Regression Models Course Project

Daniela Gracia

May 18, 2021

Synopsis

In this project we will be using the mtcars data set, this data originates from the 1974 Motor Trend US magazine. We are interested in exploring the relationship between a set of variables and miles per gallon (MPG). Particularly, we want to answer the following questions: **1.¿Is an automatic or manual transmission better for MPG?**, **2.Quantify the MPG difference between automatic and manual transmissions**. Our study concludes that when creating a multivariate regression model it is not possible to say that one transmission is better than the other, therefore it is not possible to quantify this relationship using a multivariate model. Other models should be evaluated to see if this relationship can be better explained and quantified. ### Exploratory Data Analysis First we load and explore the data (see Appendix 1).

We see our data set contains 11 variables, the units and meaning of those variables can be found in the ?mtcars page.

Now we create a box plot (see Appendix 2) to see the distribution of MPG for automatic vs manual cars (there are 19 automatic and 13 manual vehicles). In the graph we can see that manual vehicles have a higher average of 24.39 vs automatic cars which have an average of 17.15.

Model Selection & Regression Analysis

Now that we have a better understatement of the distribution of the data, we will consider some of the variables and decide which ones we want to include in our models. The models will be tested through nested modeling. Our goal is to see which variables help us best describe the relationship between mpg and transmission type.

Univariate Regression First we see the relationship between mpg and transmission by fitting a linear regression (see Appendix 3).

We interpret this model in the following way: for automatic transmission the mean mpg is $17.147 \pm 1.13\ mpg$ and for manual transmission we expect an average increase of $7.245 \pm 1.76\ mpg$. Both of these values are statistically significant with p values < 0.05. This model suggests that manual transmissions are better for mpg. Additionally, The r-squared value implies that these model is able to explain 36% of the total variation. Now we will move on to nested modeling to see what is the best way to explain more variation while maintaining a consistent model.

Multivariate Regression As we have a relatively small number of independent variables we will run the anova() function for the complete model and select the variables that have the most statistically significant values (see Appendix 4).

We see that the variables cyl, weight and disp are, in that order, statistically significant (pvalue < 0.05). So we will use them to create 3 models we believe will help explain the relationship between MPG and automatic/manual vehicles. Then we compare these models using anova() (see Appendix 5).

We can see that including cyl in the model produces highly statistically significant results and so does the addition of weight. However when adding disp the pvalue of 0.5 (»0.05) suggests that this is not necessary in the model.

We believe that model 3: **lm(mpg ~ transm + cyl+ weight)** is the best/most parsimonious relationship we can provide. Now, lets take a look at the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

