Regression Models Course Project

Johns Hopkins University - Coursera

by Massimo Malandra

Overview:

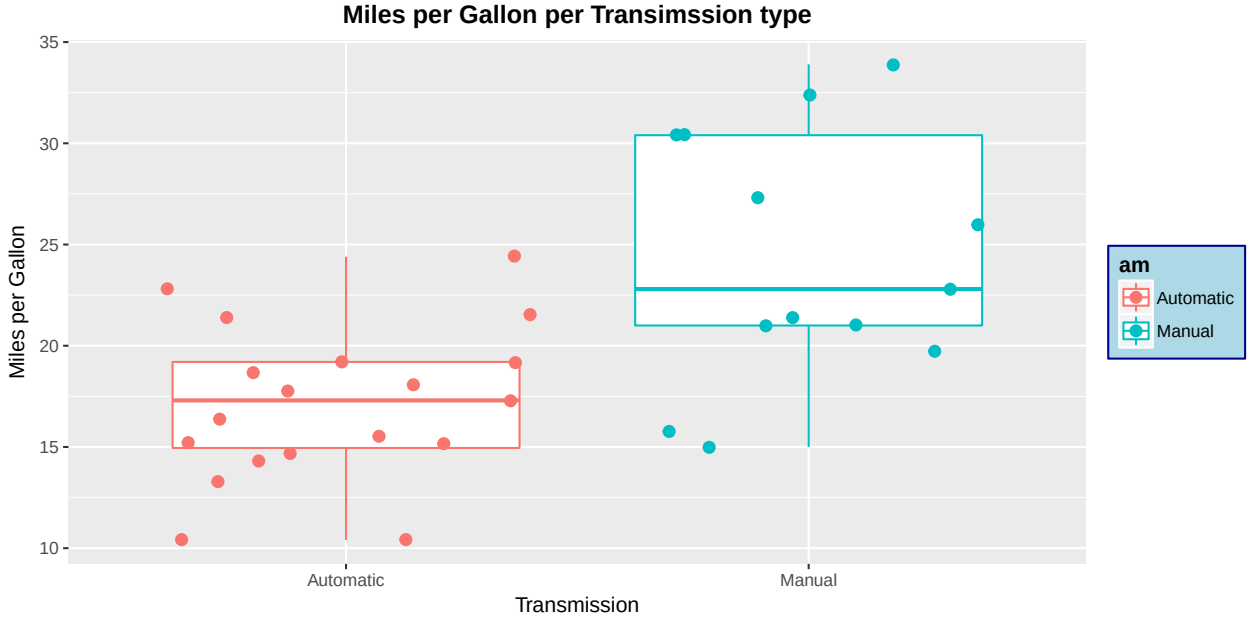
In this project we will conduct a fictional work for Motor Trends, a magazine about the automobile industry. We are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). In particular, we are interested in knowing if an automatic or manual transmission is better for mpg and in quantifying the mpg difference between automatic and manual transmissions.

The provided data set includes 11 variables: more specifically fuel consumption (miles per gallon - mpg) and 10 aspects of automobile design and performance.

First of all, we can have a look at the structure of the data. We can plot the matrix of scatterplots, so that we can have a first and quick visualization of all the variables and the relations that might exist between them (see appendix.)

To proceed to the analysis as requested by Motor Trends, we first have to perform some data preparation. We can reassign the values of the 'am' variable, which represents the Transmission type of the car: 0 will be shown as Automatic, and 1 as Manual. After that we pass the variable (together with 'vs') from numeric to factor.

We can plot the mpg variable in a boxplot that represents the data by transmission type. We can see from the figure above how automatic transmission cars are in general associated with a lower fuel consumption compared to manual: automatic have a lower median compared to manual cars (17.3 vs 22.8), and the 1st quantile of manual cars is 21, which is higher than the 3rd quantile of automatic cars. It can be noted also that the variabilty of the manual cars fuel consumption is much higher, especially in the 2 last quantiles, hence half of the observed data.



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

If we test the null hypothesis that the means for 'am' are the same with respect to 'mpg', the resulting p-value (0.001374) is less than 0.05, or - said in other words - the 95% confidence inverval does not include 0 (-11.280194 -3.209684): so we reject the null hypotesis that the 'am' means are equal, hence correlation might exist.

We can fit a linear model trying to interpret the entity of the relation between miles per gallon and transmission type. The AIC is 196.4844. The adjusted R-squared indicares that only 33.85% of the variability of the data is captured. We can try to fit other models and compare their AIC.

```
##              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
```

We can now fit another model that also aims to explain the mpg variable, but this time not limiting ourselves only to the transmission variable but including all the other variables in the data set.

The AIC indicates that the model including all variables is preferable to the one including the transmission variable only. But the too many variables involved, generate multicollinearlity problems. So we can try to improve our model, applying some variable selection methods to it.

