Three Regression Models to Assess the Impact of Different Variables on mpg-values in the mtcars-Dataset

Peer-graded Assignment: Regression Models Course Project

Luis David Bedon Gomez

12/9/2017

Executive Summary

This report presents three different regression models to quantify influencing factors in the *mpg*-values of the mtcars-dataset. The first one examines the *mpg*-Values as a function of the transmission type *am* as a unique predictor, in which cars with manual transmission show on average a 7 mpg better performance. However, this model can only explain about 34% of the variance of the *mpg*-Variable. The second model incorporates the car's weight *wt* as a predictor. Now weight has a stronger negative influence on *mpg* than *am*. Though, ANOVA results shows significance for both variables and an increase in the $adj.R^2$ up to 74% is achieved. In a third model, the interaction term between *am* and *wt* is considered, as well as main effects and interaction between *qsec* and *hp*. About 89% of the *mpg*'s variance can be explained with this final model, that shows a consistent random distribution of residuals.

1. Exploratory Data Analysis

We begin by taking a look at the documentation for the mtcars-dataset, which contains data taken from the Motor Trend US magazine comparing 32 different automobiles (1973–74 models), published in 1974.

The mtcars-dataset is a tidy dataset with 11 variables and 32 observations per variable, one for each of the cars. The variables comprise performance and technical configuration data. The performed exploration can be found in section A1 of the Appendix.

2.1 Model 1: Variable "am" as unique predictor

For the first model we take the variable *am* as a unique, categorical predictor.

- Regression Coefficients

The regression coefficients are calculated with the standard *lm*-function.

The variable *am* is a binary categorical variable with $am = 0 \Leftrightarrow$ automatic transmission and $am = 1 \Leftrightarrow$ manual transmission.

We know from the lecture, that in this case the coefficients have the meaning $\beta_0 = \overline{mpg}_{\text{automatic}}$ and $\beta_1 + \beta_0 = \overline{mpg}_{\text{manual}}$ (s. Appendix A2.1) and obtain:

```
##       Estimate Std..Error   t value     Pr(>|t|)
## beta0    17.15       1.12 15.247492 1.133983e-15
## beta1     7.24       1.76  4.106127 2.850207e-04
```

The coefficients are thus statistically significant different from 0, and from the coefficients we can affirm, that cars with manual transmission achieve in general a $\beta_1 = 7.24 mpg$ higher mpg-value than cars with automatic transmission, holding all other variables constant.

- Residuals

However, we get as the adjusted R^2 for this regression model:

```
## [1] "adj. R^2 = 0.338"
```

This means, that the classification in *automatic transmission* and *manual transmission* can only explain about 34% of the variance of *mpg*.

2.2 Model 2: Variables "am" and "wt" as predictors

To perform better than in the past model, now we add the variable *wt* as a continuous predictor. We expect a better result, since the fuel efficiency should be higher in a lighter car than a heavier one.

- Regression Coefficients The *lm*-function gives us:

```
##             Estimate Std..Error     t value     Pr(>|t|)
## (Intercept)    37.32       3.05 12.21799285 5.843477e-13
## factor(am)1    -0.02       1.55 -0.01527855 9.879146e-01
## wt             -5.35       0.79 -6.79080719 1.867415e-07
```

Having now include the *wt* variable, the transmission type looses protagonism, being this a signal, that weight has a considerable importance in the full efficiency of a car. The negative sign confirms the assumption, that a heavier car has a worse fuel efficiency than a lighter one.

- Residuals

Performing an ANOVA-analysis (s. Appendix A2.2), we see that, although the variable *wt* has a major influence on *mpg* than *am*, the F-statistic shows us, that including both variables is statistically significant. As we know from the lecture, this mean that both variables show linear independence from each other.

The $adj.R^2$ for this model is:

```
## [1] "adj. R^2 = 0.7358"
```

The $adj.R^2$ shows a huge improvement of nearly 118% comparing to the only classification of model 2.1. Motivated by this improvement, we will look in a third model, looking for an even better solution.

2.3 Model 3: Variables $am * wt$ and $hp * qsec$ as predictors

We consider now the main effects and interactions of *am* and *wt* by using the $am * wt$ in the *lm*-function, as well as the variables *hp* and *qsec* and their interaction, too. This model is able to explain about 89% of the variance of *mpg*, achieving an increase of nearly 22% in $adj.R^2$ in comparison to the Model 2 above:

```
## [1] "adj. R^2 = 0.88997"
```

- Regression Coefficients and Residuals

In contrast to the model 2.2, the *wt*-coefficient is no longer significantly different from 0. This does however the negative interaction term $wt : factor(am)$, validating the trend on the model above.

