Regression Models: Project

Thinkers Park

2023-05-23

1. Executive Summary

This report presents the Regression Models project, part of Coursera Data Science with R specialisation, by Johns Hopkins University.

In this project, "mtcars" dataset is analysed (part of R datasets package), with the aim of answering two questions in particular: 1. Is an automatic or manual transmission better for MPG? 2. Quantify the MPG difference between automatic and manual transmissions.

These questions are addressed in the remainder of this report, organised as follows: Section 2 contains a general overview of the data and basic exploratory analysis; Section 3 - discussion of model fit, adjustments and interpretations; Section 4 - analysis of residuals; Section 5 - predictions; Section 6 concludes.

2. Overview of data

Detailed description of the variables in dataset "mtcars" can be found using R help function.

For the purpose of this analysis, we are particularly interested in (numeric) variables "mpg" (Miles/(US) gallon), and "am" (Transmission, 0 = automatic, 1 = manual), and the relationship - if any - between the two. We will also look at potentially confounding variables and possibility of interaction terms.

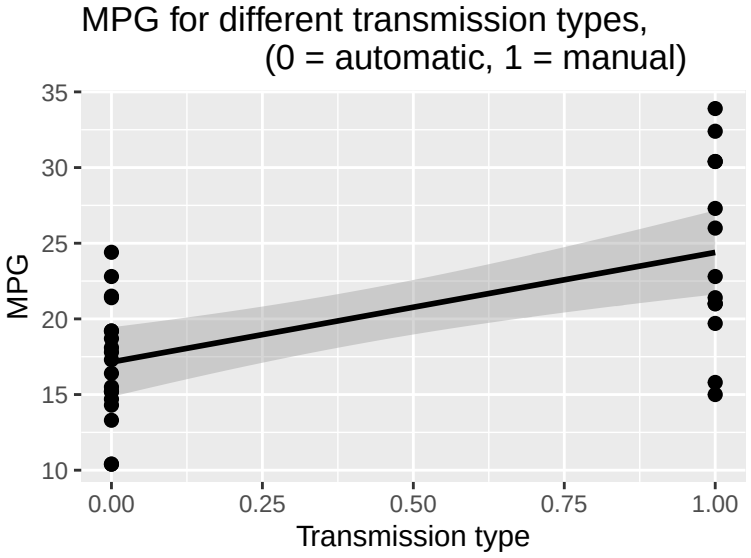

Figure 1: Miles/ (US) Gallon (MPG) for different transmission types.

The two groups of cars (with two transmission types, automatic and manual) in the mtcars data set, can be treated as independent samples, `sample0` and `sample1` respectively. Therefore, the hypothesis can be tested that the average MPG is lower in the automatic transmission group using t.test function:

```
sample0 <- mtcars$mpg[mtcars$am==0]; sample1 <- mtcars$mpg[mtcars$am==1]
t <- t.test(sample0,sample1,alternative="less")
sprintf("T-test p-value: %.4f", t$p.value);  t$estimate
```

```
## [1] "T-test p-value: 0.0007"
```

```
## mean of x mean of y
##  17.14737  24.39231
```

Statistically, the null hypothesis can be rejected, and with sufficient confidence level it can be stated that MPG is lower in the automatic transmission group (17.15 on average) than in the manual transmission group (24.39 on average).

The above test demonstrates commonality, but not yet causality. Intuitively, it is not clear if the specific transmission mechanism causes lower MPG - there may be confounding variables. To identify them, one can look at pair-wise relationships and correlations between variables in the mtcars data set, e.g. by using `ggpairs()` function (the plot would be too large to include, but the plot generation code is included in the Appendix).In this case, possible confounding variables are those displaying high correlation with both "mpg" (Miles/(US) gallon) and "am" (transmission type): displacement ("disp", correlation with "mpg" -0.848/ resp. with "am" -0.591), rear axle ratio ("drat", 0.681/0.713), and weight ("wt",-0.868/-0.692). When fitting a model, we will look at possible confounding variables and interaction terms. All these variables interact with "mpg" and "am" in the same direction (correlation sign is the same).

