles per gallon, namely after increasing of weigth of car on one unit, for cars with manual type of transsmission miles per gallon will decrease much faster than for cars with automatic type of transmission. Also it was made an attempt to estimate the quantify the MPG difference between automatic and manual transmissions.

Description of the data

The data for investigation was extracted from "mtcars" dataset with 32 observation. The variables are miles per gallon (mpg), number of cylinders (cyl), displacement (disp), gross horsepower (hp), rear axle ratio (drat), weight in lb/1000 (wt), typical quarter mile times (qsec), whether the car has a V engine or a straight engine (vs), type of transmission (am, 0 = automatic, 1 = manual), number of forward gears (gear) and number of carburetors (carb).

If we calculate the correlation matrix *cor(mtcars)* then it becomes evident that some pairs of variables are highly correlated with each other (correlation is in interval 0.8 - 0.9): *cyl* and *disp*, *cyl* and *hp*, *disp* and *wt*. So we can throw away variables *disp* and *cyl*. On the pairs plot for the remaining variables (see Appendix, "Pairs plot") we can see the relationships between, for example, *mpg* and *disp*, *mpg* and *hp*, etc.. Relationship between *mpg* and *am* is unclear.

Model Selection

Let's create a series of models adding variables one by one. And do the ANOVA test for these models.

```r
model1 <- lm(mpg ~ am, data = cars)
model2 <- lm(mpg ~ am + wt, data = cars)
model3 <- lm(mpg ~ am + wt + hp, data = cars)
model4 <- lm(mpg ~ am + wt + hp + drat, data = cars)
model5 <- lm(mpg ~ am + wt + hp + drat + qsec, data = cars)
model6 <- lm(mpg ~ am + wt + hp + drat + qsec + vs, data = cars)
model7 <- lm(mpg ~ am + wt + hp + drat + qsec + vs + gear, data = cars)
model8 <- lm(mpg ~ am + wt + hp + drat + qsec + vs + gear + carb, data = cars)
anova(model1, model2, model3, model4, model5, model6, model7, model8)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + hp
## Model 4: mpg ~ am + wt + hp + drat
## Model 5: mpg ~ am + wt + hp + drat + qsec
## Model 6: mpg ~ am + wt + hp + drat + qsec + vs
## Model 7: mpg ~ am + wt + hp + drat + qsec + vs + gear
## Model 8: mpg ~ am + wt + hp + drat + qsec + vs + gear + carb
```

```
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1     30 720.90                                    
## 2     29 278.32  1    442.58 67.2009 2.815e-08 ***
## 3     28 180.29  1     98.03 14.8847 0.0007998 ***
## 4     27 176.96  1      3.33  0.5050 0.4844356    
## 5     26 158.64  1     18.33  2.7827 0.1088487    
## 6     25 158.56  1      0.08  0.0121 0.9132542    
## 7     24 158.56  1      0.00  0.0005 0.9818403    
## 8     23 151.48  1      7.08  1.0750 0.3105914    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results of anova function we can see that including of *wt* and *hp* variables appears to be necessary. Other variables are not significant and can be excluded. Let's compare adjusted R-squared of *model1*, *model2* and *model3*.

```
##      [,1]                [,2]                [,3]               
## [1,] "model1"            "model2"            "model3"           
## [2,] "0.338458908206314" "0.735788906182185" "0.822735694896529"
```

It's clear that *model1* is the worst model of the three as adjusted R-squared is very small. In another two models variable *am* is not significant.

Let's check *model3* by fitting an analysis of variance.

