Regression Models Course Project

Vitaliy Baklikov

February 21, 2015

Executive Summary

Motor Trend magazine is interested in exploring the relationship between types of transmission and fuel consumption. They are interested in finding out what is better for MPG - manual or automatic transmissions. In the analysis that follows, we exlore the `mtcars` dataset - a collection of observations of various types of cars, their weight, mpg, and other attributes. We will try to use regression model for our analysis to try to identify whether a correlation exists between the type of transmission and MPG. We will also try to quantify the difference in MPG for both types of transmissions.

DataSet

`mtcars` dataset contains 32 observations on 11 aspects of automobile design and performance.

```
##                    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
```

`am` column represents the type of transmission. For clarity, let's factorize the variables and assign a meaningful name to each transmission.

```
mtcars$cyl <- factor(mtcars$cyl);mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear);mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am)
levels(mtcars$am) <- c("automatic", "manual")
```

Analysis

Let's plot the relationship between transmission type and MPG. See Figure 1 in Appendix. The plot suggests that there is a significant increase in MPG for manual types of cars.

Let's make our null hypothesis such that both manual and automatic types are from the same population and other variables have no effect on MPG vs type of transmission. We'll also assume that mileage data has normal distribution, so we will perform the two sided T-test.

```
result <- t.test(mpg ~ am, data=mtcars)
#result
```

Our T-test clearly demonstrates a p-value of `result$p.value`=0.0014, which allows us to reject our null hypothesis and conclude that manual and automatic types are from different populations.

Let's explore this dataset a bit further and check if other variables might have an effect on MPG. See Figure 2 in Appendix. From this pairs plot, we observe that variables `wt`, `cyl`, `qsec`, and others are correlated with `mpg`.

Regression

Let's fit a simple linear regression model of MPG using all variables in the dataset and analyse its accuracy.

```
lm.fitAll <- lm( mpg ~ ., data=mtcars)
#summary(lm.fitAll)
```

We observe that such model yields Adjusted R-squared value of 0.779, which only explains 78% in variance of MPG. Also, in this model none of the variables are at p-value of 0.005 and thus are statistically insignificant.

Let's find a better model by excluding some variables. We'll rely on stepwise algorithm by calling `step` function to suggest which variables to include in our model.

```
lm.fitSome <- step(lm.fitAll)
#summary(lm.fitSome)
```

The backward stepwise search suggests `mpg ~ cyl + hp + wt + am` formula for our model. So number of cylinders, horse power, and weight are confounding variables that have an effect on MPG per type of transmission. Such model has an adjusted R-squared value of 0.8401. Such model is a better fit model than a general model with all variables, however 84% variation in MPG is still far from ideal. Perhaps some variables have **interaction** between them. From the Figure 2, we notice that MPG has some correlation to Weight. Let's build a scatter plot and "zoom in" into that relationship. According to Figure 3, there appears to be some interaction between these two variables, so let's try to fit a better model and include this interaction in the model

```
lm.fitWithInteraction <- lm(mpg ~ cyl + hp + wt + am + wt:am, data=mtcars)
#summary(lm.fitWithInteraction)
```

Our new model yields an adjusted R-squared value of 0.8563. Such model is better than previous two models and explains 86% variation. However, we also observe that not all variables are at 0.05 significant level, so this model still has flaws.

Let's tweak our stepwise algorithm and introduce a penalty measure. Please refer to `?step` for details.

```
#k=2 gives genuine AIC: k=log(n) is sometime referred to as BIC or SBC
lm.fitAdjusted <- step(lm.fitAll, k=log(nrow(mtcars)))
#summary(lm.fitAdjusted)
```

The suggested formula for the model is `mpg ~ wt + qsec + am`. This model yeilds an Adjusted R-squared value of 0.8336, with all variable at p-value 0.05. We already know that there appears to be interaction between `wt` and `am` variables, so let's adjust our model for this interaction

```
lm.fitAdjustedForInteraction <- lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
#summary(lm.fitAdjustedForInteraction)
```

We observe that such model yields Adjusted R-squared of 0.8804 with all variables at 0.05 significance. Please refer to Figure 4 for Residuals plot and analysis.

That is the best fitted model out of all attempted.

Conclusion

We can conclude that based on the 5 models above, the last one explains 88% variance in MPG and thus suggests that when `wt` and `qsec` are fixed, MPG increases in average by `lm.fitAdjustedForInteraction$coef[4] + lm.fitAdjustedForInteraction$coef[2]*wt` for manual type cars.