3. Model fit

First, a "simple" model will be fitted, where MPG is an outcome depending only on the transmission type:

```
simplemodel <- lm(mpg ~ factor(am) ,data = mtcars)
s <- summary(simplemodel)
s$coefficients; sprintf("R squared: %.2f; SSE: %.2f", s$r.squared, s$sigma)
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

```
## [1] "R squared: 0.36; SSE: 4.90"
```

The model shows that the relationship is statistically significant, and is in line with the earlier calculations: For the automatic transmission type (am=0), the average MPG is 17.15 (intercept term), whereas for the manual transmission type (am=1) MPG increases by 7.25 (slope term), resulting in the average MPG of 24.39. Noting that with R squared of 0.36, the model explains relatively little of the overall MPG variability.

Additional "interaction term" models were fitted, i.e. including interaction terms with potential confounding variables identified earlier (each including one potential confounding variable, please see Appendix for model fit & summary code). Noting that all these variables are highly correlated with each other as well, and including more than one/ all of them may lead to overfit and not be optimal.

Based on (i) statistical significance of model coefficients, (ii) highest explanatory power (R-squared), and (iii) lowest variability around the regression line (residual standard error), we select the model with the weight

variable ("wt") and the transmission type ("am), both as standalone explanatory variables, as well as their interaction term. The model is saved down with variable `mymodel`, with an illustration of fitted values as below (please see Appendix for plot generation code).

```
mymodel <- lm(mpg ~ factor(am)*wt ,data = mtcars)
s <- summary(mymodel)
s$coefficients; sprintf("R squared: %.2f; SSE: %.2f", s$r.squared, s$sigma)
```

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    31.416055  3.0201093 10.402291 4.001043e-11
## factor(am)1    14.878423  4.2640422  3.489277 1.621034e-03
## wt             -3.785908  0.7856478 -4.818836 4.551182e-05
## factor(am)1:wt -5.298360  1.4446993 -3.667449 1.017148e-03
```

```
## [1] "R squared: 0.83; SSE: 2.59"
```

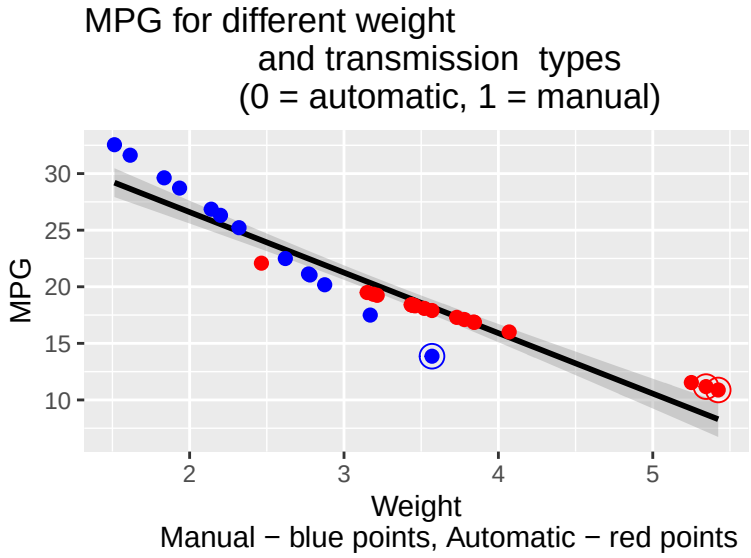

Figure 2: Model Fit: Miles/ (US) Gallon (MPG) for different transmission types. Top 3 hatvalues points encircled.

It is clear that miles per gallon ("mpg") has a decreasing linear dependence on weight ("wt"), and that the decrease is steeper in the manual transmission group than in the automatic transmission group. For a single regression mode fitted across both groups, heavy cars - notably the heaviest car in the manual transmission group, Maserati Bora - are the high-influence data points (encircled on the above plot).

```
sort(hatvalues(mymodel),decreasing=TRUE)[1:3]
```

```
##       Maserati Bora Lincoln Continental   Chrysler Imperial
##           0.3709866           0.3044512           0.2809856
```

4. Residuals

This section provides a closer look at residuals for the selected model. In particular:

(a) Are the residuals homo-scedastic (i.e. have a constant variance)? Breusch-Pagan test does not allow to reject the null hypothesis (that the model residuals have an equal variance), and therefore the variance is assumed to be constant.