```
summary(aov(mpg ~ am*wt*hp, data = cars))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## am           1  405.2   405.2  83.064 2.91e-09 ***
## wt           1  442.6   442.6  90.737 1.26e-09 ***
## hp           1   98.0    98.0  20.098 0.000155 ***
## am:wt        1   33.4    33.4   6.857 0.015057 *  
## am:hp        1   10.9    10.9   2.244 0.147179    
## wt:hp        1   16.6    16.6   3.400 0.077596 .  
## am:wt:hp     1    2.3     2.3   0.463 0.502939    
## Residuals   24  117.1     4.9                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results we can see that one more variable with dummy-variable can be added in our model - am*wt. Let's fit a new model and check the coefficients.

```
model_final <- lm(mpg ~ am + wt + hp + am:wt, data = cars)
summary(model_final)
```

```
## 
## Call:
## lm(formula = mpg ~ am + wt + hp + am:wt, data = cars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.0639 -1.3315 -0.9347  1.2180  5.0822
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.947333   2.723411  11.363 8.55e-12 ***
## am          11.554813   4.023277   2.872  0.00784 **
## wt          -2.515586   0.844497  -2.979  0.00605 **
## hp          -0.026949   0.009796  -2.751  0.01048 *
## am:wt       -3.577910   1.442796  -2.480  0.01968 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.332 on 27 degrees of freedom
## Multiple R-squared:  0.8696, Adjusted R-squared:  0.8503
## F-statistic: 45.01 on 4 and 27 DF,  p-value: 1.451e-11
```

