Analysis of transmission type differences for cars of 1973.

Executive summary.

The fulfilled analysis revealed an insufficiency of data in the used dataset to perform causal analysis and discover if changing just transmission type causes the significant changes of miles/gallon for automobiles of 1973. But results of the analysis also allow to believe that if exactly different transmission types cause changes in miles/gallon (instead of some 3rd force causes changes in both of them) then manual transmission type has more miles/gallon than automatic one does.

Analysis.

During exploratory analysis I plot 2 graphics and looked at the first rows of data.

```
##    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## 1 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## 2 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## 3 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## 4 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
```

From the boxplot (look at appendix) I can assume that manual transmission type must have bigger amount of miles/gallon, but there are many other variables here, so may be there is some variable which causes both changes in transmission type and miles/gallon. From the pairs graphic it is clear that transmission type ("am" on the plot) has strong correlation with many variables which also says for high probability of existence another force which causes changes in both transmission type and miles/gallon.

Because of the little quantity of observations only simple models (e.g. with 1-2 terms without intersection) have low p-values for all coefficients, which means that there coefficients are truly not zeros. I am interested only in models with am term (transmission type) included. Below there is the list of maximum of p-values among all coefficients for each model with 2 terms: am and the one specified in column name.

```
##    cyl   disp     hp   drat     wt   qsec     vs   gear   carb
## 0.0585 0.2118 0.0000 0.7848 0.9879 0.0077 0.0000 0.9894 0.3360
```

So, only the models with am and one term of qsec (1/4 mile time), hp (Gross horsepower), vs (V/S) have all coefficients which are significantly different from zeros (p-value < 0.05). All other models are uninterpretable because some of their coefficients probably equal to zero (We cannot reject this hypothesis).
These 3 models and the model with only 1 term (am) I compared by ANOVA to find the one with the least sum of residual's squares (RSS column):

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + qsec
## Model 3: mpg ~ am + hp
## Model 4: mpg ~ am + factor(vs)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 352.63  1    368.26 30.285 6.271e-06 ***
## 3     29 245.44  0    107.19
```

```
## 4     29 353.49  0   -108.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, the best model is the one with am and hp terms.

Must to say that for all those 4 models the coefficient for am1 term is positive. This coefficient defines differences between manual (labeled as 1) and automatic (labeled as 0) transmission types given all other predictors (if they present) stay constant. So I can assume that if changes in transmission type can cause changes in miles/gallon values then for manual transmission type the value of miles/gallon is bigger than for automatic type. Below there are actual values of coefficients for each of those models with 95% confidence intervals:

```
##                 Formula  CoefVal lowerConfInt upperConfInt
## 1              mpg ~ am 7.244939     3.641510    10.848369
## 2       mpg ~ am + qsec 8.876331     6.238672    11.513990
## 3         mpg ~ am + hp 5.277085     3.069177     7.484994
## 4 mpg ~ am + factor(vs) 6.066667     3.459322     8.674012
```

So, with the best model am + hp we can expect increasing in miles/gallon on 5.2770853 when we replace automatic transmission by manual one given the gross horsepower value is fixed. The uncertainty of this estimation for 95% confidence interval is (3.0691769; 7.4849937).

To perform some diagnostics I have placed residual plot and fitted value plot for the model mpg ~ am + hp in the appendix and below there are influence measures for mpg ~ am and mpg ~ am + hp models:

```
## Potentially influential observations of
##   lm(formula = mpg ~ am, data = mtcars) :
## NONE
```

```
## numeric(0)
```

