Regression Models Course Project

Ignacio Ojea

July 28, 2018

Assignment

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:

1. "Is an automatic or manual transmission better for MPG"
2. "Quantify the MPG difference between automatic and manual transmissions"

Exploratory data analisis

```
data("mtcars")
?mtcars
# am variable accounts for transmission, and mpg for miles per gallon.
# am, cyl, vs, gear and carb are factors
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
cols <- c("cyl", "vs", "gear", "carb")
mtcars[cols] <- lapply(mtcars[cols], factor)
fit <- lm(mpg ~am, mtcars)
summary(fit)$coefficients
```

```
##              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
```

```
summary(fit)$r.squared
```

```
## [1] 0.3597989
```

See appendix for exploratory analysis plot.

From the plot, as well as the first exploratory linear regression **fit**, the answer to the first question seems positive (coefficient is positive, and p-value indicates significance). But we need to include other variables in order to avoid bias, since the R squared value is 0.36 thus telling us this model only explains us 36% of the variance.

Regression Models

Strategy for model selection

To begin, the type of data does not suggest the need to use either binomial or Poisson generalized linear models, since no variables correspond to binary (or sets of binary) outcomes, nor to counting, number of occurences, or rates.

Hence we need to do a regular multivariable regression. The question now is which are the relevant predictor variables to take into account. Too few will lead to bias, too many to an increase in variance. We need a strategy for model selection. We start with an initial model **fitall** that takes into account all variables, and perfom stepwise model selection to select significant predictors for the final model which is the best model **fitbest** using the *step* function in R, which builds several regression models and makes a selection using the AIC algorithm. I will hide the results for the sake of simplicity, the code is below.

```r
fitall <- lm(mpg ~ ., data = mtcars)
fitbest <- step(fitall, direction = "both")
```

Since the assignment needs to be short, I will avoid studying variance inflation factors (via the *vif* function). Also, the *step* function implicitly exhausts a nested model strategy. Rather, I will *anova* to compare the three models we explored:

```r
anova(fit,fitbest,fitall)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 17.7489 1.476e-05 ***
## 3     15 120.40 11     30.62  0.3468    0.9588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for **fitbest** is significant, therefore we reject the null hypothesis that the added variables cyl, hp and wp are unnecessary. On the contrary, we see that **fitall** has a p-value larger than 0.05, which suggests that the added predictors may not be necessary.

Since the **fitbest** model is accurate, we want to explore how much of the variance it explicates. We do this computing R squared.

```r
summary(fitbest)$r.squared
```

```
## [1] 0.8658799
```

We therefore see that the linear relationship explains about 86% percent of the variance, a great improvement from the 36% corresponding to **fit**.

Inference

We also perform a t-test assuming that the transmission data has a normal distribution and we clearly see that the manual and automatic transmissions are significatively different (by looking at p-value and confidence interval).

```r
t.test(mpg ~ am, data = mtcars)
```

```
##
##  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
```

Residual diagnostics

See the appendix for Residual Plot.

```
cov(fitbest$residuals, hatvalues(fitbest))
```

```
## [1] -0.008314563
```

The points in the Residuals vs. Fitted plot, as well as the analysis of covariance suggest that the variables are independent, namely that there is not correlation between them, as desired. From the rest of the plots we also see that residuales all normally distributed and homoskedastic.

We also see that there are some outliers. Let us study leverage and influence.

```
leverage <- hatvalues(fitbest)
tail(sort(leverage),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora 
##           0.2777872           0.2936819           0.4713671
```

```
influential <- dfbetas(fitbest)
tail(sort(influential[,6]),3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona 
##         0.3507458         0.4292043         0.7305402
```

Answers to the questions

