Regression Models Project

V. Boodram

October 22, 2014

Exectuive Summary

Based on the 1974 statistics for fuel consumption and 10 aspects of automobile design for 32 (1973-4) automobile models, provided by the magazine *Motor Trend* (1974), it would like to be determined if automatic or manual transmissions are better for fuel economy, measured in miles per gallon, and what the quantitative difference in mpg between the two transmission types is.

The appropriate predictive model for determining fuel economy with a set of potential predictors from those provided in the `mtcars` data set was determined to be

$$\text{mpg} = 9.68 - 3.917 \times \text{weight} + 1.226 \times \text{quarter-mile time} + 2.936 \times \text{transmission:manual}$$

With this model, it was determined with 95% confidence, all other variables being equal, that cars with a manual transmission have a fuel economy of 0.04107 mpg to 5.83093 mpg higher than cars with an automatic transmission.

Methodology

The full model (i.e: with every variable) was first evaluated from the skeptical point of view that none of the slopes β_i in the set of potential predictors was in fact a statistically significant predictor of the response variable, mpg; the alternative hypothesis was that at least one predictor is a significant predictor of the response variable

$H_0 : \beta_i = 0,\ \forall i$
$H_A : \exists i \ni \beta_i \neq 0$

```
# full model
summary(lm(mpg ~. , data = mtcars))$fstatistic
```

```
## value numdf dendf
## 13.93 10.00 21.00
```

The small p-value indicates that the model as a whole is significant, so the null hypothesis is rejected.

Having found evidence that there is at least one statistically significant predictor of mpg, an initial model was fit to determine if the transmission type was among these.

```
mtcars$am<-as.factor(mtcars$am)
levels(mtcars$am)<-c("automatic", "manual")
summary(lm(formula = mpg~am, mtcars))$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.134e-15
## ammanual       7.245      1.764   4.106 2.850e-04
```

The percentage in variability of fuel economy explained by transmission type, R^2 is 36%, and with a small p-value (0.00029), transmission type was determined to be a significant predictor of fuel economy. The above summary indicates that automatic transmission cars achieve 17.15 mpg, and that all else being held constant, manual cars achieve 7.24 more mpg. The initial model

$$\text{mpg} = 17.17 + 7.24 \times \text{transmission: manual}$$

reflects patterns in the box plots in Appendix Figure 1, composed in the exporatory data analysis.

The plots suggest that higher fuel efficiency is achieved by cars with manual transmissions; however, it may be the case that some other features may have been influencing the fuel consumption. This possibility was evaluated in reference to the bivariate plots in the Appendix Figure 2. These plots show scatter plots between each pair of variables in the data set, as well as the correlation coefficients on the upper half of the matrix. It shows a moderate correlation (0.66) between transmission and fuel efficiency, and significantly stronger correlations between mpg and the number of cylinders, the engine displacement, and the weight of the car, suggesting interactions between the potential set of predictors that may affect the amount of variability in mpg explained by transmission type. The collinearities between transmission type and these other variables are moderately high, with the strongest being between transmission type and weight, and transmission type and rear-axle ratio, which means that some of the information in the latter variables is already being captured by the variable transmission.

The remaining variables were incorporated into the model through a *backwards elimination* process with reference to the p-value. Beginning with the full model (which includes every variable), in each iteration the variable with the highest p-value was discarded, and the model was refit, until all the variables contained in the model were significant, and this was declared the parsimonious model. Two iterations are included below; the complete algorithm is included in the Appendix.

```r
require(nnet); require(dplyr)
# iteration 1
fit<-lm(mpg~., data = mtcars)
ind<-which.is.max(x = summary(fit)$coef[,4])
c(rownames(summary(fit)$coef)[ind], summary(fit)$coef[ind, 4])
```

```
## [1] "cyl"               "0.916087375515962"
```

The p value of the eliminated variable, cyl, is non-significant, and so the variable cyl is discarded as a potential predictor

```r
# iteration 2
fit<-lm(mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb,
         data = mtcars)
ind<-which.is.max(x = summary(fit)$coef[,4])
c(rownames(summary(fit)$coef)[ind], summary(fit)$coef[ind, 4])
```

```
## [1] "vs"                "0.843258496576132"
```

The p-value of vs is also non-significant, so it is eliminated as a potential predictor

