Regression Analysis - mtcars

Executive Summary

Looking at a data set of a collection of cars, we are interested in investigating:

- whether there is a difference in mileage (MPG) between automatic or manual transmissions, and
- quantify the difference in MPG between automatic and manual transmissions.

We investigate the mileage by analyzing both single variable- and multi variable regression models on the transmissions, using data from `mtcars`, as well as making an ANOVA table and residual plots. We can conclude that there is a signficant difference between transmission and that the MPG is 2.94 higher for cars with manual transmission compared to those with automatic, and that there are other variables that influence the MPG.

Processing data

Load in the mtcars data:

```
data(mtcars)
```

Exploratory data

Single variable regression

We are interested in finding out whether there is a significant difference between automatic and manual transmissions on the mileage. We check first with a single variable whether this might imply a significant difference:

```
fit <- lm(mpg ~ as.factor(am), mtcars)
summary(fit)$coef
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    17.147368   1.124603 15.247492 1.133983e-15
## as.factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

As the p-value is very low (<0.05) it seems to imply that there is a significant difference between automatic and manual transmissions when looking to a single variable. However, as R-squared is only 33.85% of the total variability by the linear relationship with the predictor, it also implies that the model is of poor fit.

Multi variable regression

In order to find a better fitting model a linear model is fit against all the variables in `mtcars`, see appendix.

The R-squared has improved (80.66%), however, none of the variables are significant (>0.05). This suggest there are covariate correlations between the variables (that the variables are nested).

Let's investigate this by checking the correlation for `mtcars`. The higher the absolute covariate value is between the two variables, the stronger the correlation between them. This means that we can exclude the

other variables, if we include a variable that might imply to be significant. See appendix for plot of the correlation. Since we didn't have a variable that was significant (<0.05) in the model we pick the smallest one to begin with, which is `wt`:

```
cor(mtcars)[6,]
```

```
##        mpg        cyl       disp         hp       drat         wt
## -0.8676594  0.7824958  0.8879799  0.6587479 -0.7124406  1.0000000
##       qsec         vs         am       gear       carb
## -0.1747159 -0.5549157 -0.6924953 -0.5832870  0.4276059
```

When comparing the correlaion of `wt` with the other variables, we exclude all variables which have a correlation > 0.5 (excluding `am`):

We see that we are left with 4 variables: `wt`, `am`, `qsec` and `carb`. However, since there is a strong correlation between `qsec` and `carb` ($0.6562 > 0.5$), we exclude `carb` as well.

As automated covariate selection is a difficult topic, which depends heavily on how rich of a covariate space one wants to explore, this also affects the selection of variables to use in the regression model.

We do a new multi-variable regression model with the three variables:

```
fit_new <- lm(mpg ~ as.factor(am) + wt + qsec, mtcars)
summary(fit_new)$coef
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     9.617781  6.9595930  1.381946 1.779152e-01
## as.factor(am)1  2.935837  1.4109045  2.080819 4.671551e-02
## wt             -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec            1.225886  0.2886696  4.246676 2.161737e-04
```

We can check with an ANOVA table if the variables should be added with the single regression model, see appendix for the table.

According to the p-value in the ANOVA table we should include `wt` and `qsec` to the model.

Before concluding, we also check the residuals in order to be sure that the model is normally distributed, as well as homoskedastic, see appendix for plots.

After checking the plots, we can confirm that the residuals are normally distributed and are homoskedastic.

Conclusion

Based on the outcome there is a significant difference in the mileage between automatic and manual transmissions, however, there are other variables that are also influencing the mileage (such as weight and acceleration). On average, manual transmissions are 2.94 MPG better than automatic transmissions.

Appendix

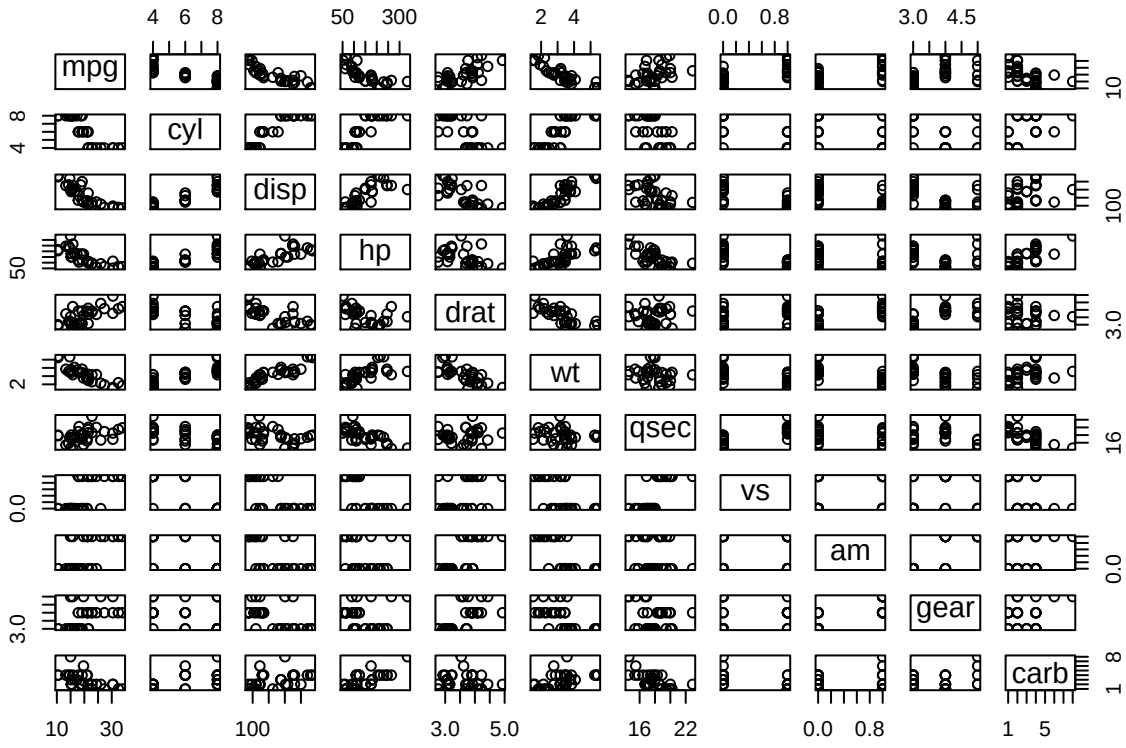
Multi variable regression - all variables:

```
fit_all <- lm(mpg ~ ., mtcars)
summary(fit_all)$coef
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

Pairwise scatter plot:

```
pairs(mtcars)
```



ANOVA table:

```
anova(fit, fit_new)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ as.factor(am)
## Model 2: mpg ~ as.factor(am) + wt + qsec
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Plot of the multi-variable regression model:

```
par(mfrow = c(2,2))
plot(fit_new)
```