Appendix

Figure 1. MPG vs Transmission Type

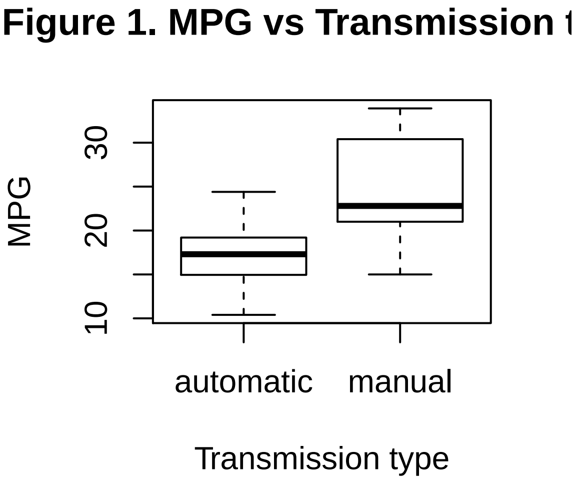


Figure 2. Pairs of variables

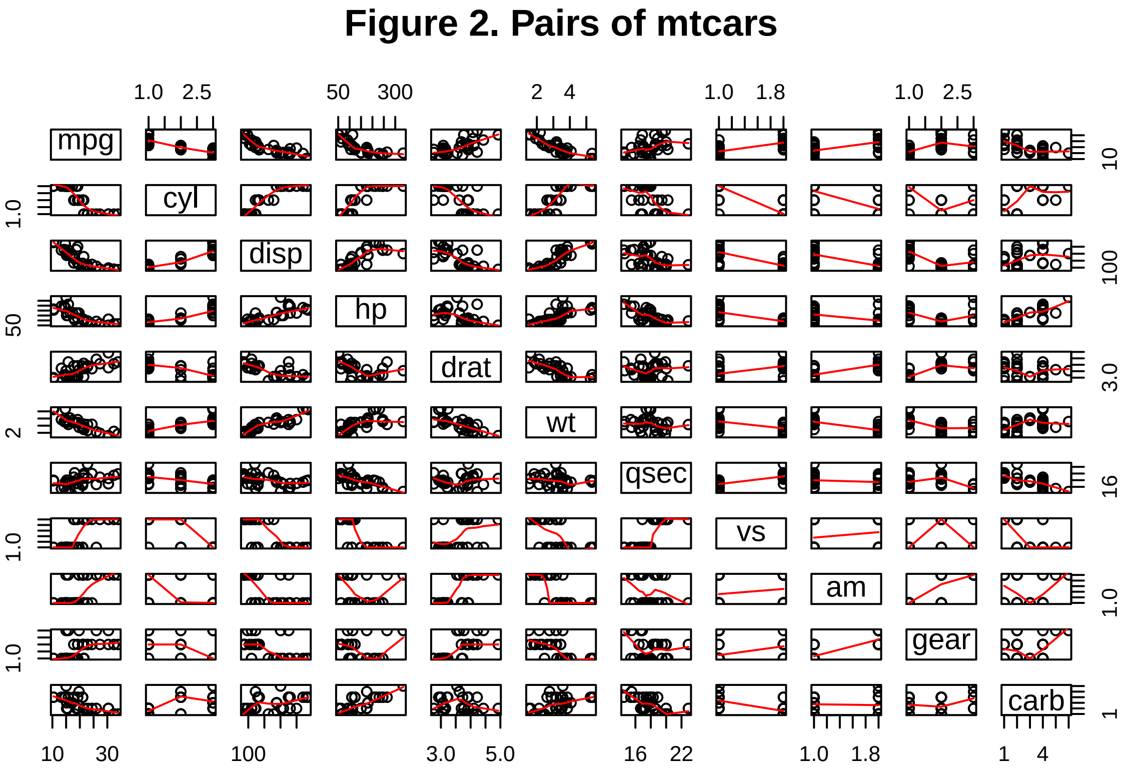


Figure 3. MPG vs Weight

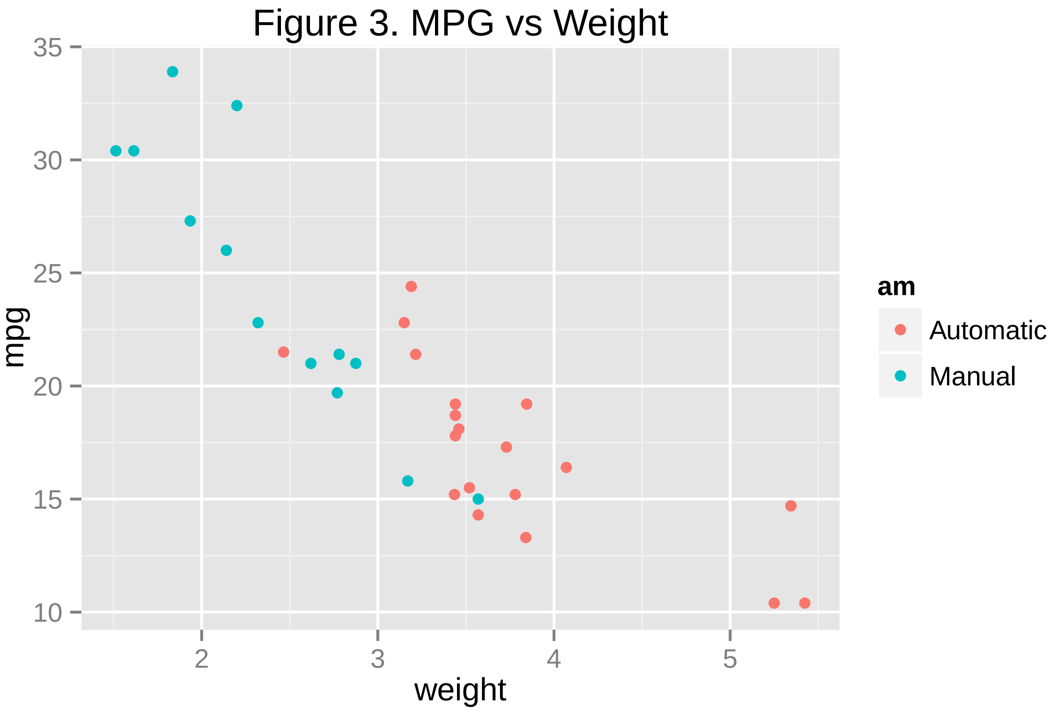
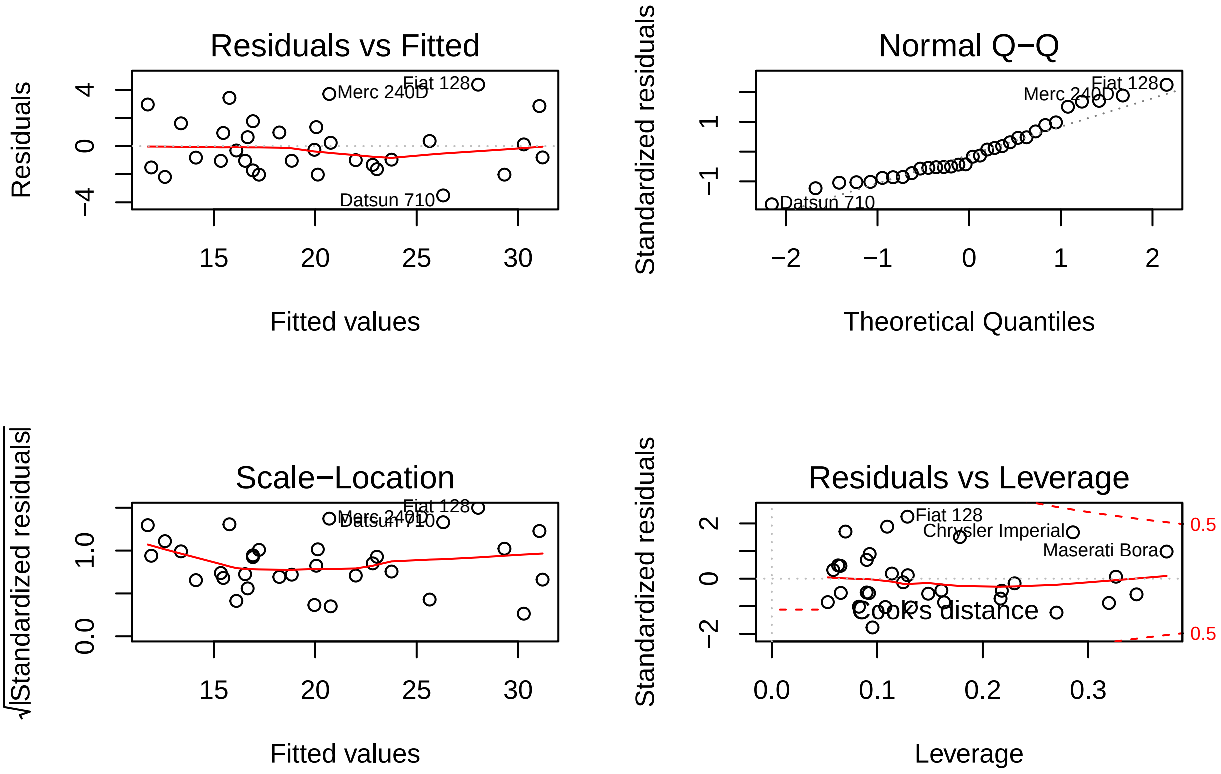


Figure 4. Residuals plot



Residuals vs Fitted
Residuals
4
0
−4
Merc 240D
Fiat 128
Datsun 710
15
20
25
30
Fitted values
Normal Q−Q
Standardized residuals
1
−1
Merc 240D
Fiat 128
Datsun 710
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
1.0
0.0
Merc 240D
Datsun 710
Fiat 128
15
20
25
30
Fitted values
Residuals vs Leverage
Standardized residuals
2
0
−2
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5
0.0
0.1
0.2
0.3
Leverage