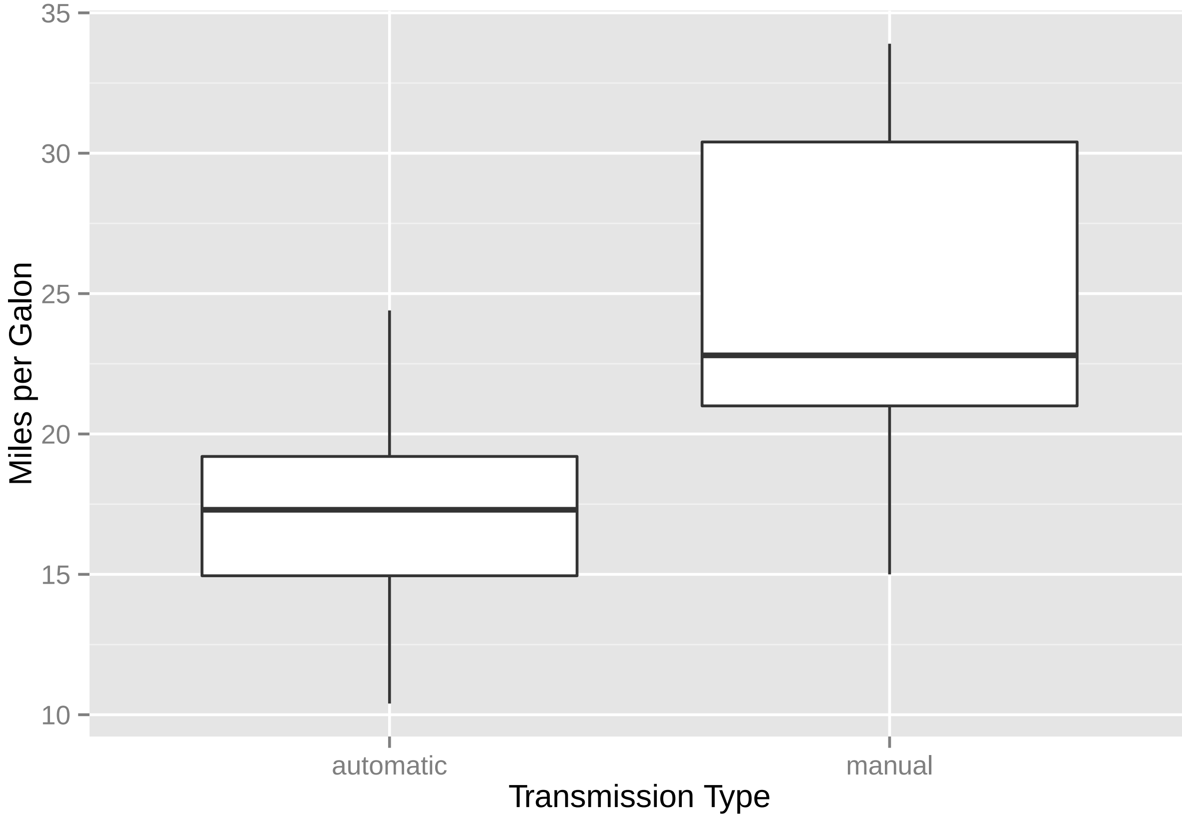
Conclusion

Fuel efficiency is predicted to be between 0.04107 mpg to 5.83093 mpg higher in manual transmission cars than automatic transmission cars with 95% confidence. All else being held equal, a manual transmission car improves the number of miles per gallon by 2.936, compared to a car with an automatic transmission.