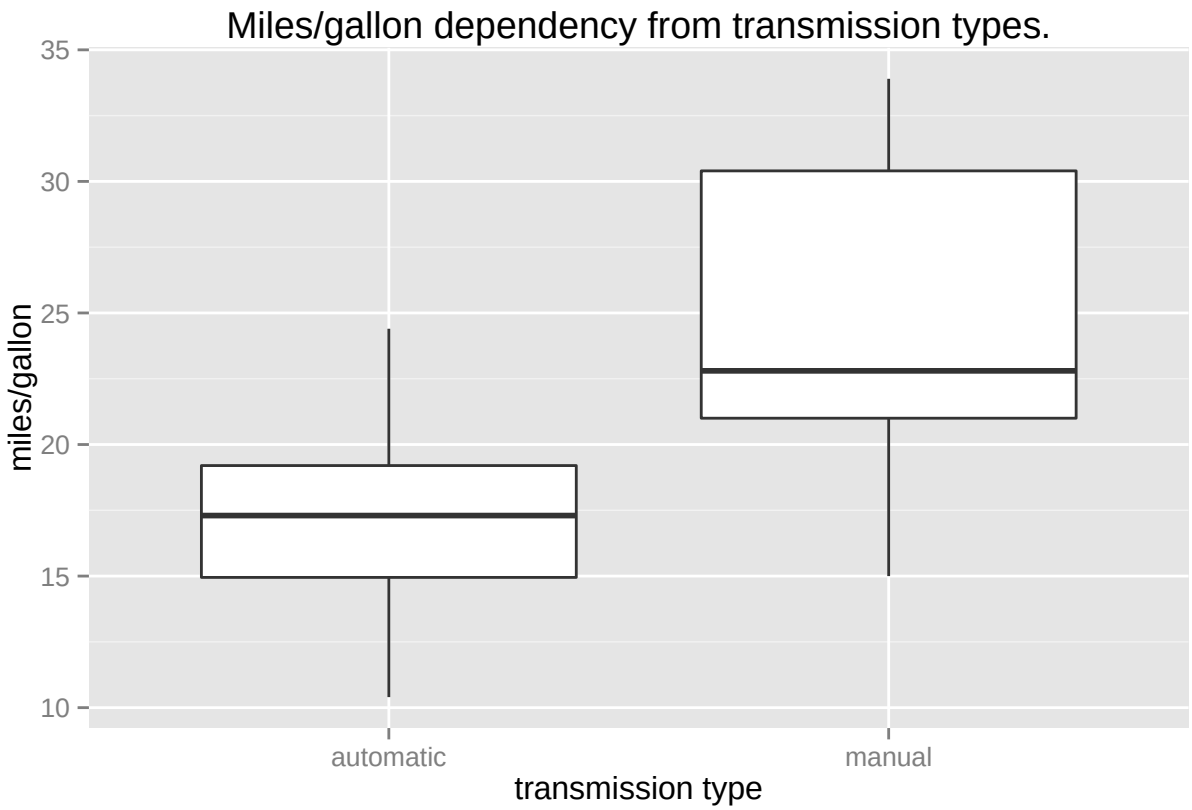
```
## Potentially influential observations of
##   lm(formula = mpg ~ am + hp, data = mtcars) :
##
##    dfb.1_ dfb.am1 dfb.hp dffit   cov.r   cook.d hat
## 29  0.07  -0.07   -0.08  -0.10    1.41_*  0.00   0.21
## 31 -0.81   0.56    0.92   1.03_*  1.54_*  0.34   0.39_*
```

I didn't exclude influential observations because there is too little data to reduce it further.

