Regression Models Project

Vesa Kuoppala

20 marraskuuta 2015

Executive summary

This analysis is performed for the *Motor Trend*, a magazine about the automobile industry. By looking at a data set of a collection of cars, we are interested in exploring the relationship between a set of variables and miles per gallon (MPG) as outcome. We are particularly interested to explore:

- *"Is an automatic or manual transmission better for MPG?"*
- *"Quantify the MPG difference between automatic and manual transmissions"*

In order to answer to these questions we performed exploratory data analyses and used hypothesis testing and linear regression as methodologies to make inference. We established both simple and multivariate linear regression analysis. However, the results of the multivariate regression model is more promising as it includes the potential effect of other variables on MPG.

To answer to the question we found out that **using simple model manual transmission is better for MPG and it is about 7 MPG more for cars with manual transmission.** However, further analysis indicate that we must consider other variables in our analysis to understand better how the transmission affects MPG.

Using model selection strategy, we found out that **among all variables weight and quarter mile time (acceleration) have significant impact in quantifying the difference of MPG between automatic and manual transmission cars, on average, manual transmission cars have 2.94 MPGs more than automatic transmission cars.**

This report is longer than recommended 2 page content + 3 page appendix for pictures. The author felt it necessary to use more space to evaluate models and to come to conclusion about the chosen model.

Data Set

For the purpose of this analyis we use `mtcars` dataset, which is a dataset that was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973 - 1974 models). First six records of the dataset are shown below:

```
##                    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
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Exploratory data analyses

We begin the analysis by performing some initial explaratory data analysis to get better idea of the existing patterns between variables in the data set. Pairwise scatterplot of variables are in Figure 1 in Appendix

It is also worthwhile check how MPG varies by automatic versus manual transmission. Figure 2 in Appendix has boxplot of MPG by automatic and manual transmissions.

We can make a hypothesis from visualization: it appears that automatic cars have a lower miles per gallon a.k.a lower fuel efficiency, than manual cars. Is this just a random chance, do we just picked a group of automatic cars with low efficiency and a group of manual cars with higher efficiency ? Let's do statistical test.

Model fitting and hypothesis testing

Two samples t-test

We are interested to know if an automatic or manual transmission is better for MPG. So we test hypothesis that cars with an automatic transmission use more fuel than cars with manual transmission. To compare two samples to see if they have different means, we use two sample T-test.

```
##
##  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 0 mean in group 1
##        17.14737        24.39231
```

p-value for the probability that the difference between the two groups could appear by chance is very low. The confidence interval tells us how much lower the miles per gallon is in manual cars than it is in automatic cars. We can be confident that the true difference is between 3.2 and 11.3.

Simple linear regression model

We can also fit factor variables as regressors and come up with analysis of variance as a special case of linear regression models.This case we use transmission (`am`) as factor variable.

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

The intercept of `17.14` is the mean MPG for automatic transmission. The slope `7.24`is the change in the mean between manual and automatic transmission. The P value for the slope is significant. **So we can conclude according to this model that the manual transmission is more fuel efficient.**

Multivariate linear regression model

Modelling based on only one predictor variable does not seem to be sufficient and good enough (Adjusted R-squared explains only `0.3385` of residuals in the model). In this part we develop a model that include the effect of other variables.

Model selection

We want to know what combination of predictors will best predict the fuel efficiency. If we include all the variables in the model, none of them will be a significant predictor of MPG (Based on p-value at 0.95 confidence level).

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## factor(vs)1  0.31776281  2.10450861  0.1509915 0.88142347
## factor(am)1  2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

Detecting collinearity

A major problem with multivariate regression is collinearity. If two or more predictor variables are highly correlated, and they are both entered into a regression mode, it increases the true standard error and you get a very unstable estimates of the slope. We can assess the collinearity by variance inflation factor (VIF).

```
## Warning: package 'car' was built under R version 3.2.2
```

```
##        cyl       disp         hp       drat         wt       qsec
##  15.373833  21.620241   9.832037   3.374620  15.164887   7.527958
## factor(vs) factor(am)       gear       carb
##   4.965873   4.648487   5.357452   7.908747
```

Values for the VIF larger than 10 are considered large, so we have lot of variables, which maybe have collinearity.

Stepwise selection method

Among available methods we decided to perform stepwise selection to help select a subset of variables that best explain the MPG.

