Regression Model - Final Project

Pierre Baudin

February 21, 2017

Regression Model - Final Project

Data exploration

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).

```
# Import library
library(ggplot2)

# Import data set
data(mtcars)

# Tidy data and reclass to factor
ind <- c("cyl", "vs", "am", "gear", "carb")
mtcars[ind] <- lapply(mtcars[ind], as.factor)
```

Exploratory plots have been placed in the annex section. The preliminary exploration helps to visualize that an automatic transmission yield a lower miles per gallon consumption.

Model fitting and analysis

Linear model - 2 variables

```
mpgam_model$coefficients
```

```
## (Intercept)         am1
##   17.147368    7.244939
```

```
mpgam_model$df.residual
```

```
## [1] 30
```

```
# summary(mpgam_model)
```

This simple linear model shows that on average, a car consumes **17.147 mpg with automatic transmission**, and **7.245 mpg for a manual transmission**. This model has a **Residual standard error of 4.902** (not displayed) on 30 degrees of freedom. The **Adjusted R-squared value is 0.3385** (not displayed). This means that the model can **explain about 34% of the variance** of the MPG variable.

The Adjusted R-squared value is low and indicates that we need to consider other variables to the model in order to better explain what influence the fuel consumption for this set of data.

Adjusting the model

The process for finding a better linear model to explain the change in fuel consumption is as follow: 1. Create a model that include all the variables available 2. Using the R step function, determine the best model in our case.

See https://en.wikipedia.org/wiki/Stepwise_regression)

Summary of the model is included in the annex.

This more complex linear model includes in addition to the transmission type, the variables weight and the 1/4 mile time. The fuel consumption is not explainable anymore by the transmission type only. This model has a **Residual standard error of 2.459** on 28 degrees of freedom. The **Adjusted R-squared value is 0.8336**. This means that the model can **explain about 83% of the variance** of the MPG variable.

Using this model on this dataset, we have a better way of predict the fuel consumption of a car. We can easily understand that the weight of the cars plays a significant role to explain the fuel consumption. The heavier the car, the higher the fuel consumption. As for the 1/4 mile time, I would speculate that as the car get faster for this distance, the fuel consumption goes up.

Going further

Can we explain more variance using other variable in the model?

To go further on the influence of the weight and the transmission type on the fuel consumption, the exploratory data is helping us to find another interesting relationship. Refer to the annex for the plot. By exploring further on this relationship, we can see that the cars with manual transmission weight less that the automatic transmission cars.

Given this information, we can fit it in our model as follow:

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am + wt:am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## wt            -2.937      0.666  -4.409 0.000149 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## am1           14.079      3.435   4.099 0.000341 ***
## wt:am1        -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

This model has a **Residual standard error of 2.084** on 28 degrees of freedom. The **Adjusted R-squared value is 0.8804**. This means that the model can **explain about 88% of the variance** of the MPG variable. Using this model on this dataset, we have continued to improve the way of predict the fuel consumption of a car.