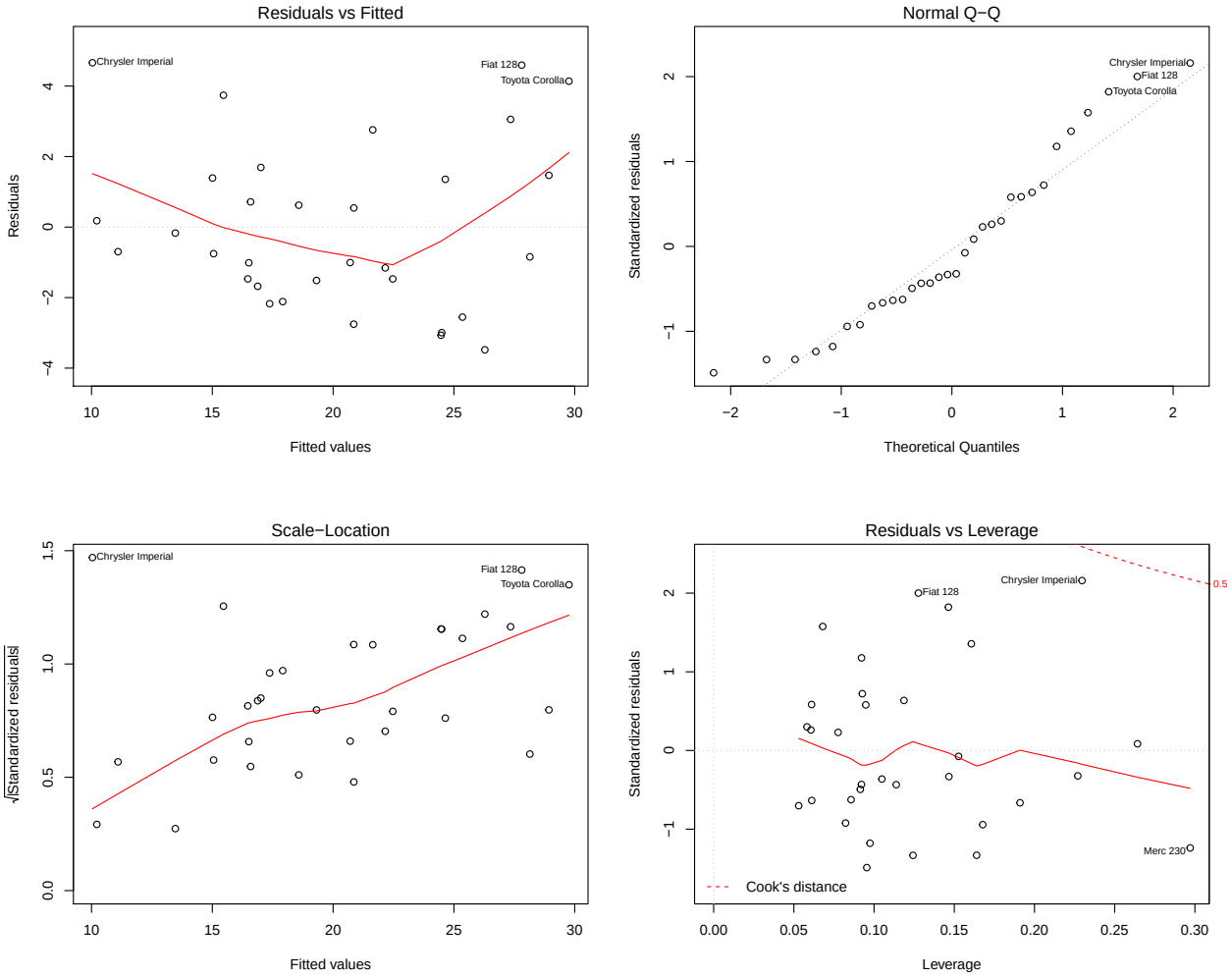
The Backward Elimination procedure, which starts from the global model that combines all the 10 variables and iteratively removes one variable at a time from the model. According to this method, we can keep 3 variables into the model: am, qsec and wt. The AIC is 154.1194.

Another procedure we can use to select variables is the Forward Selection, which starts from the minimal model that doesn't include any predictor variable - so an intercept-only model - and iteratively includes one variable at a time. As for the previous method, this procedure ends with a 3 variables model, but this time the chosen predictors are: wt, cyl, hp. The AIC is 155.4766.

Finally we can try the Stepwise Selection, which combines Backward and Forward method. With this method, the variables we keep in the model are: am, qsec, wt. The same variables selected with the Backward Elimination procedure. The AIC is exactly the same too: 154.1194.

```
##              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
## amManual     2.935837  1.4109045  2.080819 4.671551e-02
```

So we can keep the Stepwise model (which is the same as the Backward in this case), which selects the following subset of variables as most significant to explain the fuel consumption: transmission type, 1/4 mile time and weight. According to this model manual transmission has an impact over mpg 2.9358 higher than automatic transmission. The p-value is below 0.05. The anova test shows a p-value less than 0.05, rejecting the null hypotesis that there is no improvement between the two models. The residuals vs. fitted and the Normal Q-Q for the Stepwise selected model suggest normality, only with a slight deflection.