Adjusted R-squared for *model_final* is greater than for *model3*.

```r
anova(model3,model_final)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + wt + hp
## Model 2: mpg ~ am + wt + hp + am:wt
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     28 180.29
## 2     27 146.84  1    33.446 6.1496 0.01968 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And the results of anova-function shows that it seems that variable *am:wt* necessary in model. So, if we put *am* in *model_final*, then we'll get the next model:

- for case, if we have a car with automatic transmission then our model is the next: mpg = 30.95 - 2.52wt - 0.03hp
- for case of car with manual transmission: mpg = 42.50 - 6.10wt - 0.03hp

To quantify the uncertainty in coefficients let's check the plot of coefficients estimates with confidence interval (see Appendix, "Coefficients estimates with confidence interval").

It's clear that hypothesys about that coefficients are equal to zero is rejected as confidence intervals don't contain zero.

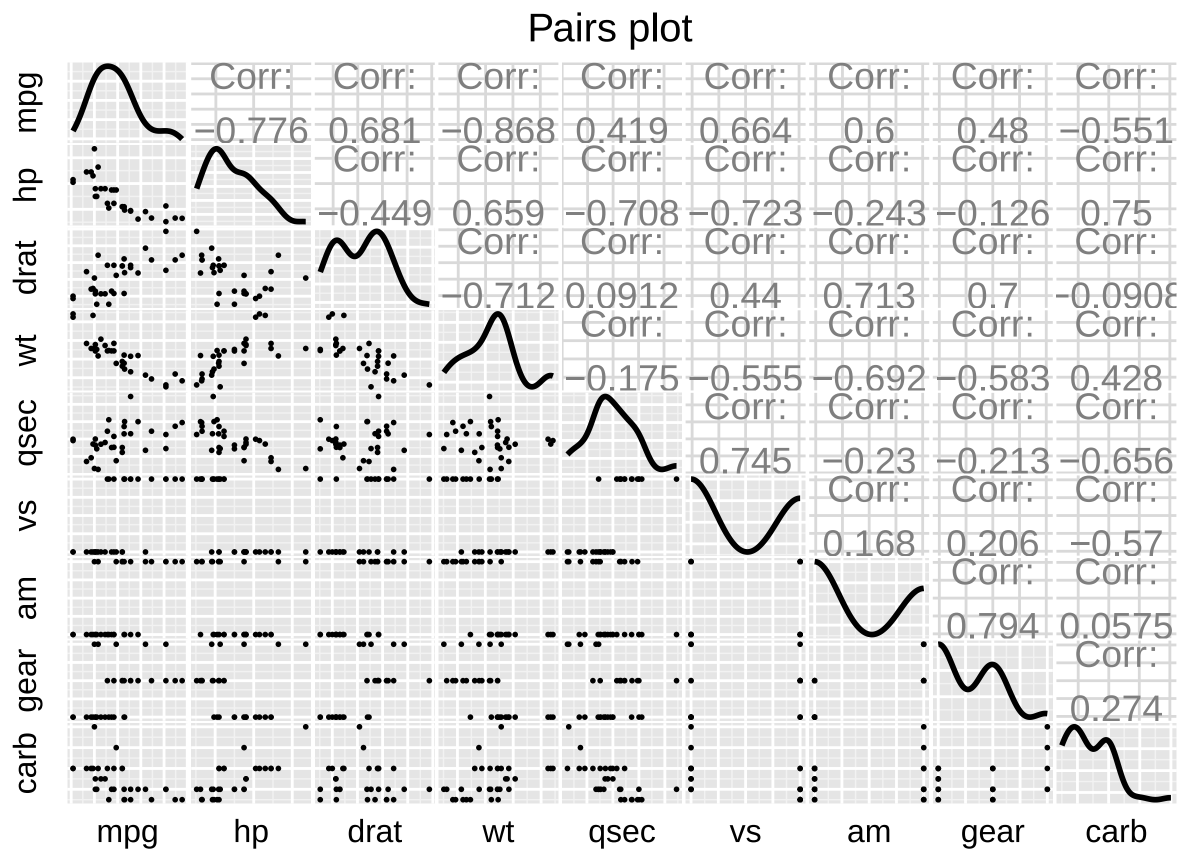
So our model means that after increasing of weigth of car on one unit for cars with manual type of transsmission miles per gallon will decrease on approximatly 6 units, while for cars with automatic type of transmission on 2,5 units only. This dependence is also visible on the plot (see Appendix, "Miles from weigth for automatical and manual types of transmission").