Executive Summary

Comparing our four models using the anova function:

```
anova(mpgam_model, mpgam_model_all, mpgam_model_step, mpgam_model_final)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 3: mpg ~ wt + qsec + am
## Model 4: mpg ~ wt + qsec + am + wt:am
##   Res.Df    RSS  Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     15 120.40  15    600.49 4.9874 0.001759 **
## 3     28 169.29 -13    -48.88 0.4685 0.911413
## 4     27 117.28   1     52.01 6.4795 0.022398 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can observe from the Anova analysis that the residual sum of square (RSS) which is a measure of the discrepancy between the data and the estimation model, decreases as we fine-tune the model.

Residual plots and visual diagnostics are displayed in the annex section.

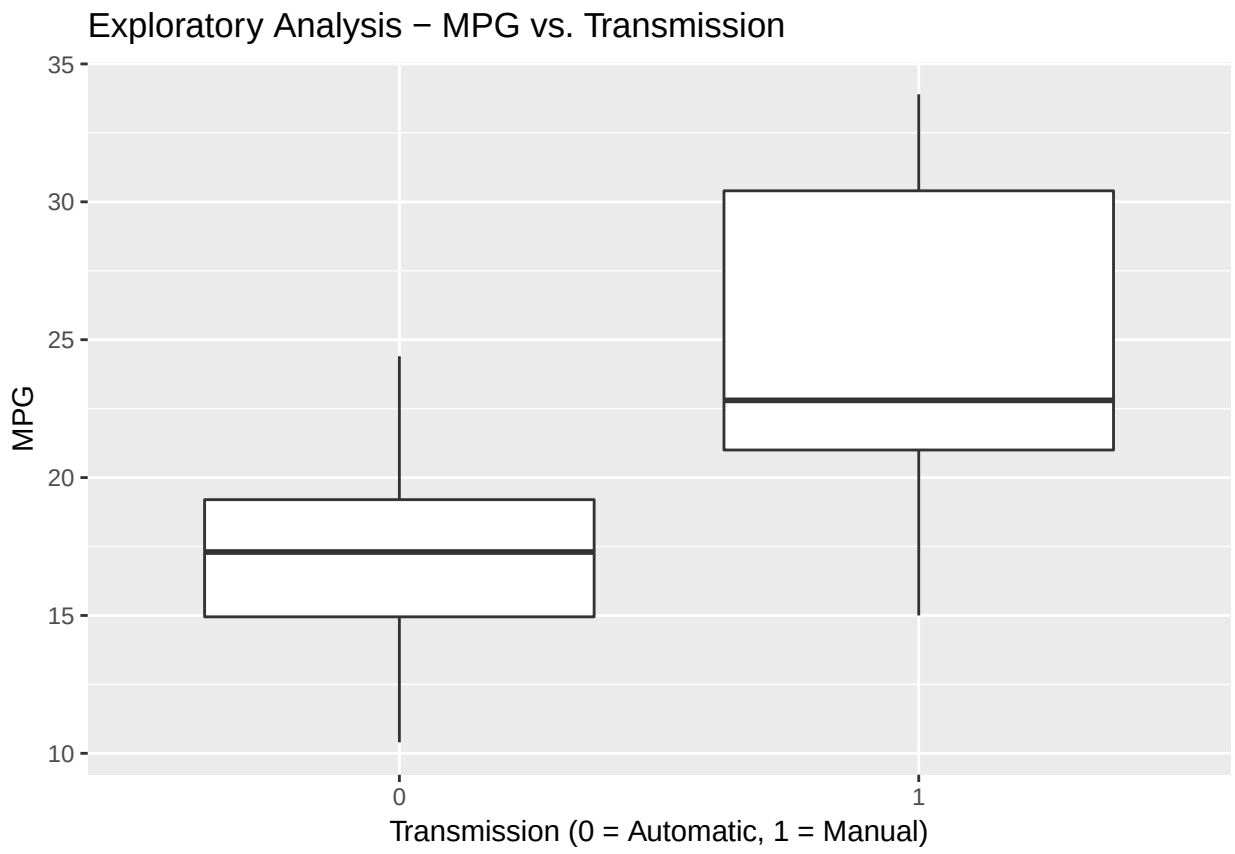
In conclusion, the regression model in this project that best explain the fuel consumption from this dataset includes, on top of the transmission type, the car weight, the 1/4 mile time and the relationship between the weight and the transmission type.

Annexes

EXploratory data analysis

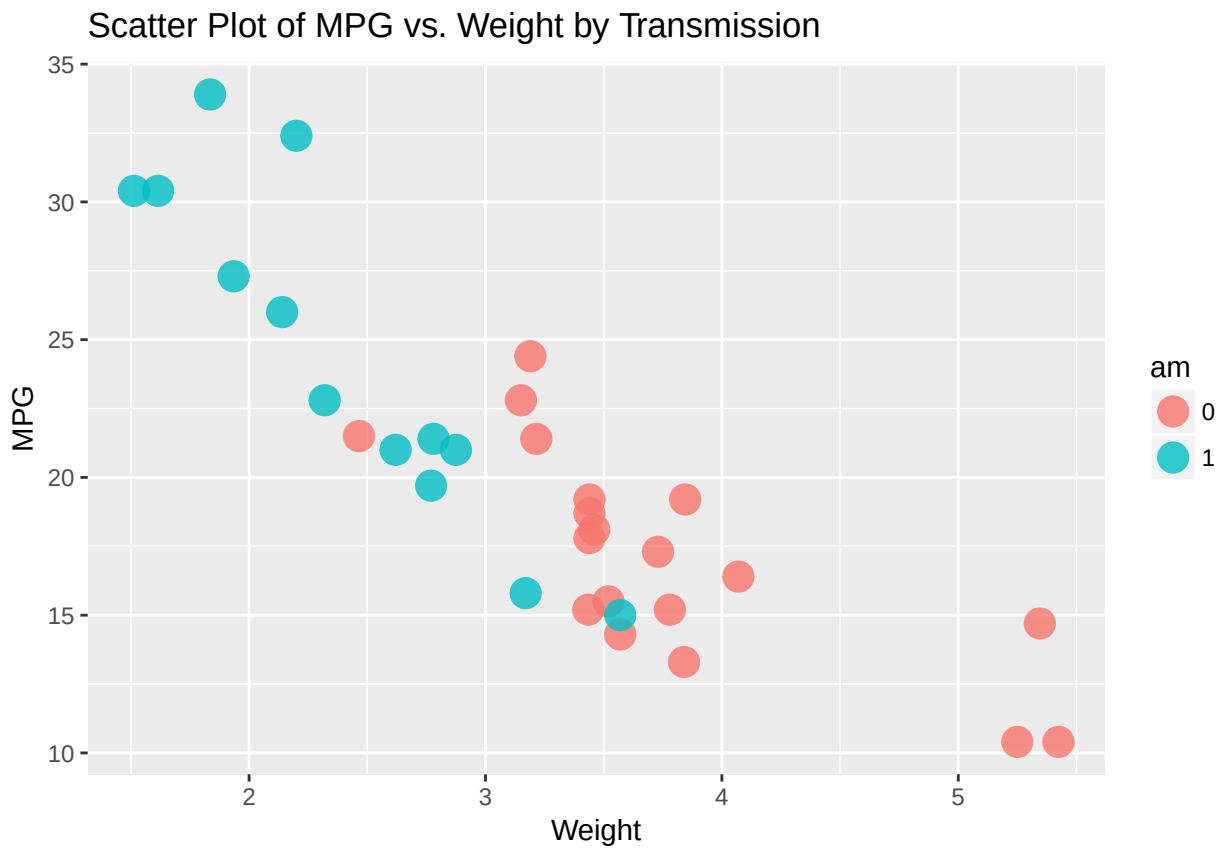
MPG versus Transmission type:

