Regression Models Course Project. Author: Sergii Sorokolat. Date: 7/1/2018

Executive summary

This analysis explores the relationship between a set of variables in `mtcars` dataset and addresses the following questions:

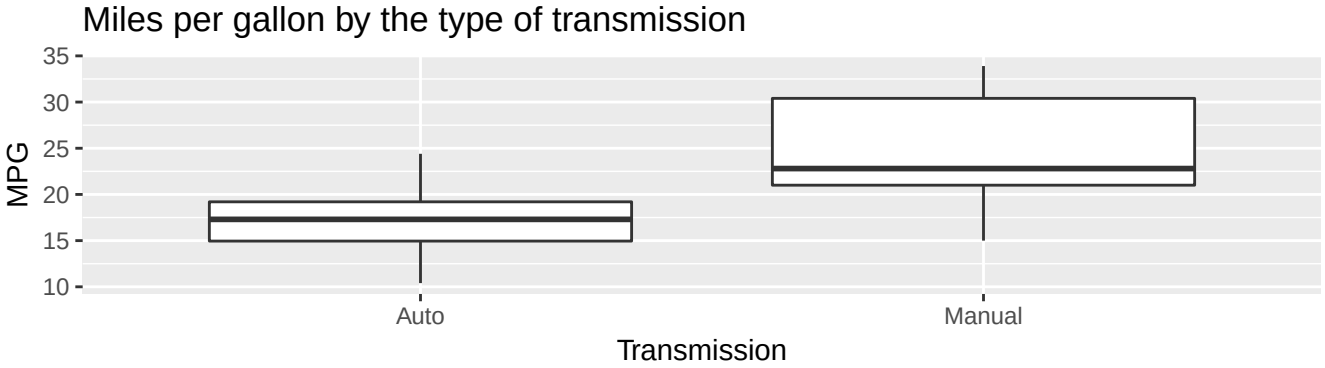
- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions.

The analysis will show, that even though there is difference between the cars with manual and automatic transmission with respect to MPG, the manual transmission is no better than automatic transmission judging by information collected in the dataset.

Exploratory analysis

The data `mtcars` is available in R `datasets` packages. The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). Format: A data frame with 32 observations on 11 variables. Features description: mpg - Miles/(US) gallon, cyl - Number of cylinders, disp - Displacement (cu.in.), hp - Gross horsepower, drat - Rear axle ratio, wt - Weight (1000 lbs), qsec - 1/4 mile time, vs - Engine type V/Straight, am - Transmission (0 = automatic, 1 = manual), gear - Number of forward gears, carb - Number of carburetors.

```
## # A tibble: 6 x 11
##     mpg cyl    disp    hp  drat    wt  qsec vs           am    gear  carb
##   <dbl> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <fct>        <fct> <fct> <fct>
## 1  21   6       160   110  3.9   2.62  16.5 V/Vee        Manu~ 4     4
## 2  21   6       160   110  3.9   2.88  17.0 V/Vee        Manu~ 4     4
## 3  22.8 4       108    93  3.85  2.32  18.6 Straight/in~ Manu~ 4     1
## 4  21.4 6       258   110  3.08  3.22  19.4 Straight/in~ Auto  3     1
## 5  18.7 8       360   175  3.15  3.44  17.0 V/Vee        Auto  3     2
## 6  18.1 6       225   105  2.76  3.46  20.2 Straight/in~ Auto  3     1
```



The boxplot shows that MPG is higher for the manual transmission. Let's check this with a t-test.

```
##    estimate estimate1 estimate2 statistic     p.value parameter  conf.low
## 1 -7.244939  17.14737  24.39231 -3.767123 0.001373638  18.33225 -11.28019
##   conf.high                 method alternative
## 1 -3.209684 Welch Two Sample t-test   two.sided
```

The t-test confirms our assumption (p-value of 0.001374 <= 0.05, therefore we reject H0) and shows that there's statistically significant difference between automatic and manual transmission. The corellation matrix shows that weight is the most strongly correlated with MPG (-0.8677). Let's take this into account during the model selection process. *The corellation matrix is available in Appendix A.*

Modeling

The naive approcah would be to fit a linear model with am as a predictor and mpg as an outcome. Let's try this first.

```r
fit_1 <- lm(mpg ~ am, data = dataset)
summary(fit_1)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## amManual     7.244939   1.764422  4.106127 2.850207e-04
```

```r
summary(fit_1)$r.squared
```

```
## [1] 0.3597989
```

The single variable model explains only 36% of the variance. The Stepvise Algorithm will be used to find the best fitting model automatically.

```r
stepvise_model <- step(lm(mpg ~ . , data = dataset ), trace = 0)
summary(stepvise_model)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## amManual     1.80921138 1.39630450  1.295714 2.064597e-01
```

```r
summary(stepvise_model)$r.squared
```

```
## [1] 0.8658799
```

The algorithm suggests that this model explains 87% of the variance. **am, cyl, wt, hp** variables are used as predictors for mpg. Let's do a quick check and compare our first model and the one we just discovered.

```r
tidy(anova(fit_1, stepvise_model))
```

```
##   res.df      rss df   sumsq statistic      p.value
## 1     30 720.8966 NA      NA        NA           NA
## 2     26 151.0256  4 569.871  24.52671 1.688435e-08
```

F statistic is large, p-value is small, so we can confirm that there's a significant difference between the models and the one that the stepwise algorithm suggested is in fact better.

