Regression Models Course Project

Yifan XIA

11 Sep 2017

Executive summary

Based on data collected from different cars, *Motor Trend* magazine investigated potential influences of car characteristics, in particular that of the transmission system, on its Mile Per Gallon (MPG). A regression analysis is performed through model selection and inferential analysis on each model's relevance. We found that a car's MPG strongly depends on its weight, number of cylinders, the transmission system, and the interaction thereof. Regarding the transmission system, our study concludes that manual transmission generally yields higher MPG than automatic one, although some exception may potentially exist.

Exploratory Data Analysis

We start by loading the *mtcars* dataset and show first entries.

```
data("mtcars"); head(mtcars, 3)
```

```
##                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
```

Several variables of numerics type are converted to factors (code not shown).

Next, we compute statistics of cars' **mpg** grouped by **am**:

- Mean MPG

```
aggregate(mtcars$mpg, by = list(mtcars$am), FUN = mean)
```

```
##   Group.1        x
## 1       0 17.14737
## 2       1 24.39231
```

- Median MPG

```
aggregate(mtcars$mpg, by = list(mtcars$am), FUN = median)
```

```
##   Group.1    x
## 1       0 17.3
## 2       1 22.8
```

Both statistics suggest that manual transmission yields higher MPG. The same observation can be made from **Fig. 1** in **Appendix**.

Inference

A quick T-test confirms that the two transmission systems yield different MPGs:

```
TTest <- t.test(mpg ~ am, data = mtcars); TTest$p.value
```

```
## [1] 0.001373638
```

The resulting *p-value* is small (< 0.05), suggesting that there is a non-zero difference between the two means. The variable MPG does not follow a normal distribution regardless of the transmission system.

More inferential analyses will also be performed throughout the model selection process.

Regression and Linear Model Fitting

We fit different linear models to the dataset and assess the robustness of each. 1. The full model

```
fullModel <- lm(formula = mpg ~ ., data = mtcars)
summary(fullModel)
```

The results (not shown in the report) give the following observations:
1. $R^2 = 0.779$, suggesting that the full model explains 77.9% of variations;
2. All coefficients have excessively large P-values (> 0.05), suggesting none are statitcally significant.

We need to investigate other linear models.

Stepwise regression

We apply a stepwise model selection by exact AIC algorithm. This method is arguably capable of selecting the "best" subset of regressor variables from a larger set.

```
require(MASS)
```

```
## Loading required package: MASS
```

```
stepA <- stepAIC(fullModel, direction = 'both', k = 2)
summary(stepA)
```

The model selected by the algorithm AIC is mpg ~ cyl + hp + wt + am, with $R^2 = 0.8401$. However, the summary (not shown) suggests that not all coefficients are statistically significant (in particular *Intercept*). Therefore some diagnosis should be performed. We first look at the VIF of the retained model:

```
require(car); vif(stepA)
```

```
## Loading required package: car
```

```
##         GVIF Df GVIF^(1/(2*Df))
## cyl 5.824545  2        1.553515
## hp  4.703625  1        2.168784
## wt  4.007113  1        2.001778
## am  2.590777  1        1.609589
```

It shows that **hp** and **wt** has similar variance inflation factor, implying they might be correlated. This can also be observed in **Fig. 2** in the Appendix. We decide to remove **hp** since the weight is a more commonly used characteristic. This figure show also a likely interaction between **wt** and **am** (Cars with automatic transmission are seemingly heavier). Thus an interaction term **wt:am** is added in the new model.

The intercept, *i.e.* the mpg for a vehicle of zero weight, makes no practical sense and should be removed.

Our selected model is therefore: **mpg** ~ **cyl** + **wt** + **am** + **wt:am - 1**, *i.e.* a car's MPG depends strongly on its number of cylinders, weight, transmission and the interactions thereof.

```
finalModel <- lm(mpg ~ cyl + wt + am + wt:am - 1, data = mtcars)
summary(finalModel)$adj.r.squared
```

```
summary(finalModel)$coef
```

```
##          Estimate Std. Error   t value     Pr(>|t|)
## cyl4   29.774836  2.8403415 10.482836 7.870715e-11
## cyl6   27.065059  3.0485831  8.877914 2.378052e-09
## cyl8   24.998726  3.4848285  7.173589 1.282823e-07
## wt     -2.398713  0.8439884 -2.842116 8.603904e-03
## am1    11.568790  4.0877912  2.830083 8.853842e-03
## wt:am1 -4.067981  1.3974151 -2.911075 7.295503e-03
```

The selected model describes almost 99% of variation in the outcome ($R^2 = 0.9879$), and all coefficients are statistically significant (maximum of P-value is $0.00885 < 0.05$). The model is a considerably satisfactory one, as one can see from **Fig. 4** in Appendix.