```
ggplot(mtcars, aes(x = am, y = mpg)) +
        geom_boxplot() +
        xlab("Transmission (0 = Automatic, 1 = Manual)") +
        ylab("MPG") +
        ggtitle("Exploratory Analysis - MPG vs. Transmission")
```



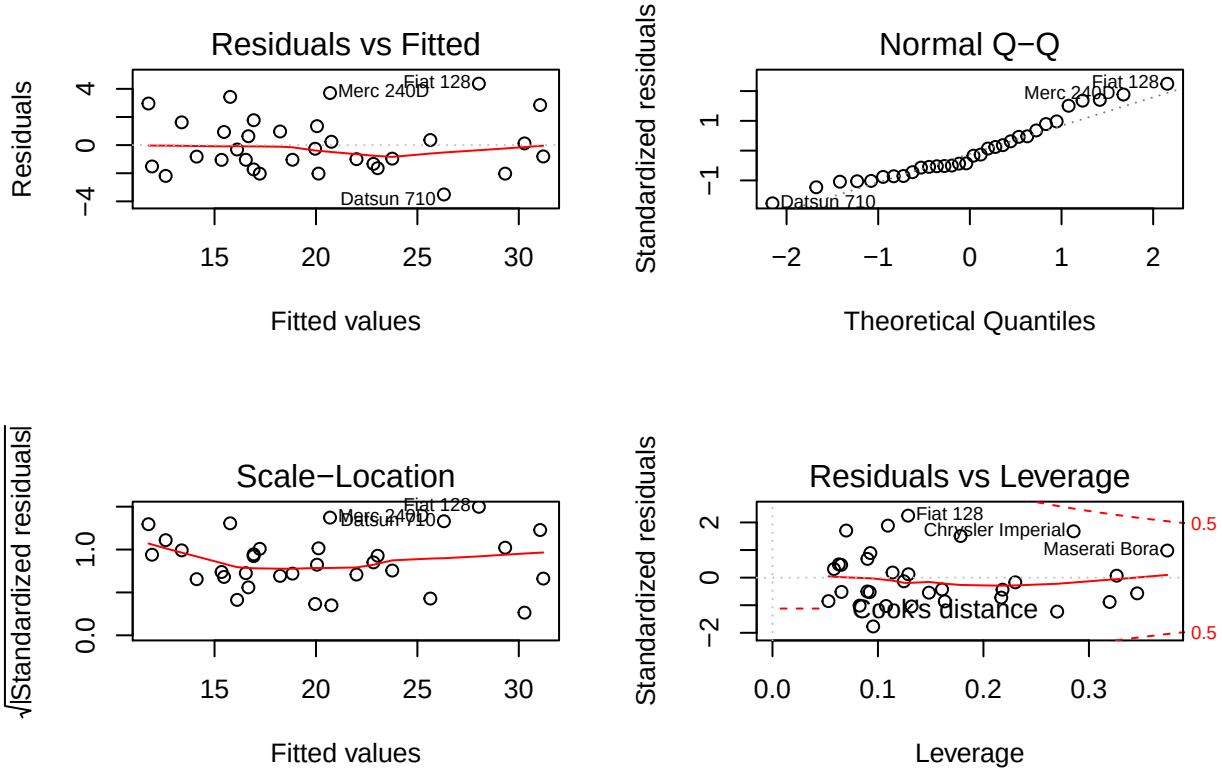
Relationship between MPG, weight and transmission type

```
ggplot(mtcars, aes(x = wt, y = mpg, group = am, color = am)) +
        geom_point(size = 5, alpha = 0.8) +
        xlab("Weight") +
        ylab("MPG") +
        ggtitle("Scatter Plot of MPG vs. Weight by Transmission")
```



Residual plot and visual diagnostics

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



Session info for reproducibility