Regression models Course Project

Alessandro Spilotros

Wednesday, November 02, 2016

Automatic vs Manual Transmission: the best MPG performances

Summary and conclusions

This paper explores the mtcars data set to determine whether an automatic or manual transmission has a better mile-per-gallon (MPG) output. The difference between the two cases is quantified using linear models. The Analysis section of this document presents inference with a simple linear regression model (MPG as a function of transmission type predictor) and a multiple regression model. Both models support the conclusion that the cars with manual transmissions have on average significantly higher MPG's than cars with automatic transmissions. According to the first model, the mean MPG difference is 7.245 MPG; the average MPG for cars with automatic transmissions is 17.147 MPG, and the average MPG for cars with manual transmissions is 24.392 MPG. The multiple linear regression model includes transmission type together with 2 other predictors: wt (weight) and qsec (1/4) mile time. In the multiple regression model, the MPG difference is 2.9358 MPG at the mean weight and qsec. Exploratory analysis and visualizations are located in the Appendix to this document.

Analysis

Exploratory plots are reported in Appendix. Appendix - Plot 1, shows the impact on MPG by transmission: Automatic transmissions has a lower MPG.

Linear Regression model 1 = lm(mpg ~ am, data = mtcars)

```
data(mtcars)
mtcarsdata<-mtcars
# Converting the am variable to factor (0=automatic, 1=manual)
mtcars$am   <- factor(mtcars$am,labels=c("Automatic","Manual"))
fit <- lm(mpg ~ am, data = mtcars)
coef(summary(fit))
```

```
##              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
```

```
cat("R-squared=", summary(fit)$r.squared)
```

```
## R-squared= 0.3597989
```

From this model we observe an increase of 7.245 in MPG of manual transmission in respect to automatic. If we build a 95% confidence interval for the am coefficient we observe that it does not contain the 0. This fact together with the small p-value found (2 10^-4) ensures the statistical significance of the coefficient.

```
predictor_coef=coef(summary(fit))[2,1]
se=coef(summary(fit))[2,2]
alpha=0.05
n=length(mtcars$mpg)
tstat <- qt(1 - alpha/2, n - 2)  # n - 2 degrees of freedom
predictor_coef+c(-1,1)*(tstat*se)
```

```
## [1]  3.64151 10.84837
```

This model explains only 36% (R-squared) of the total variation in MPG. We can improve the model including other predictors. The best choice of the predictors has been taken following this logic: 1. Look for the variables that have the max correlations with MPG but are as orthogonal as possible between them. 2. Try to maximize the R-squared with the minimum number of variables. This is a quite general problem: we have used the algorithm "Leaps and bound" (1974 G.Fournival and R. W. Wilson Technimetrics Vol.16, NO.4, 499) implemented in the R package "bestglm".

Multiple regression model=lm(mpg ~ am+wt+qsec, data = mtcars)

```
#Preparing the matrix for bestglm
X=mtcarsdata[,-1]
Xy=cbind(X,mtcarsdata$mpg)
#Running bestglm with the maximum entropy criteria
#(http://www2.uaem.mx/r-mirror/web/packages/bestglm/vignettes/bestglm.pdf) for documentation
fit2=bestglm(Xy,IC="AIC")
coef(summary(fit2$BestModel))
```

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

```
cat("R-squared=", summary(fit2$BestModel)$r.squared)
```

```
## R-squared= 0.8496636
```

Calculating the 95% confidence interval for the am coefficient

```
predictor_coef=coef(summary(fit2$BestModel))[4,1]
se=coef(summary(fit2$BestModel))[4,2]
alpha=0.05
n=length(mtcars$mpg)
tstat <- qt(1 - alpha/2, n - 4)  # n - 2 degrees of freedom
predictor_coef+c(-1,1)*(tstat*se)
```