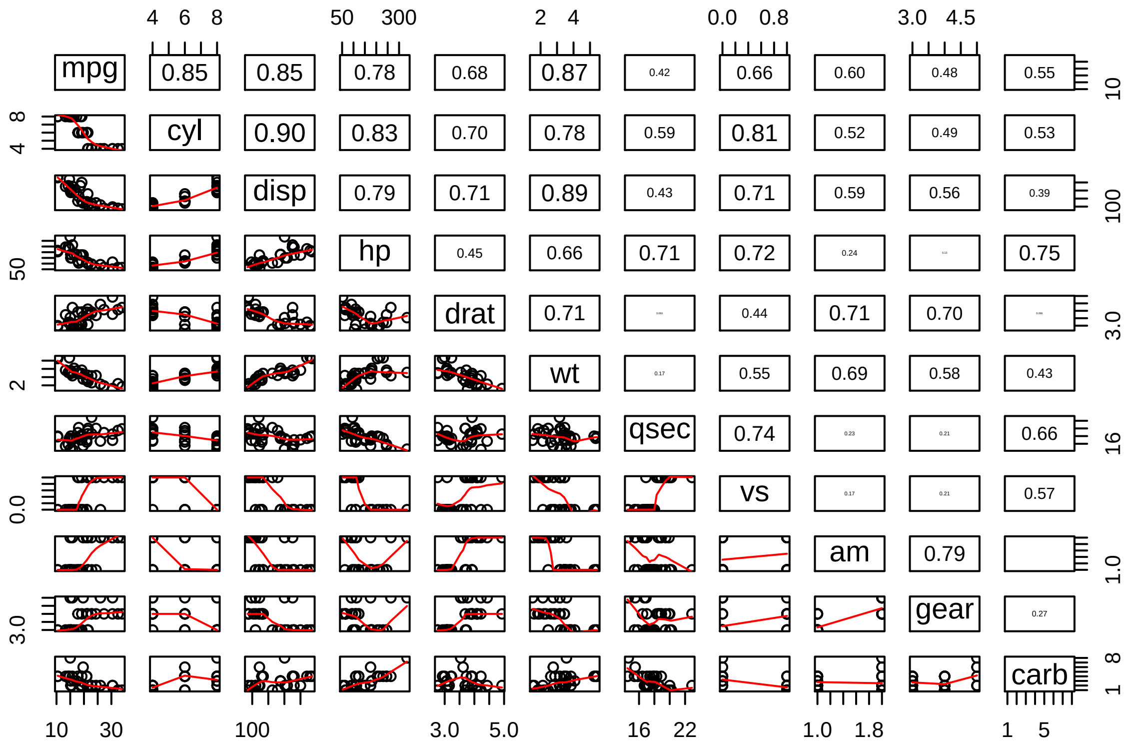
Appendix

Figures

```r
library(ggplot2)
options(height=60)
ggplot(mtcars, aes(factor(am), mpg))+geom_boxplot()+labs(x="Transmission Type", y="Miles per Galon")
```




```
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{   usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)}
pairs(mtcars, lower.panel = panel.smooth, upper.panel = panel.cor)
```



2. Computations

Full Backwards Elimination procedure

```
# iteration 2
fit<-lm(mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb,
         data = mtcars)
ind<-which.is.max(x = summary(fit)$coef[,4])
c(rownames(summary(fit)$coef)[ind], summary(fit)$coef[ind, 4])
```

```
## [1] "vs"                 "0.843258496576132"
```

```
# iteration 3
fit<-lm(mpg ~ disp + hp + drat + wt + qsec + am + gear + carb,
         data = mtcars)
ind<-which.is.max(x = summary(fit)$coef[,4])
c(rownames(summary(fit)$coef)[ind], summary(fit)$coef[ind, 4])
```

```
## [1] "carb"               "0.746958210119909"
```

```
# iteration 4
fit<-lm(mpg ~ disp + hp + drat + wt + qsec + am + gear,
         data = mtcars)
ind<-which.is.max(x = summary(fit)$coef[,4])
c(rownames(summary(fit)$coef)[ind], summary(fit)$coef[ind, 4])
```

```
## [1] "gear"               "0.619640615802269"
```

```r
# iteration 5
fit<-lm(mpg ~ disp + hp + drat + wt + qsec + am,
         data = mtcars)
ind<-which.is.max(x = summary(fit)$coef[,4])
c(rownames(summary(fit)$coef)[ind], summary(fit)$coef[ind, 4])
```

```
## [1] "drat"              "0.462401184664443"
```

```r
# iteration 6
fit<-lm(mpg ~ disp + hp + wt + qsec + am,
         data = mtcars)
ind<-which.is.max(x = summary(fit)$coef[,4])
c(rownames(summary(fit)$coef)[ind], summary(fit)$coef[ind, 4])
```

```
## [1] "disp"              "0.298972149878453"
```

```r
# iteration 7
fit<-lm(mpg ~ hp + wt + qsec + am,
         data = mtcars)
ind<-which.is.max(x = summary(fit)$coef[,4])
c(rownames(summary(fit)$coef)[ind], summary(fit)$coef[ind, 4])
```

```
## [1] "hp"                "0.223087931975388"
```

```r
fit<-lm(mpg ~ wt + qsec + am,
         data = mtcars)
ind<-which.is.max(x = summary(fit)$coef[,4])
c(rownames(summary(fit)$coef)[ind], summary(fit)$coef[ind, 4])
```

```
## [1] "(Intercept)"       "0.177915165458584"
```

This terminates the backwards elimination process.

```r
summary(fit)$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    9.618     6.9596   1.382 1.779e-01
## wt            -3.917     0.7112  -5.507 6.953e-06
## qsec           1.226     0.2887   4.247 2.162e-04
## ammanual       2.936     1.4109   2.081 4.672e-02
```

```r
# confidence interval
2.936+c(-1,1)*qt(p = .975, df = 27)*1.4109
```

```
## [1] 0.04107 5.83093
```