Residual diagnosis

Some residual diagnosis is performed and plotted in **Fig. 3** in Appendix. We can draw the following conclusion:
1. The Residual vs. Fitted plot shows no remarquable pattern, suggesting independence between the two;
2. The Normal Q-Q plot shows the residuals are normally distributed;
3. The Scale-location plot shows randomly distributed points, confirming constant variance;
4. The Residuals vs. Leverage shows the absence of outlier.

Conclusion: Is an automatic or manual transmission better for MPG?

In order to answer this question, we perform some simulations using the final model. We create a mock dataset that contains only two "dummy cars", whose characteristics are identical except the transmission system variable **am**.

```
predictedMpg <- predict(finalModel, DummyCars, interval = 'confidence')
cbind(DummyCars, predictedMpg)
```

```
##         cyl    wt am      fit      lwr      upr
## Dummy.1   4 1.835  0 25.37320 22.16487 28.58153
## Dummy.2   4 1.835  1 29.47724 27.63101 31.32348
```

The above results suggest:
1. Manual transmission is predicted to yield higher Mile Per Gallon (**fit** variable);
2. Statistically, the above conclusion could not be drawn with absolute certainty, as the confidence interval of *Dummy 1* overlaps (although slightly) that of *Dummy 2*.

So the final conclusion is: manual transmission would in most cases yield higher MPG; however, some exceptions may exist.