The variable *qsec* - the 1/4 mile time - having a positive coefficient suggests that a slower car achieve a better *mpg*-value. Surprisingly, the variable *hp* also shows a positive sign, though being not far for zero, this suggests a slightly better fuel performance for cars with a higher power. The full list of coefficients is:

```
##                Estimate Std..Error    t value     Pr(>|t|)
## (Intercept)       -4.09      10.52 -0.3890436 0.7005379023
## wt                -1.45       1.02 -1.4282594 0.1655920110
## factor(am)1       16.35       3.74  4.3704984 0.0001904780
## hp                 0.16       0.08  2.0392292 0.0521309002
## qsec               1.78       0.53  3.3879881 0.0023356297
## wt:factor(am)1    -5.22       1.36 -3.8304988 0.0007646997
## hp:qsec           -0.01       0.01 -2.0823257 0.0477016205
```

The residuals (s. Appendix A2.3) show a random distribution, pointing that the chosen linear model gives a good representation of the data. Furthermore, the ANOVA shows significance for all variables with exception of *am*, though necessary for the significant interaction term $wt * am$.

```
## Analysis of Variance Table
##
## Response: mpg
##               Df Sum Sq Mean Sq  F value    Pr(>F)
## wt             1 847.73  847.73 212.1137 1.019e-13 ***
## factor(am)     1   0.00    0.00   0.0006  0.981299
## hp             1  98.03   98.03  24.5282 4.214e-05 ***
## qsec           1  20.22   20.22   5.0605  0.033528 *
## wt:factor(am)  1  42.82   42.82  10.7150  0.003103 **
## hp:qsec        1  17.33   17.33   4.3361  0.047702 *
## Residuals     25  99.91    4.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix

A1. Exploratory data analysis

```
data("mtcars") #load the mtcars-data

?mtcars #read documentation
str(mtcars) #structure of data
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
rownames(mtcars) #cars included
```

```
##  [1] "Mazda RX4"           "Mazda RX4 Wag"       "Datsun 710"
##  [4] "Hornet 4 Drive"      "Hornet Sportabout"   "Valiant"
##  [7] "Duster 360"          "Merc 240D"           "Merc 230"
## [10] "Merc 280"            "Merc 280C"           "Merc 450SE"
## [13] "Merc 450SL"          "Merc 450SLC"         "Cadillac Fleetwood"
## [16] "Lincoln Continental" "Chrysler Imperial"   "Fiat 128"
## [19] "Honda Civic"         "Toyota Corolla"      "Toyota Corona"
## [22] "Dodge Challenger"    "AMC Javelin"         "Camaro Z28"
## [25] "Pontiac Firebird"    "Fiat X1-9"           "Porsche 914-2"
## [28] "Lotus Europa"        "Ford Pantera L"      "Ferrari Dino"
## [31] "Maserati Bora"       "Volvo 142E"
```

A2.1 - Model 2.1

```
# Regression one categorical predictor
fit0<-lm(mpg~factor(am),data=mtcars)
sumfit<-summary(fit0)
# Rename the coefficients for better understanding:
row.names(sumfit$coefficients)<-c("beta0","beta1")
tablecoef<-data.frame(round(sumfit$coefficients[,1:2],2))
tablecoef$"t value"<-(sumfit$coefficients[,3])
tablecoef$"Pr(>|t|)"<-(sumfit$coefficients[,4])
```

A2.2 - Model 2.2

```
# Regression one categorical predictor and one continuous variable
fit2_0<-lm(mpg~factor(am)+wt,data=mtcars)
sumfit2_0<-summary(fit2_0)

anova(fit2_0)
```

```
## Analysis of Variance Table
##
## Response: mpg
##             Df Sum Sq Mean Sq F value    Pr(>F)
## factor(am)   1 405.15  405.15  42.215 4.110e-07 ***
## wt           1 442.58  442.58  46.115 1.867e-07 ***
## Residuals   29 278.32    9.60
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

