Regrssion models - Course Project

Fabrice Tereszkiewciz

16 Oct 2015

Introduction

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions.

Executive summary

Exploratory Analysis

```r
dim(mtcars)
```

```
## [1] 32 11
```

```r
head(mtcars, 5)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
```

```r
# make am a factor, and set better level names
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Automatic", "Manual")
# convert variables to factor
mtcars$vs <- as.factor(mtcars$vs)
mtcars$cyl <- factor(mtcars$cyl)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
glimpse(mtcars)
```

```
## Observations: 32
## Variables:
## $ mpg  (dbl) 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19....
## $ cyl  (fctr) 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4,...
## $ disp (dbl) 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 1...
## $ hp   (dbl) 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, ...
## $ drat (dbl) 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.9...
## $ wt   (dbl) 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3...
## $ qsec (dbl) 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 2...
## $ vs   (fctr) 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1,...
```

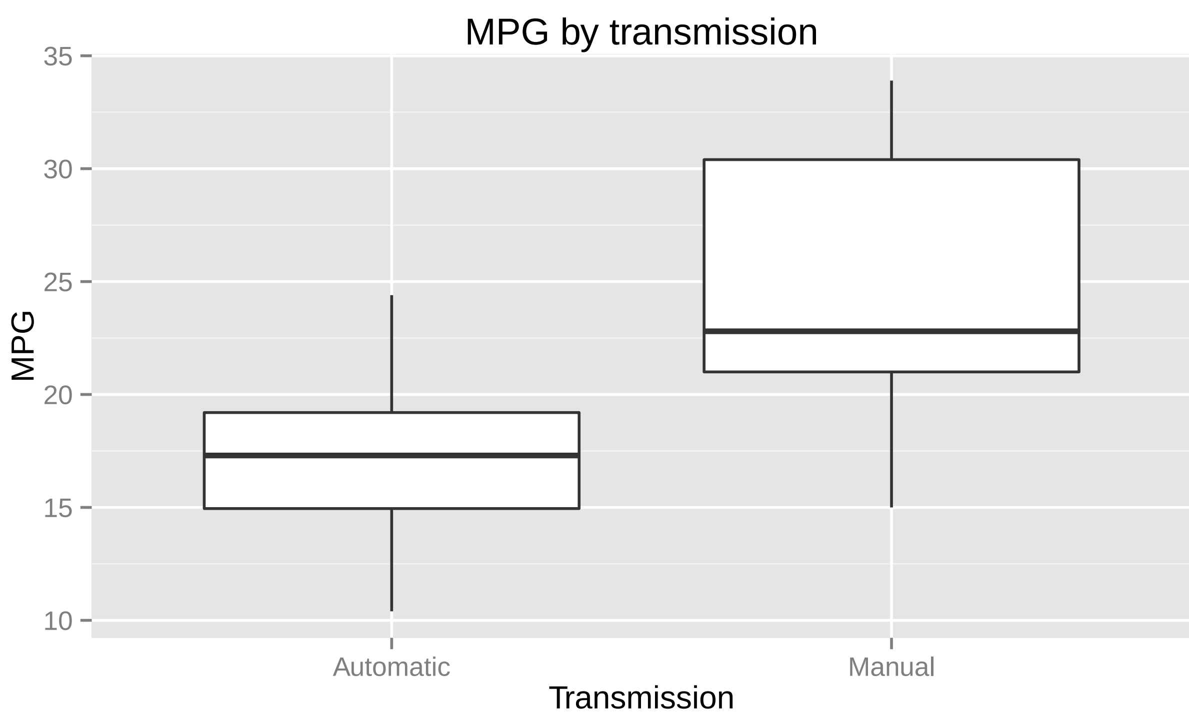
```
## $ am   (fctr) Manual, Manual, Manual, Automatic, Automatic, Automatic,...
## $ gear (fctr) 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4,...
## $ carb (fctr) 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2,...
```

```
summary(mtcars[,'am'])
```

```
## Automatic    Manual
##        19        13
```

The dataset contains 32 observations and 11 variables.

```
g <- ggplot(data=mtcars, aes(x=am, y=mpg)) +
    geom_boxplot() +
    xlab("Transmission") +
    ylab("MPG") +
    ggtitle("MPG by transmission")
g
```



This plot shows the MPG values increasing for the manual transmissions.

Inference

To first determin if the transmission has an impact on the MPG, let H_0 be the null hypothesis that it has no impact : $H_0 : \mu_{auto} = \mu_{manual}$

```
var(mtcars[mtcars$am == 'Automatic',]$mpg)
```

```
## [1] 14.6993
```

```r
var(mtcars[mtcars$am == 'Manual',]$mpg)
```

```
## [1] 38.02577
```

The variance difference is not negligible, we will assume the 2 groups to have unequal variance for the T test.

```r
t.test(mpg ~ am, data=mtcars, paired=FALSE, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

As the 95% interval is `[-11.28 -3.21]` doesn't contain 0 and the P-value is < 0.5, the hypothesis that the transmission is not important can be rejected.

Regression

We can try a marginal linear regression first to see the effect of the transmission type on MPG, holding all other variables constant.

```r
mdl.mar <- lm(mpg ~ am, data=mtcars)
summary(mdl.mar)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

We can see that on average manual transmission vehicules can do 7.245 miles per gallon more than automatic transmission vehicules. Looking at the R-squared error, we can see that only 35.98% of the variance in MPG is explained by the type of transmission. We must therefore look at other variables in order to explain the MPG change.

We can do a step search to find the most optimal model to fit the MPG value.

```
mdl.opt = step(lm(data = mtcars, mpg ~ .),direction = "both")
```

```
summary(mdl.opt)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

The MPG can be better explained by including the weight (`wt`), horse power (`hp`), and cylinder 16 and 8 (`cyl16`, `cyl18`) to the model. With this model, a manual transmission increases the MPG by 1.81 holding all other variables constant.

```
anova(mdl.mar, mdl.opt)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The P-value is highly significant (almost 0), which confirms that the model `mdl.opt` is more accurate than the marginal model `mdl.mar`.

Residuals and Diagnostics

The plots are in appendix :

- there are no signs of correlation between the residuals and fitted values, which is good for a in a homoscedastic linear model with normally distributed errors.
- the QQ plot support normality of the residuals, as all points are close to the line.
- the points look randomly distributed on the Scale-Location plot, so we can assume constant variance.
- all the residuals are well away from the 0.5 Cook's distance, which doesn't indicate any proble with excessive leverage.

Appendix

Models diagnostic

```r
#autoplot(mdl.opt, label.size = 3)
par(mfrow = c(2,2))
plot(mdl.opt)
```