CourseProject_RegressionModels

Instructions

You 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"

Analysis

```
data(mtcars)
#summary(mtcars)
```

Let us first plot mpg (Miles/gallon) versus am (Transmission(0 = automatic, 1 = manual)). This exploratory plot is shown as the first plot in the appendix. The boxplot shows that the average mpg for cars with manual transmission (~24) is significantly higher than for cars with automatic transmission (~17).

To get the exact numbers I first do a simple linear regression.

```
simple_model <- lm(mpg ~ am, data = mtcars)
summary(simple_model)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

As mentioned before the automatic cars have an average mpg of 17.7 while the manual cars have 24.4. The slope coefficient is significant (p-value < 0.01) from which I conclude that the mpg is indeed different. But the R-squared value is rather smal, which means that the regression line explains only a small part of the variance. Before perform a multiple regression we first look at the correlations between mpg and the other variables.

```
cor(mtcars)[1,]
```

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

The absolute correlations are highest for cyl, disp, hp and wt. Let us use these variables and add “am” to the model

```r
multiple_regr <- lm(mpg ~ am + cyl + disp + hp + wt, data = mtcars)
summary(multiple_regr)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## am           1.55649    1.44054   1.080  0.28984
## cyl         -1.10638    0.67636  -1.636  0.11393
## disp         0.01226    0.01171   1.047  0.30472
## hp          -0.02796    0.01392  -2.008  0.05510 .
## wt          -3.30262    1.13364  -2.913  0.00726 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```

This model explains 85% of the variance, is therefore significant better than the first model. However, there are strong correlations between some of the variables, for instance, between cylinders and horse power, which may blow up the variance of the model.

```r
cor(mtcars$cyl, mtcars$hp)
```

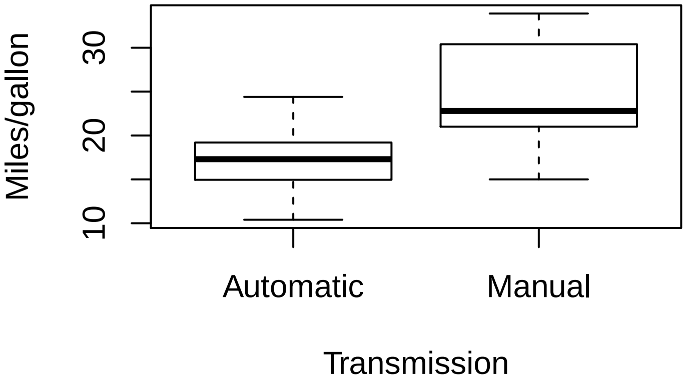
```
## [1] 0.8324475
```

The second plot of the appendix shows the residuals of this model to confirm that the residuals are normally distributed around zero.

In conclusion, the model that explains 85% of the variance. It predicts an average increase in MPG by 1.5 between automatic and manual cars. This is however not significant, and therefore one can not say whether manual or automatic cars are better.

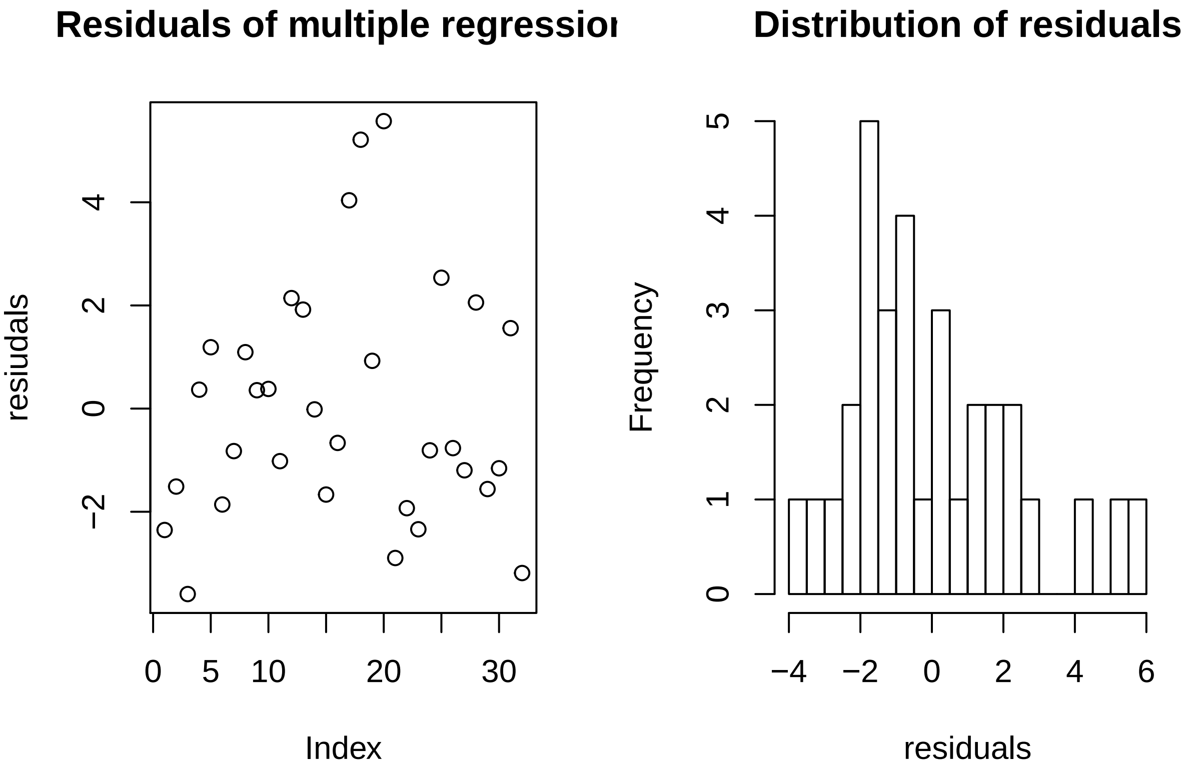
Appendix

```r
boxplot(mpg ~ am, data = mtcars, names =
            c("Automatic", "Manual"), xlab = "Transmission", ylab = "Miles/gallon")
```



```
par(mfrow = c(1,2))

plot(multiple_regr$residuals, ylab = "resiudals", main = "Residuals of multiple regression")
hist(multiple_regr$residuals, breaks = 20, xlab = "residuals", main = "Distribution of residuals")
```


Residuals of multiple regressio
resiudals
Index
Distribution of residuals
Frequency
residuals