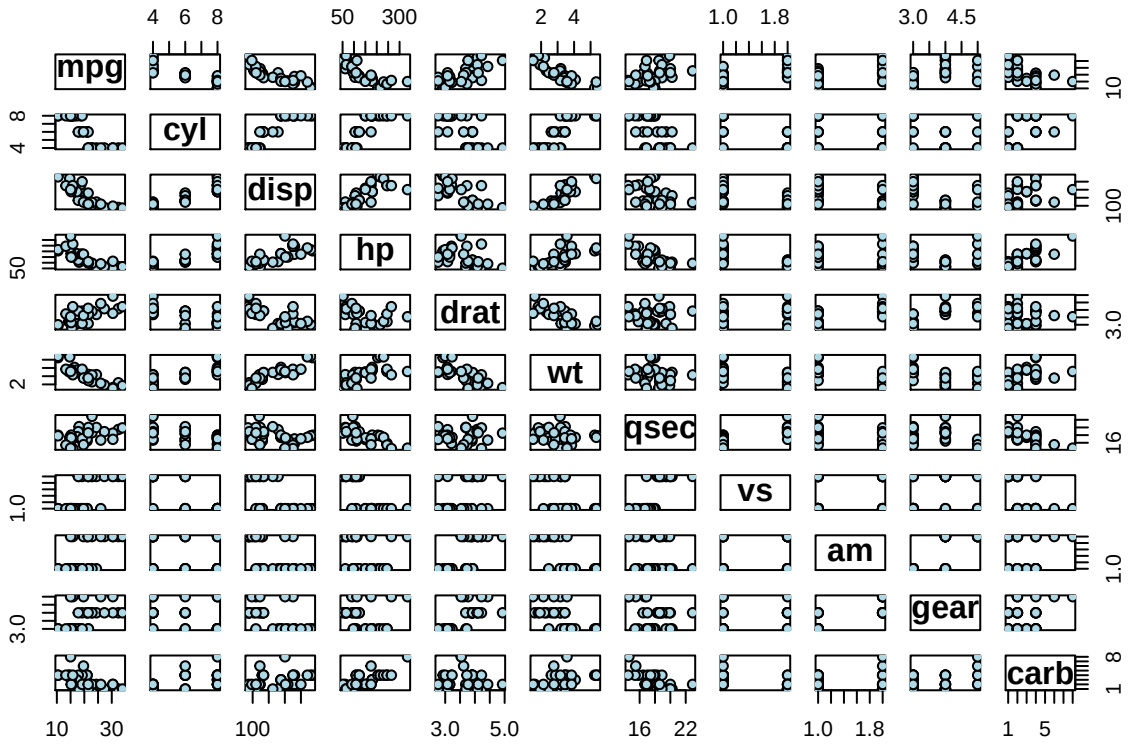


Appendix:

R code:

```
data(mtcars)
?mtcars
str(mtcars)
summary(mtcars)
```

```
pairs(mpg ~ ., data = mtcars,
      cex = 1, pch = 21, bg = "light blue",
      cex.labels = 1.5, font.labels = 2)
```




```
mtcars[mtcars$am == 0, "am"] <- "Automatic"
mtcars[mtcars$am == 1, "am"] <- "Manual"

mtcars$am <- as.factor(mtcars$am)
mtcars$vs <- as.factor(mtcars$vs)


library(ggplot2)
# Boxplot:
ggplot(data = mtcars, aes(x = am, y = mpg, color = am)) +
    ggtitle("Miles per Gallon per Transimssion type") +
    geom_boxplot() +
    geom_jitter(size = 2.5) +
    xlab("Transmission") +
    ylab("Miles per Gallon") +
    theme(plot.title = element_text(face = "bold", hjust = 0.5),
          legend.title = element_text(face = "bold"),
              legend.background = element_rect(fill = "lightblue",
              size = 0.5, linetype = "solid", colour = "darkblue"))

summary(mtcars[mtcars$am == "Automatic", "mpg"])
summary(mtcars[mtcars$am == "Manual", "mpg"])

# t-test:
test <- t.test(mpg ~ am, data = mtcars)
test

# am-onlt Model:
modelLin <- lm(mpg ~ am, data = mtcars)
summary(modelLin)
```

```r
AIC(modelLin)

# All-In Model:
modelAllIn = lm(mpg ~ ., data = mtcars)
summary(modelAllIn)
AIC(modelAllIn)

# Backward Elimination:
modelBackward <- step(modelAllIn, direction = "backward", trace = 1)
summary(modelBackward)
AIC(modelBackward)

# Starting Model:
modelStart <- lm(mpg ~ 1, data = mtcars)
summary(modelStart)
AIC(modelStart)

# Forward Selection:
modelForward <- step(modelStart, direction = "forward",
                     scope = formula(modelAllIn))
summary(modelForward)
AIC(modelForward)

# Anova test:
anova(modelStepwise, modelLin)

# Diagnostics:
par(mfrow = c(2,2))
plot(modelStepwise)
```