(b) Are the residuals normally distributed? Shapiro-Wilk test shows p-value of ab. 0.09, which allows to reject the null hypothesis (that the model residuals are normally distributed) at 90% confidence level, but not at 95%, which is not very conclusive. QQ-plot of the residuals indicates they look reasonably close to normally distributed in the left tail, however with some probability mass distributed differently in the right tail (plot generation code int he Appendix).

```
library(lmtest)
sprintf("BP Test p-value: %.2f; SW Test p-value: %.2f ", bptest(mymodel)$p.value,
        shapiro.test(mymodel$residuals)$p.value)
```

```
## [1] "BP Test p-value: 0.69; SW Test p-value: 0.09 "
```

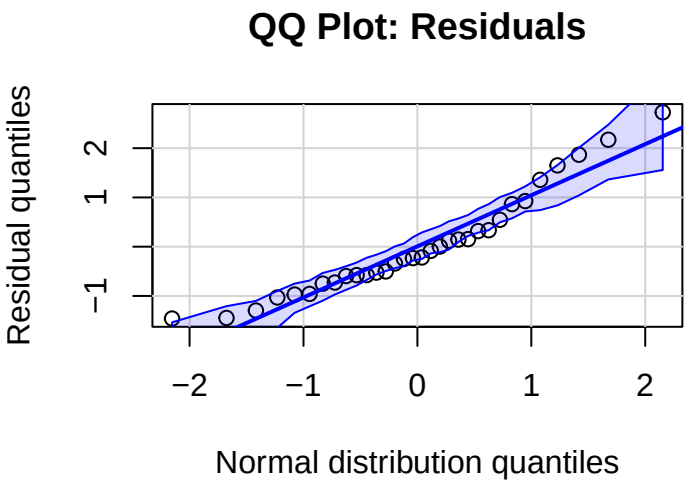

Figure 3: QQ Plot: Residuals

5. Prediction

When constructing the prediction set, one needs to be careful not to include excessively heavy cars with manual transmission (i.e. with weight not representative of this transmission group); otherwise the predicted MPG may be pushed into negative territory, which does not make sense. With this constraint the model predicts reasonably well (compare the plot below vs. Figure 1).

6. Conclusions

So, is an automatic or manual transmission better for MPG? Statistically, MPG is higher in the manual transmission group -so those looking for higher MPG, may prefer to look among manual transmission cars. However, they should primarily look at the car's weight - MPG linearly decreases with increasing weight, and cars with manual transmission are lighter, which then drives up MPG. The selected model includes both the transmission mechanism and the weight, and their interaction term, as explanatory variables; diagnostics and prediction look reasonably well, bearing in mind certain constraints to keep MPG in the positive range.