```
##              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
## factor(am)1  2.935837  1.4109045  2.080819 4.671551e-02
```

This shows that in addition to transmission, weight of the vehicle as well as acceleration speed have the highest relation to explaining the variation in mpg. The adjusted Rˆ2 is 83 % which means that the model explaisns 83 % of the variation in mpg indicating a robust and highly predictive model.

Nested likelihood ratio test

Let verify our result from stepwise seletion model with nested likelihood test

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + wt
## Model 3: mpg ~ factor(am) + wt + qsec
## Model 4: mpg ~ factor(am) + wt + qsec + hp
## Model 5: mpg ~ factor(am) + wt + qsec + hp + drat
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 72.536 5.362e-09 ***
## 3     28 169.29  1    109.03 17.870 0.0002579 ***
## 4     27 160.07  1      9.22  1.511 0.2299925
## 5     26 158.64  1      1.43  0.234 0.6326111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result is consistent with stepwise selection model and adding more variables will dramatically increase variation in the model and the p-value immediately becomes insignificant.

Fitting the final model

We can fit the final model

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## factor(am)1   2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

All the variables are now statistically significant. This model explains 83% of variances in miles per gallon (mpg). When we read the coefficient for `am`, we can say that, **on average, manual transmission cars have 2.94 MPGs more than automatic transmission cars.** However, this effect was much higher when we did not adjust the weight and qsec in our model.

Regression diagnostics

We perform some diagnostics on our final mode. Plots are found in Appendix

Detecting collinearity