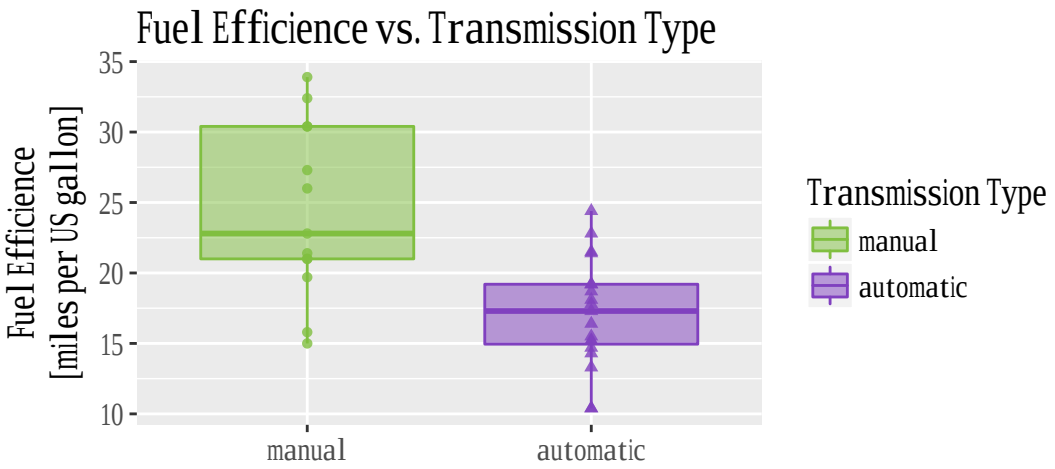

Figure 1: Model 2.1 as a classification of the mpg-values by transmission type.

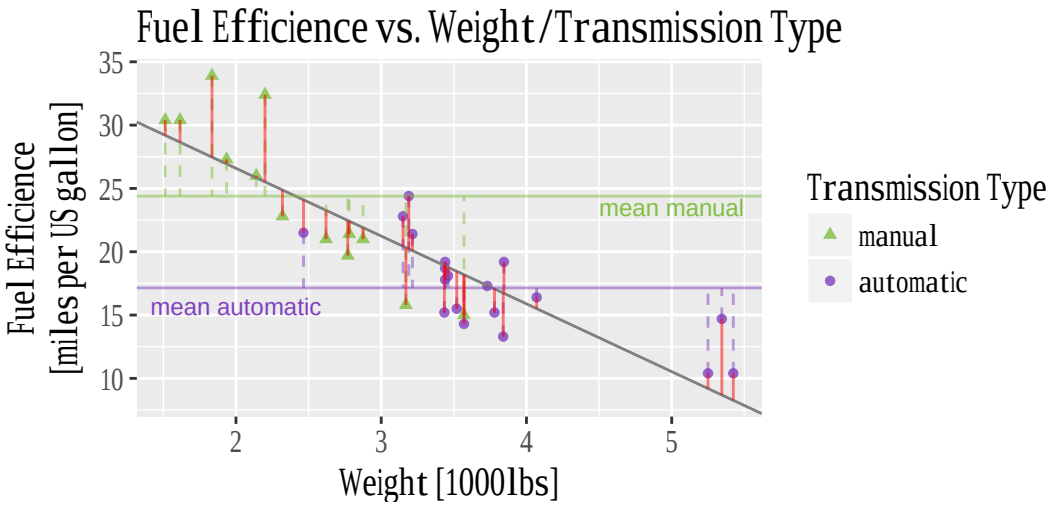

Figure 2: Model 2.2 is based on Model 2.1, but includes the continuous predictor wt. Note how the regression line, which represent the predicted values (grey), reduces the residuals (red lines) with respect to the ones generated by only taking the mean for each group as the predicted value (pointed lines). This illustrates the huge increase in $adj.R^2$ with respect to Model 2.1.

A2.3 - Model 2.3

```
# Regression: main effects and interactions
fit3_0<-lm(mpg~wt*factor(am)+hp*qsec,data=mtcars)
sumfit3_0<-summary(fit3_0)
tablecoef<-data.frame(round(sumfit3_0$coefficients[,1:2],2))
tablecoef$"t value"<-(sumfit3_0$coefficients[,3])
tablecoef$"Pr(>|t|)"<-(sumfit3_0$coefficients[,4])
print(tablecoef)
```

```
##                   Estimate Std..Error    t value     Pr(>|t|)
## (Intercept)          -4.09      10.52 -0.3890436 0.7005379023
## wt                   -1.45       1.02 -1.4282594 0.1655920110
## factor(am)1          16.35       3.74  4.3704984 0.0001904780
## hp                    0.16       0.08  2.0392292 0.0521309002
## qsec                  1.78       0.53  3.3879881 0.0023356297
## wt:factor(am)1       -5.22       1.36 -3.8304988 0.0007646997
## hp:qsec              -0.01       0.01 -2.0823257 0.0477016205
```

```
a<-plot(fit3_0)
```