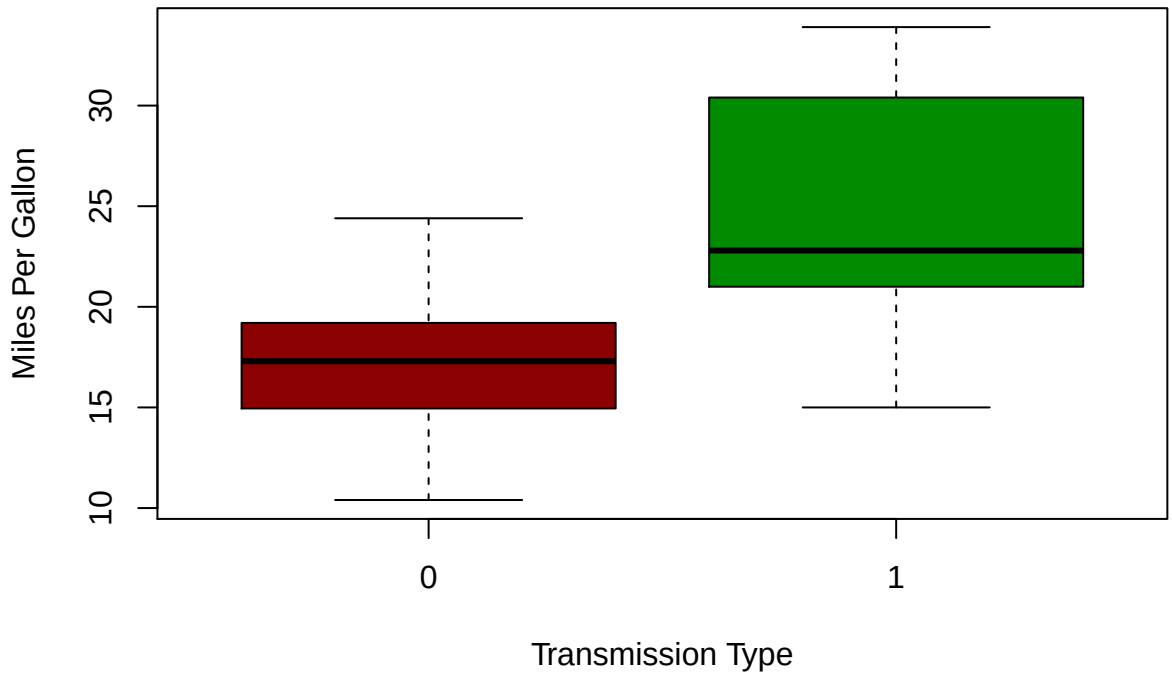
```
## [1] 0.04573031 5.82594408
```

The am coefficient is about 3 and its confidence interval (95%) does not contain 0. The model explains the 85% of the total variation in MPG.

Appendix

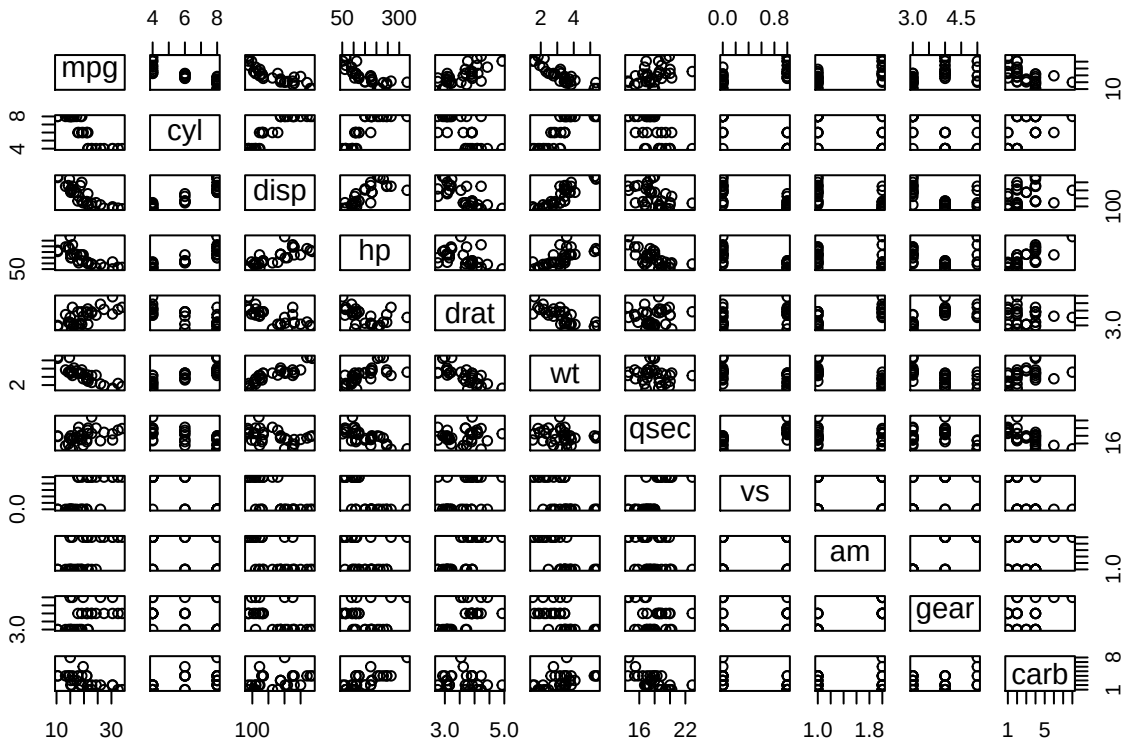
Plot 1: Exploratory data analysis on the mpg and am variables of mtcars dataset

```
boxplot(mpg ~ am, data = mtcarsdata, col = (c("darkred","green4")), ylab = "Miles Per Gallon",
        xlab = "Transmission Type")
```