```
##         wt       qsec factor(am) 
##   2.482952   1.364339   2.541437
```

This time VIF numbers are reasonable

Residuals versus the fitted values

By plotting residuals versus fitted values, we are looking for any sort of pattern. Plots show that there are no specific pattern in the residuals.

Normality of residuals

The normal Q-Q plot shows us a reasonable linearity in residuals when when quantiles is changed.

Influental Observations

Residuals versus leverage has no outliers, as all values are within the 0.5 bands.

Appendix

Figure 1. Pair Graph of Motor Trend Car road Tests

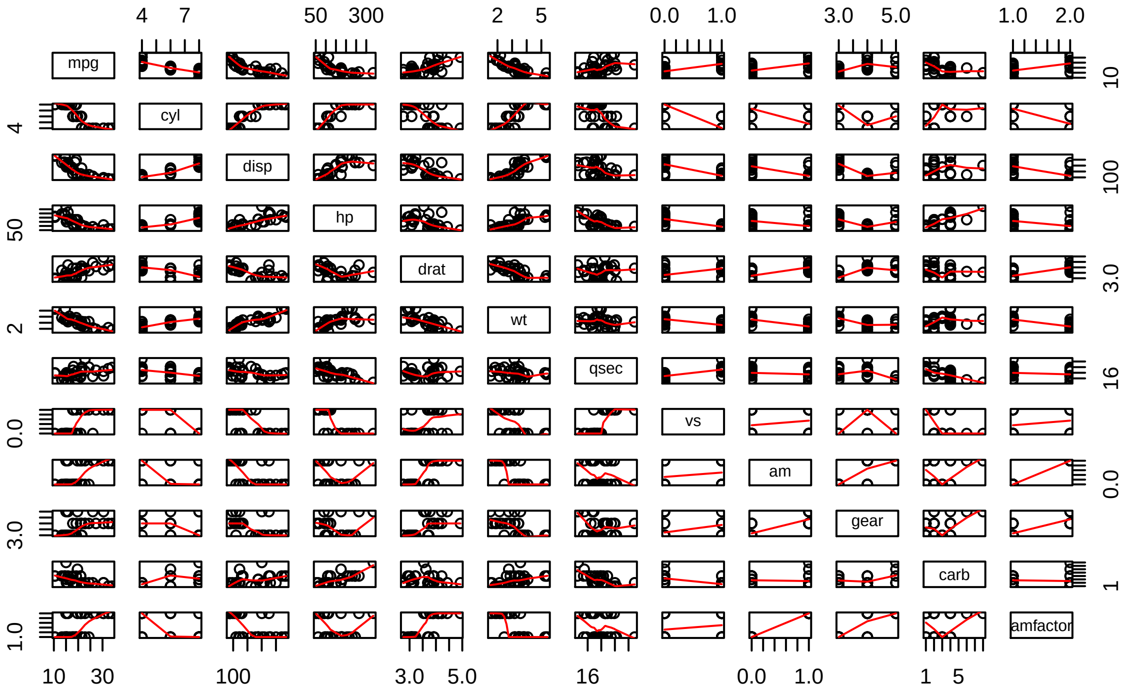


Figure 2. Boxplot of MPG vs. Transmission

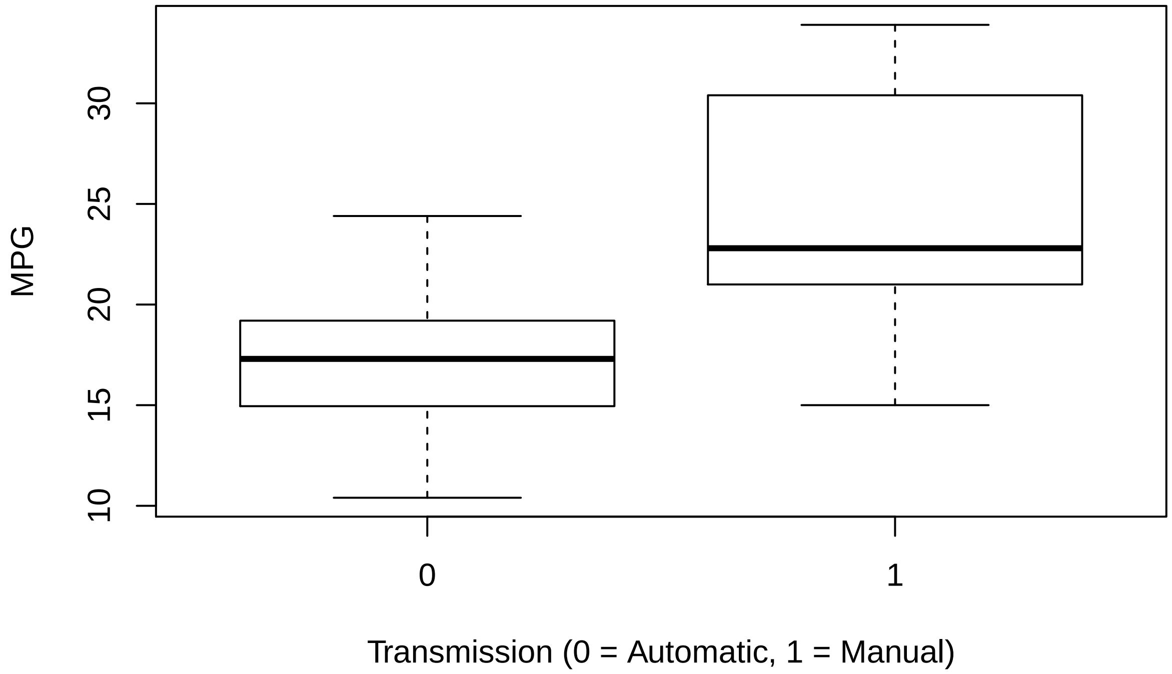
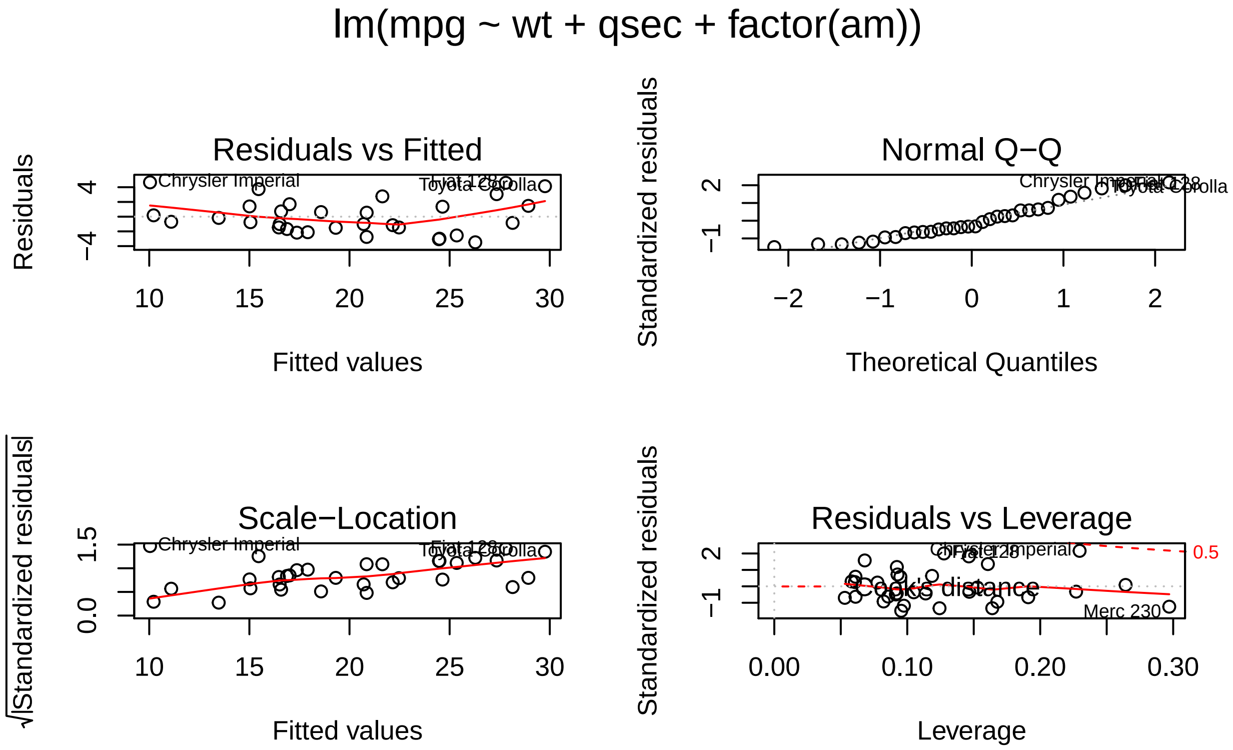


Figure 3. Residual analysis of final model


lm(mpg ~ wt + qsec + factor(am))
Residuals vs Fitted
Residuals
4
−4
10
15
20
25
30
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
2
−1
−2
−1
0
1
2
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Toyota Corolla
Scale−Location
√|Standardized residuals|
1.5
0.0
10
15
20
25
30
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals vs Leverage
Standardized residuals
2
−1
0.00
0.10
0.20
0.30
Leverage
Chrysler Imperial
Fiat 128
Merc 230
Cook's distance
0.5