Residuals / diagnostics plots

Residuals plots are available in Appendix A. From the plots of residuals, we can see that there is no pattern in residuals and they are homoscedastic. All the outliers do not have significant influence on the model.

```r
head(sort(dfbetas(stepvise_model)[,'amManual'], decreasing = TRUE), n = 2)
```

```
##        21        18
## 0.7305402 0.4292043
```

```r
head(sort(hatvalues(stepvise_model), decreasing = TRUE), n = 2)
```

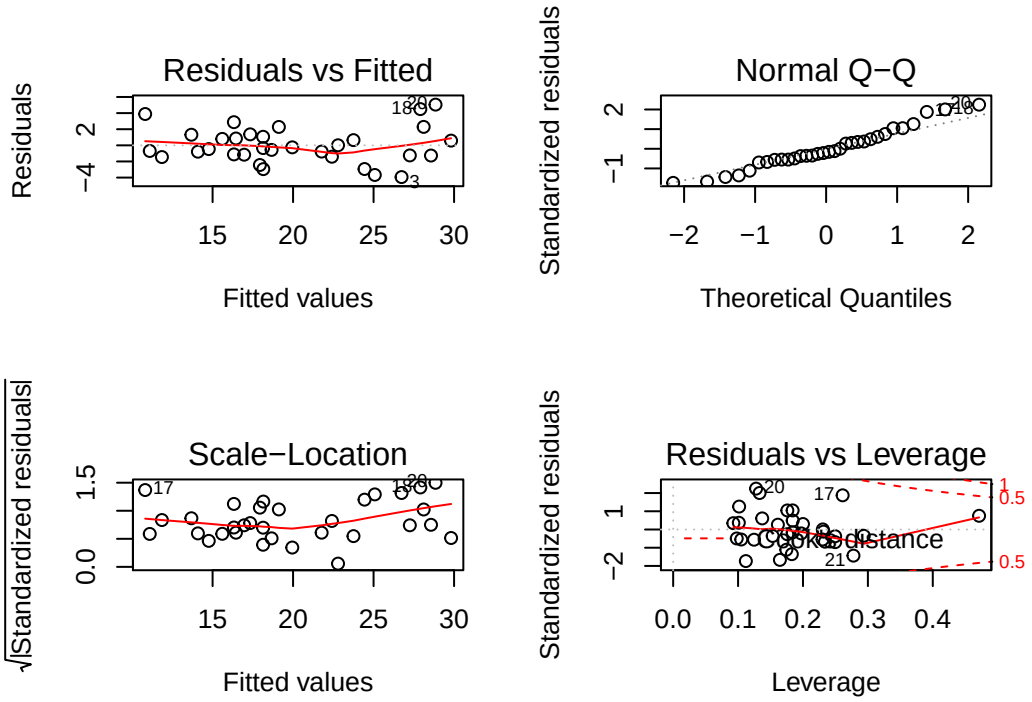
```
##        31        16
## 0.4713671 0.2936819
```

Conclusions

From the model output, the manual transmission results in +1.8 mpg compared to the automatic transmission (**however p value of 0.206 suggests that the result is not statistically significant**). Every 1000 lbs of the weight of the car result in decrease of mpg by -2.5

It's not clear, however, whether it's an effect of the type of transmission itself, or cars in the dataset just tend to be lighter and have less cylinders. More thorough analysis is required.

Appendix A



Corellation matrix

	mpg	disp	hp	drat	wt	qsec
mpg	1.0000000	-0.8475514	-0.7761684	0.6811719	-0.8676594	0.4186840
disp	-0.8475514	1.0000000	0.7909486	-0.7102139	0.8879799	-0.4336979
hp	-0.7761684	0.7909486	1.0000000	-0.4487591	0.6587479	-0.7082234
drat	0.6811719	-0.7102139	-0.4487591	1.0000000	-0.7124406	0.0912048
wt	-0.8676594	0.8879799	0.6587479	-0.7124406	1.0000000	-0.1747159
qsec	0.4186840	-0.4336979	-0.7082234	0.0912048	-0.1747159	1.0000000