X_1 is the transmission, 0 if automatic and 1 if manual, X_2 is the number of cylinders (4, 6 or 8) and X_3 is the weight of the vehicle (1000 lbs). Let´s take a look at the estimated coefficients.

```
##
## Call:
## lm(formula = mpg ~ transm + cyl + weight, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.4179     2.6415  14.923 7.42e-15 ***
## transmmanual   0.1765     1.3045   0.135  0.89334
## cyl           -1.5102     0.4223  -3.576  0.00129 **
## weight        -3.1251     0.9109  -3.431  0.00189 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

We interpret this model in the following way: The r-squared value implies that this model is able to explain 83% of the total variation, which is a significant improvement from the 36% the previous model explained. Considering all other predictors remain constant:
β_0: Suggests an average mpg usage for automatic transmissions of 39.42 ± 2.64 *mpg*.
β_1: Suggests an estimated average 0.18 ± 1.30 *mpg* increase for manual transmissions, however the error is much larger than the estimate and the p-value suggests this is not a significant result.
β_2: Suggests a strong negative relationship between mpg and cyl, for every one unit increase in cyl there is an average 1.51 ± 0.42 *mpg* decrease.
β_3: Suggests a strong negative relationship between cyl and mpg, for every 1000 lbs increase in weight there is an average 3.13 ± 0.91 *mpg* decrease.

β_0, β_2 and β_3 provide significant results with p-values under 0.05, however β_1 does not seem to be a significant value as it fails to reject the null hypothesis: $H_0 : \mu_a = \mu_m$. Lets take a look at the residuals to see if they suggest anything about our model.

Residual Analysis and Conclusions

First we produce some plots to study the residuals (see Appendix 6). From the plots we can see:

Residuals vs Fitted: This plot doesn't have any major trends, the only thing to note is that there are slightly higher residuals for smaller and higher fitted values. **Normal Q-Q:** This plot suggests that the errors follow a normal distribution, which is how we want this plot to look. **Scale-Location:** This plot has a semi-horizontal trend with the data points equally distributed around it, which suggests nothing abnormal about it. **Residuals vs Leverage:** None of the points in this plot are outside Cooks distance, which means none of the points are influential.

In conclusion, the residual analysis suggests the model fit is okay, it would be interesting to look at what is causing the slight pattern in the residuals vs fitted values plot, this might help better explain an quantify the relationship between MPG and transmission in vehicles.Finally, we answer the questions:
1.¿ Is an automatic or manual transmission better for MPG ?: The univariate linear regression model has statistically significant evidence suggesting that manual transmission is better for MPG, however this model fails

to explain a lot of the variation. The multivariate linear regression is able to explain a lot of the variation, however it fails to prove that one transmission is better than the other, in this model it seems like the number of cylinders and weight have a more significant impact on MPG.

2. Quantify the MPG difference between automatic and manual transmissions For the univariate regression there is an average expected increase of 7.245 ± 1.76 *mpg* for manual vehicles. For the multivariate model this relationship cannot be quantified as the results are not significant.

In conclusion the multivariate regression model created provides a consistent and parsimonios explanation of the relationship between MPG, transmission, cyl and weight. It would be interesting to run more model comparison tests and residual analysis to see if there is a model that explains this relationship better.

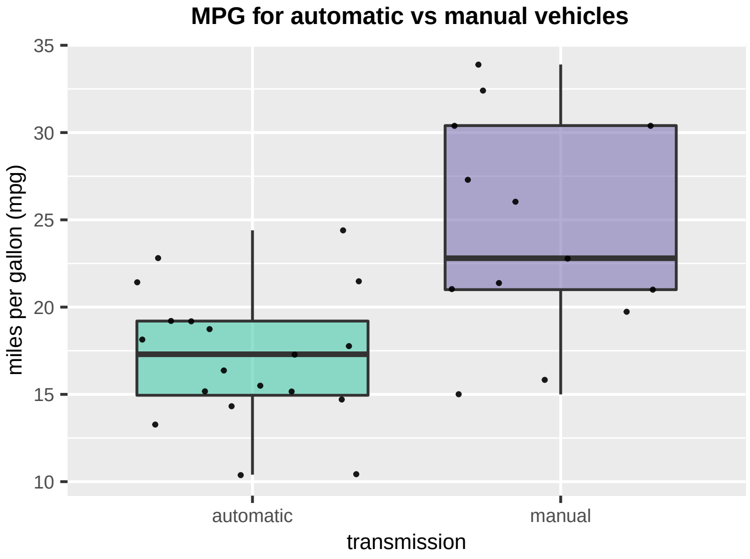
Appendix

Below is all additional information needed to understand the report.

Appendix 1

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Appendix 2



Appendix 3

```r
# create first model
model1 <- lm(mpg ~ transm, data=mtcars2)
summary(model1)
```

```
##
## Call:
## lm(formula = mpg ~ transm, data = mtcars2)
##
## 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 ***
## transmmanual    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
```

Appendix 4

```r
# we build the complete model
model <- lm(mpg~.,data=mtcars2)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## cyl        1 817.71  817.71 116.4245 5.034e-10 ***
## disp       1  37.59   37.59   5.3526  0.030911 *
## hp         1   9.37    9.37   1.3342  0.261031
## drat       1  16.47   16.47   2.3446  0.140644
## weight     1  77.48   77.48  11.0309  0.003244 **
## qsec       1   3.95    3.95   0.5623  0.461656
## engine     1   0.13    0.13   0.0185  0.893173
## transm     1  14.47   14.47   2.0608  0.165858
## gear       1   0.97    0.97   0.1384  0.713653
## carb       1   0.41    0.41   0.0579  0.812179
## Residuals 21 147.49    7.02
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