Appendix: Figures

Figure 1: Boxplot of MPG

```
## Loading required package: ggplot2
```

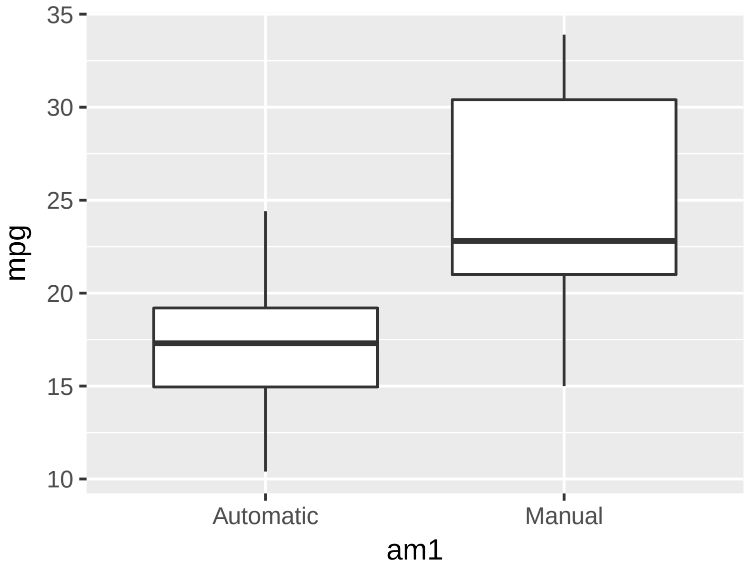


Figure 2: Pair graph of mpg ~ cyl + hp + wt + am variables

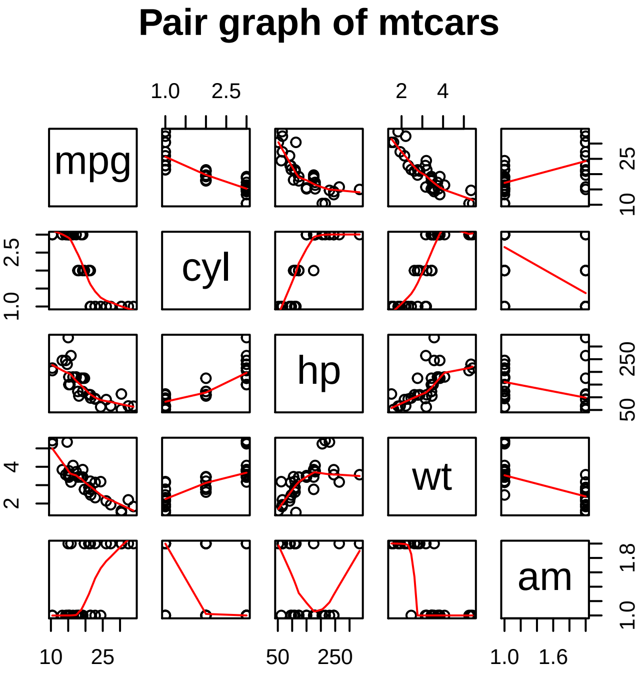


Figure 3: Residual plots

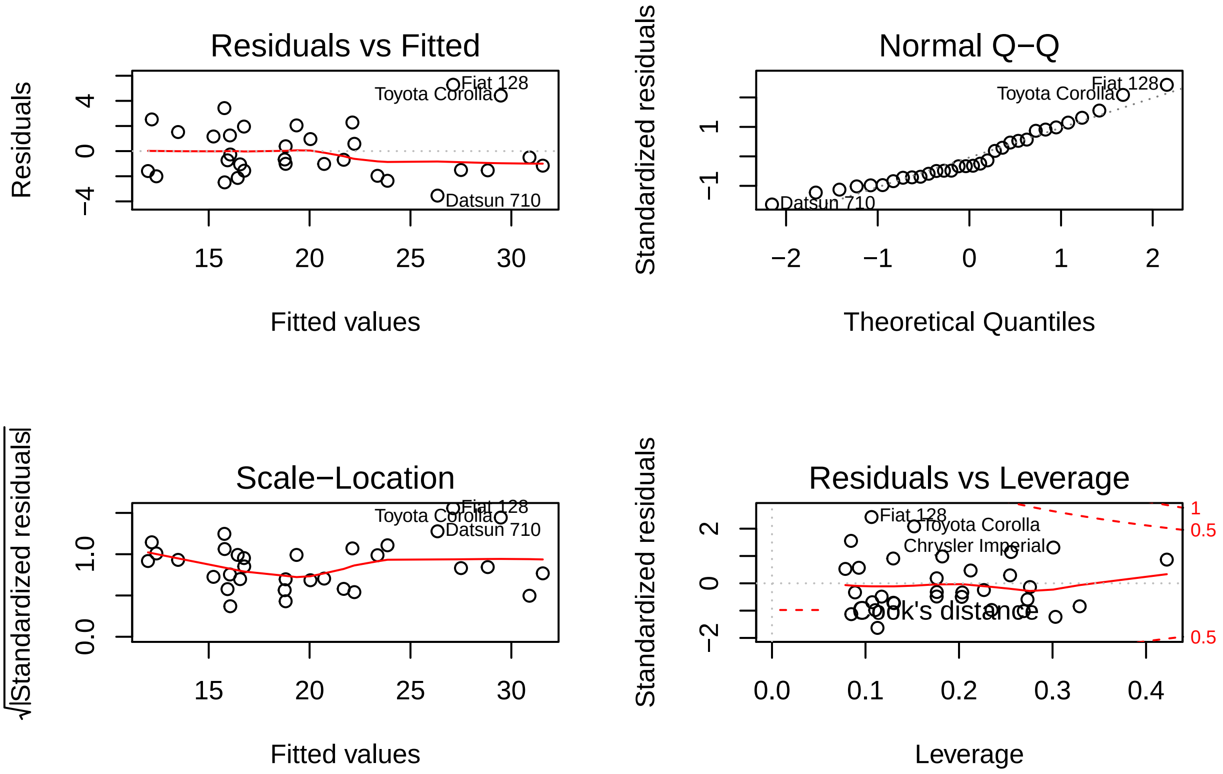
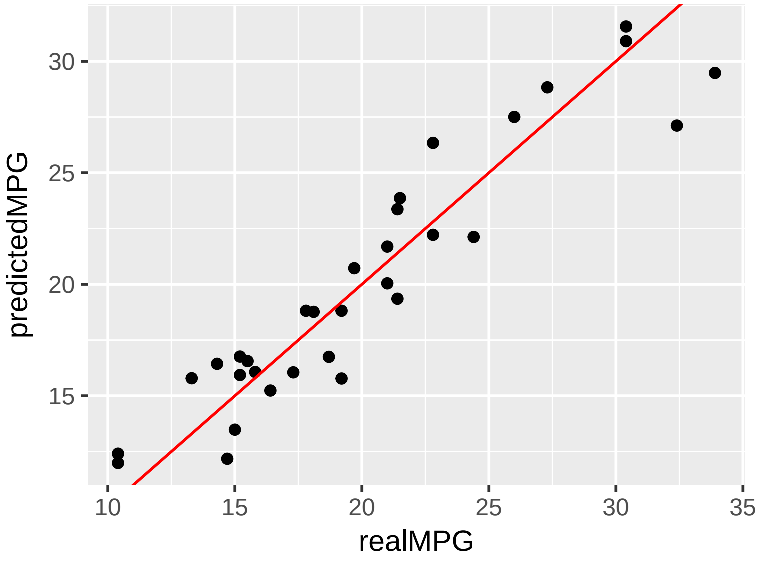


Figure 4: Prediction *vs.* Data



The ideal linear relation *predictedMPG=realMPG* is plotted as reference.