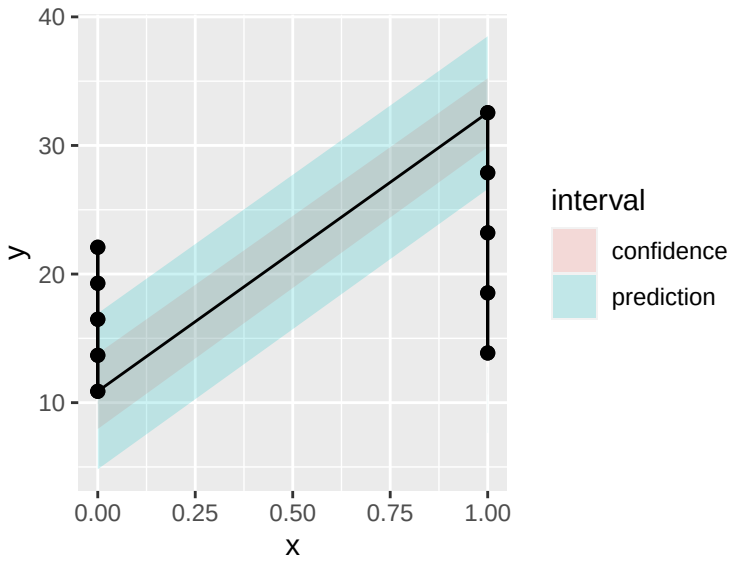

Figure 4: Model prediction

Appendix

Plot generation code

Figure 1: "Miles/ (US) Gallon (MPG) for different transmission types."

```
data(mtcars)
## help("mtcars")
library(ggplot2)
g = ggplot(mtcars, aes(x = am, y = mpg))
g = g + geom_smooth(method = "lm", colour = "black")
g = g + geom_point(size = 2, colour = "black")
g = g + ggtitle("MPG for different transmission types,
                 (0 = automatic, 1 = manual)")
g = g + ylab("MPG") + xlab("Transmission type")
g
```

Pair-plot: Pair-wise relationships and correlations between variables

```
library(GGally)
g = ggpairs(mtcars, lower = list(continuous = "smooth"))
g = g + ggtitle("Pair plot - illustrating linear relationships and correlations
                between pairs of variables in the mtcars data set)")
g = g + ylab("") + xlab("")
g
```

Figure 2: "Model fit: Miles/ (US) Gallon (MPG) vs. weight and transmission type".

```
mtcars = transform(mtcars,y = mymodel$fitted.values)
g = ggplot(mtcars, aes(x = wt, y = y))
g = g + geom_smooth(method = "lm", colour = "black")
```

```
g = g + geom_point(data = subset(mtcars, am == 0),size = 2, colour = "red")
g = g + geom_point(data = subset(mtcars, am == 1),size = 2, colour = "blue")
g = g + geom_point(data = subset(mtcars, rownames(mtcars)=="Maserati Bora"),
                   size = 4, pch=21, colour = "blue")
g = g + geom_point(data = subset(mtcars, rownames(mtcars)=="Lincoln Continental"),
                   size = 4, pch=21, colour = "red")
g = g + geom_point(data = subset(mtcars, rownames(mtcars)=="Chrysler Imperial"),
                   size = 4, pch=21, colour = "red")
g = g + ggtitle("MPG for different weight and
                 transmission types
               (0 = automatic, 1 = manual)")
g = g + ylab("MPG") + xlab("Weight
        Manual - blue points, Automatic - red points")
g
```

Figure 3: "QQ Plot: Residuals".

```
library(car)
qqp(mymodel,distribution="norm", ylab="Residual quantiles - selected model",
    xlab="Normal distribution quantiles",
    main="QQ Plot: Residuals")
```

Figure 4: "Model prediction".

```
x1 <- mtcars$wt[mtcars$am==0]; x2 <- mtcars$wt[mtcars$am==1]
newx = data.frame(cbind(am = c(rep(0,5),rep(1,5)),
                  wt = c(seq(min(x1),max(x1),length = 5),seq(min(x2),max(x2),length = 5))))
p1 = data.frame(predict(mymodel, newdata= newx,interval = ("confidence")))
p2 = data.frame(predict(mymodel, newdata= newx,interval = ("prediction")))
p1$interval = "confidence"; p1$x = newx$am
p2$interval = "prediction"; p2$x = newx$am
dat = rbind(p1, p2); names(dat)[1] = "y"
g = ggplot(as.data.frame(dat), aes(x = x, y = y))
g = g + geom_ribbon(aes(ymin = lwr, ymax = upr, fill = interval), alpha = 0.2)
g = g + geom_line()
g = g + geom_point(data = dat, aes(x = x, y = y), size = 4)
g
```

Model fit & summary code

```
fit1 <- lm(mpg ~ factor(am)*disp ,data = mtcars)
summary(fit1)
fit2 <- lm(mpg ~ factor(am)*drat ,data = mtcars)
summary(fit2)
fit3 <- lm(mpg ~ factor(am)*wt ,data = mtcars)
summary(fit3)
```