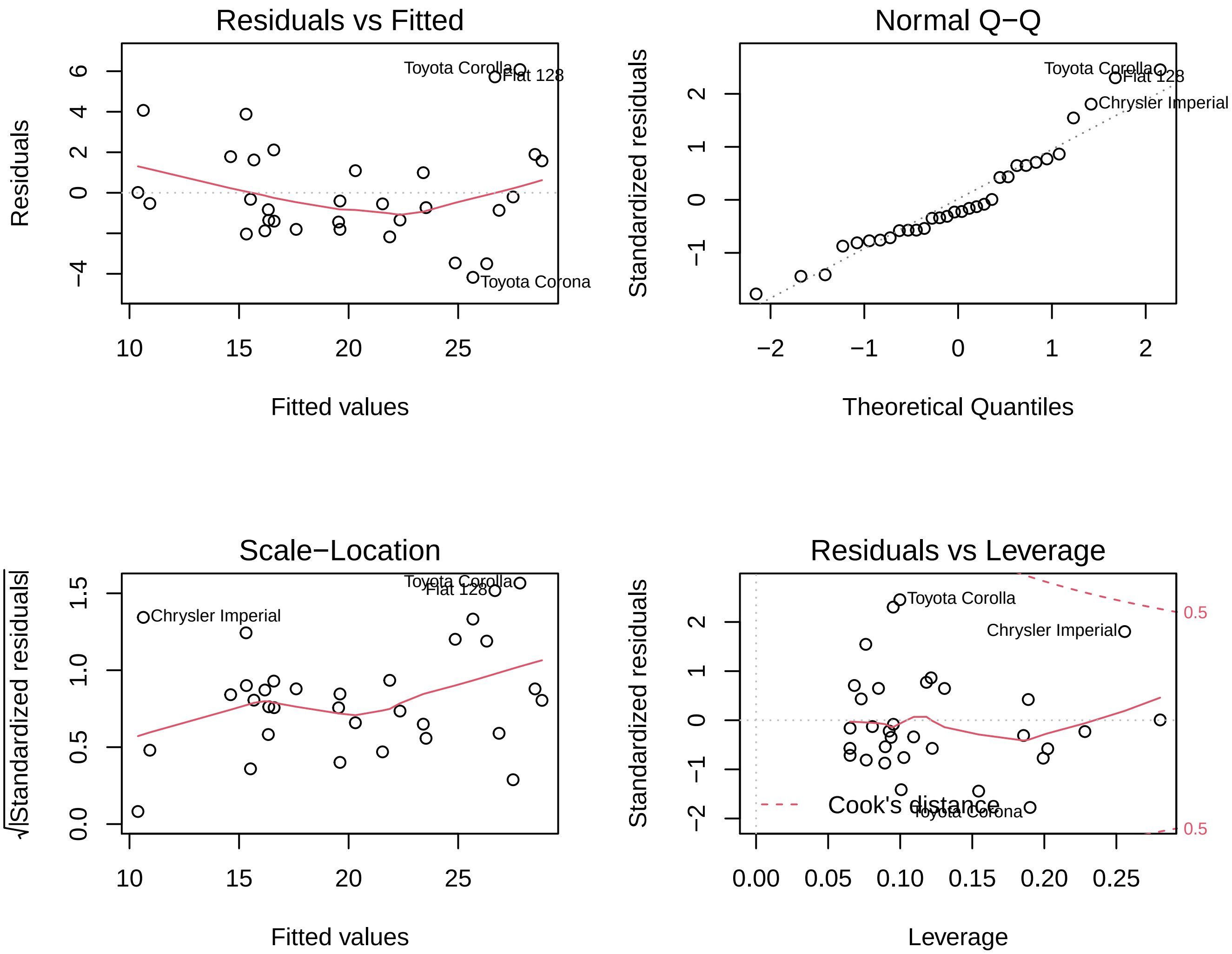
Appendix 5

```r
# first model includes cyl
model2 <- lm(mpg ~ transm + cyl, data=mtcars2)
# second model includes cyl and weight
model3 <- lm(mpg ~ transm + cyl+ weight, data=mtcars2)
# third model includes cyl, weight and disp
model4 <- lm(mpg ~ transm + cyl + weight + disp, data=mtcars2)
# we compare the models
anova(model1, model2, model3, model4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ transm
## Model 2: mpg ~ transm + cyl
## Model 3: mpg ~ transm + cyl + weight
## Model 4: mpg ~ transm + cyl + weight + disp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1     30 720.90                                   
## 2     29 271.36  1    449.53 64.4149 1.264e-08 ***
## 3     28 191.05  1     80.32 11.5085  0.002152 ** 
## 4     27 188.43  1      2.62  0.3756  0.545093    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix 6

```r
# build residual analysis plots
par(mfrow = c(2,2))
plot(model3)
```


Residuals vs Fitted
Residuals
Fitted values
10
15
20
25
−4
0
2
4
6
Toyota Corolla
Fiat 128
Toyota Corona
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−2
−1
0
1
2
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
0.0
0.5
1.0
1.5
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
0.00
0.05
0.10
0.15
0.20
0.25
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
0.5
0.5