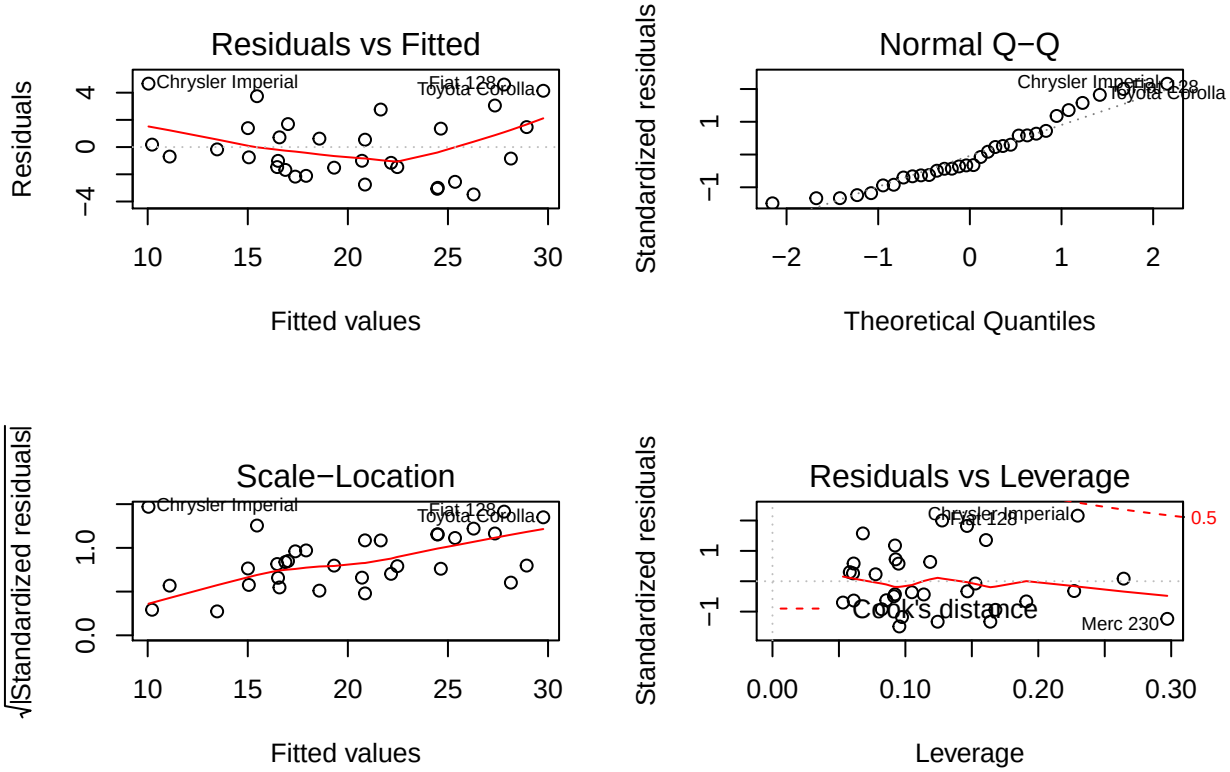
Plot 2: Relation between variables in the mtcars dataset

```
pairs(mpg ~ ., data = mtcars)
```



Plot 3: Residuals

```
par(mfrow = c(2,2))
plot(fit2$BestModel)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
0.5
Merc 230
Leverage