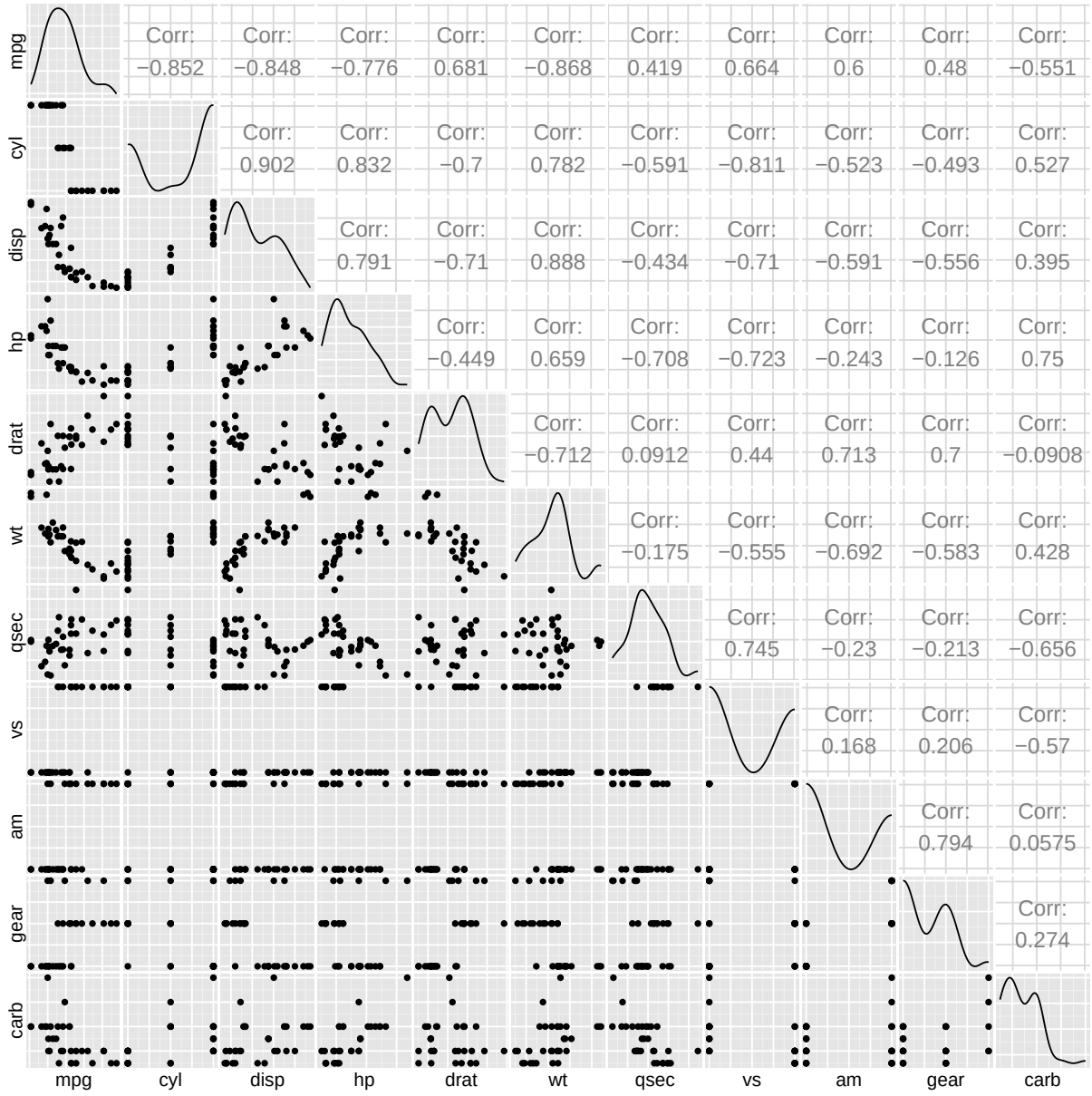
In the end I must to say that these models are not enough to conclude that changes in transmission type can cause change in miles/gallon. To proof such statement I need at least to fit the model with all possible terms, get significant p-values for all coefficients, look at coefficient for am1 (difference between manual transmission type and automatic one given all other predictors are constant) and only then claim how exaclty transmission type affect the miles/gallon. But in the given dataset there is not enough data to get significant p-value for am1 coefficient (I cannot reject hypothesis that it is not equal to zero), but may be more data will not change this and then it means that transmission type doesn't affect miles/gallon at all.

Appendix

Exploratory analysis



So called, pairs graphic with correlation values for all variables of dataset.



Model analysis.