```
summary(fitbest)
```

```
## 
## 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
```

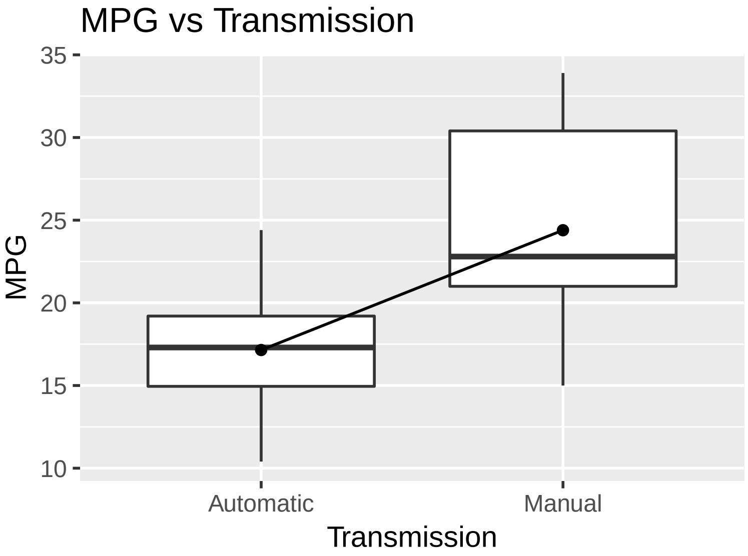
Based on the observations from our **fitbest** model, we can conclude the following,

- Cars with Manual transmission get more miles per gallon compared aganist cars with Automatic transmission. (1.8 adjusted by hp, cyl, and wt). mpg will decrease by 2.5 (adjusted by hp, cyl, and am) for every 1000 lb increase in wt.
- mpg decreases negligibly with increase of hp.
- If number of cylinders, cyl increases from 4 to 6 and 8, mpg will decrease by a factor of 3 and 2.2 respectively (adjusted by hp, wt, and am).

Appendix with plots

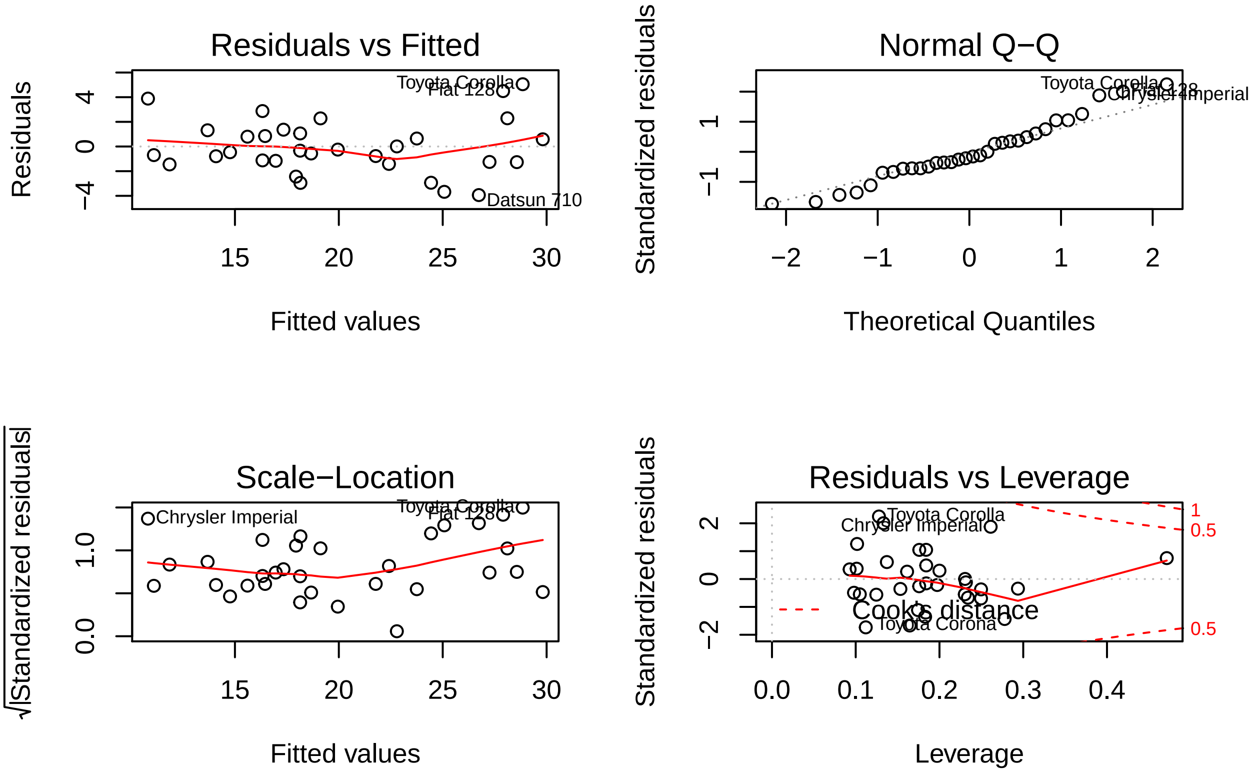
Exploratory analysis plot:

```
g <- ggplot(mtcars) + aes(x = factor(am), y = mpg) + geom_boxplot()
g <- g + stat_summary(fun.y=mean, geom="line", aes(group=1)) + stat_summary(fun.y=mean, geom="point") +
g
```



Residual Plots

```
par(mfrow = c(2, 2))
plot(fitbest)
```


Residuals vs Fitted
Residuals
Fitted values
4
0
−4
15
20
25
30
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
1
−1
−2
−1
0
1
2
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.0
15
20
25
30
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
2
0
−2
0.0
0.1
0.2
0.3
0.4
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5