Residual plot

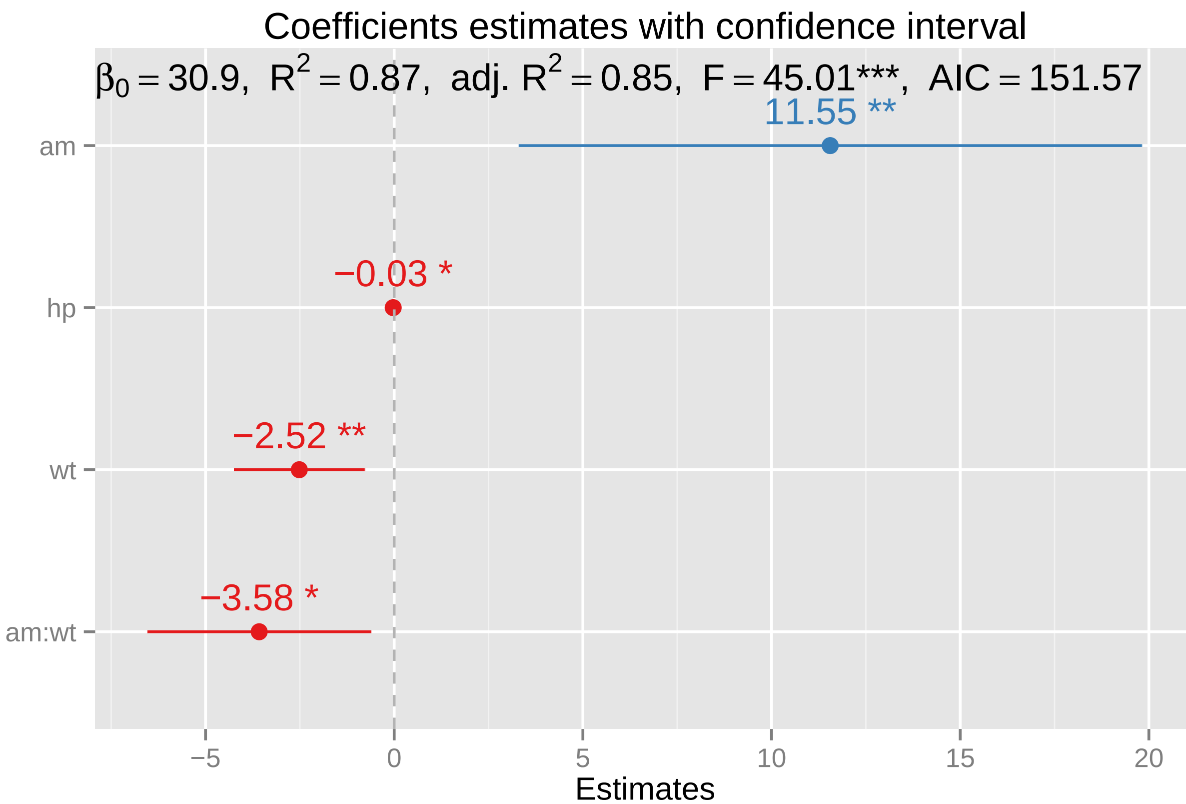
On the residual plot (see Appendix, "Residual plot") residuals are pretty symmetrically distributed and there aren't clear patterns in general. No abnormalities are observed in the residual plot.

Appendix

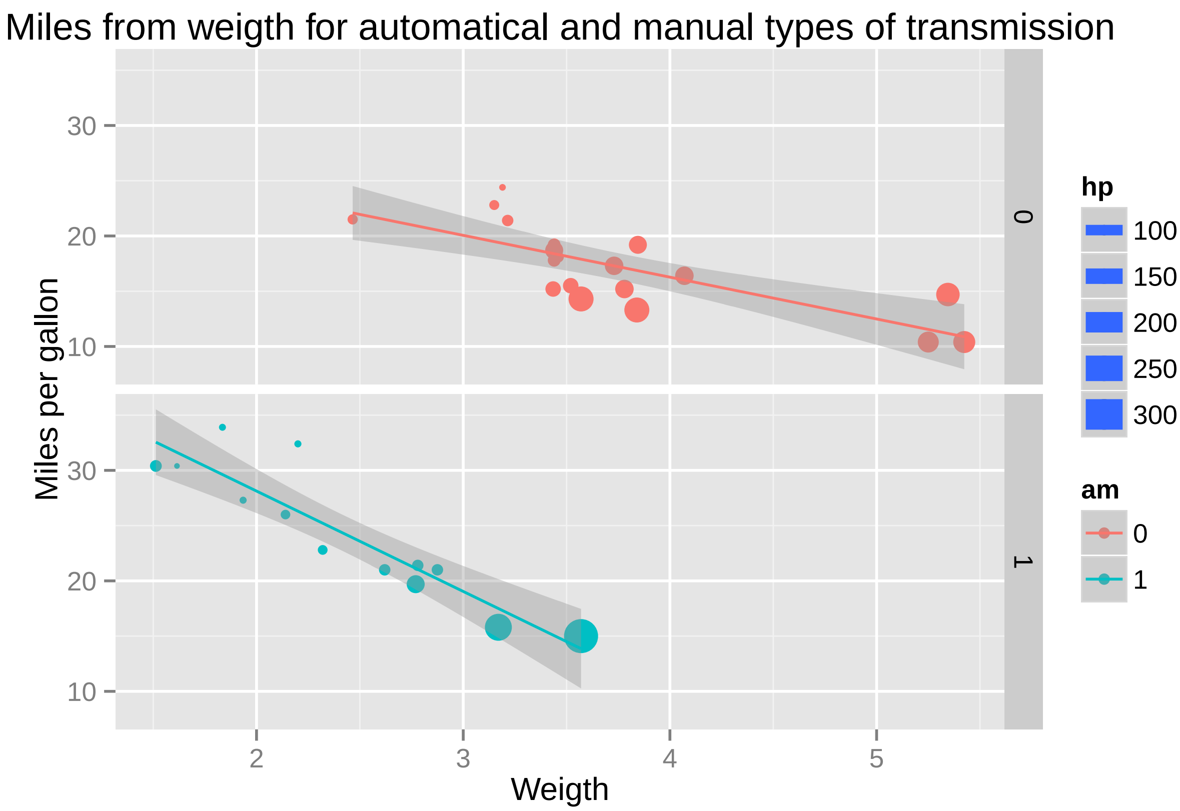
```
ggpairs(cars, columns = c(1, 4:11), title = "Pairs plot", params=c(size=1),
        upper=list(params=list(size=5)))+
        theme(axis.line=element_blank(),
              axis.text=element_blank(),axis.ticks=element_blank())
```



```r
sjp.lm(model_final, showModelSummary = TRUE,
             title = "Coefficients estimates with confidence interval")
```



```r
cars$am <- as.factor(cars$am)
ggplot(cars, aes(x = wt, y = mpg, color = am, size = hp)) + geom_point() +
        facet_grid(am ~.) + stat_smooth(method = "lm") +
        ggtitle("Miles from weigth for automatical and manual types of transmission") +
        labs(x = "Weigth", y = "Miles per gallon")
```



```
ggplot(model_final, aes(x=fitted(model_final), y=resid(model_final))) +
        geom_point() + ggtitle("Residual plot") +
        labs(x = "Fitted values", y = "Residuals") + geom_hline(aes(yintercept = 0)) +
        theme(legend.position="none")
```