Regression Modeling Project

Rabindra Swamidasan

Sunday, May 24, 2015

Motor Trend Data Mileage Analysis

1. Executive Summary

We analyse data extracted from the 1974 Motor Trend US magazine consisting of fuel consumption(mpg) and 10 other parameters for 32 cars. The analysis focuses on the question whether automatic or manual transmissions are better for mpg and to what quantifiable extent. To this end, we develop a linear regression model to explain the variation in the mileage of the cars in the data set based on the other 10 parameters.

The short answer turns out to be: It depends ... on the weight of the car. All other factors being the same, the expected mileage (in mpg) of a car with manual transmission exceeds that of a car with automatic transmission by 14.08 - 4.14 x the weight of the car in 1,000 lbs. In other words, below approximately 3,400 lbs, cars with manual transmission are more fuel efficient than automatics, all other factors being the same. Above 3,400 lbs, automatics are less thirsty.

2. Exploratory Data Analysis

A brief look at the data set shows the variable names and some values. Transmission is denoted by **am**, where 0 = automatic, 1 = manual.

```
##                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
```

To narrow our exploration of the data in line with our goal, we recall from elementary physics that the kinetic energy of a body is proportional to its mass. Therefore, the variable **wt** must be an important element affecting **mpg**. To examine this relationship, we plot **mpg** vs **wt**, color coded for **am**, our primary variable of interest, with separate fitted lines for Automatic and Manual transmissions.

Fig 1 in the Appendix shows this plot. We see from Fig 1 that the lines for Automatic and Manual transmissions have different slopes, and intersect near the center of the plot. This indicates a dependence of **mpg** on an interaction between **wt** and **am**. If there were no interaction, these lines would be parallel. Because they intersect near the center of the plot, there is a good chance that manuals may have better mileage in some weight range, but lower mileage than automatics in some other range.

3. Model Selection Strategy

Our model selection follows the iterative methodology below.

1. Start with the model we would like to test. This includes:
    - A regressor, **wt**, that we know we should include – a ***known known***.
    - A regressor, **am**, that we would like to include – a ***known unknown***.
    - The interaction term, **wt** * **am**, on the basis of the evidence in Fig 1.
2. Verify from the fit that the p-values for all regressor terms are below 1%. If not, discard the offending regressor.
3. Build the next model:
    - Find the correlation of the remaining covariates to the residual variance of the current model.
    - Choose the covariate with the highest correlation if the |correlation| >= 0.50.
    - If the covariates all have |correlation| < 0.50, we do not attempt further improvement. Go to Step 5.
4. Fit the linear model with the new regressor, go to step 2 to check p-values.
5. Perform an ANOVA test to compare the nested models.

4. Model Development

Execution of the strategy detailed above results in a rapid convergence.

1. **Model 1**: Regressors – **wt** + **am** + **wt** * **am**; p-values for all regressors were less than 1%; Multiple R-squared = 0.833, Adjusted R-squared = 0.8151. **qsec** is the only covariate with $|r| > 0.50$, with r = 0.51. So, we include it in Model 2.
2. **Model 2**: Regressors – **wt** + **am** + **wt** * **am** + **qsec**; p-values for all regressors were less than 1%; Multiple R-squared = 0.8959, Adjusted R-squared = 0.8804. The covariate with the highest correlation is **gear** with r = 0.09. So, we stop.

The summary results of Model 2 and the ANOVA results are below. A plot of the results of Model 2 are in Fig 2 in the Appendix.

```
##
## Call:
## lm(formula = mpg ~ wt + am + qsec + wt:am)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## wt            -2.937      0.666  -4.409 0.000149 ***
## am            14.079      3.435   4.099 0.000341 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## wt:am         -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt * am
## Model 2: mpg ~ wt + am + qsec + wt:am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     28 188.01
## 2     27 117.28  1    70.731 16.284 0.000403 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The low p-values of the regressors (all $<< 1\%$) and their t-statistics indicate that there is a low probabilty that any of them are superfluous. The low ANOVA p-value indicates that the regressor added in Model 2 contributes to the explanatory power of the model. Fig 2 shows that the the residuals are randomly and homoskedastically distributed about the fitted line. The Q-Q plot shows that the residuals have a close to normal distribution. The Leverage plot indicates that all residuals are within a Cook's distance of 0.5.

5. Conclusions

Using the coefficients from the results above, the final model is formalized in the equation below, which explains 89.59% of the variation in **mpg**. An error term is included to account for the residuals.

mpg = 9.72 - 2.94 * **wt** + 14.08 * **am** - 4.14 * **wt** * **am** + 1.02 * **qsec** + ϵ

By setting **am** = 1 (for manual) and **am** = 0 (for automatic) in this equation and taking the difference, we find that **mpg** for manual cars is greater than **mpg** for automatics by 14.08 - 4.14 * **wt**. This difference is 0 at the crossover point: **wt** = (14.08 / 4.14) x 1000 lbs = 3399.7 lbs. Above this weight automatics are more fuel efficient than manual transmissions.

6. Appendix

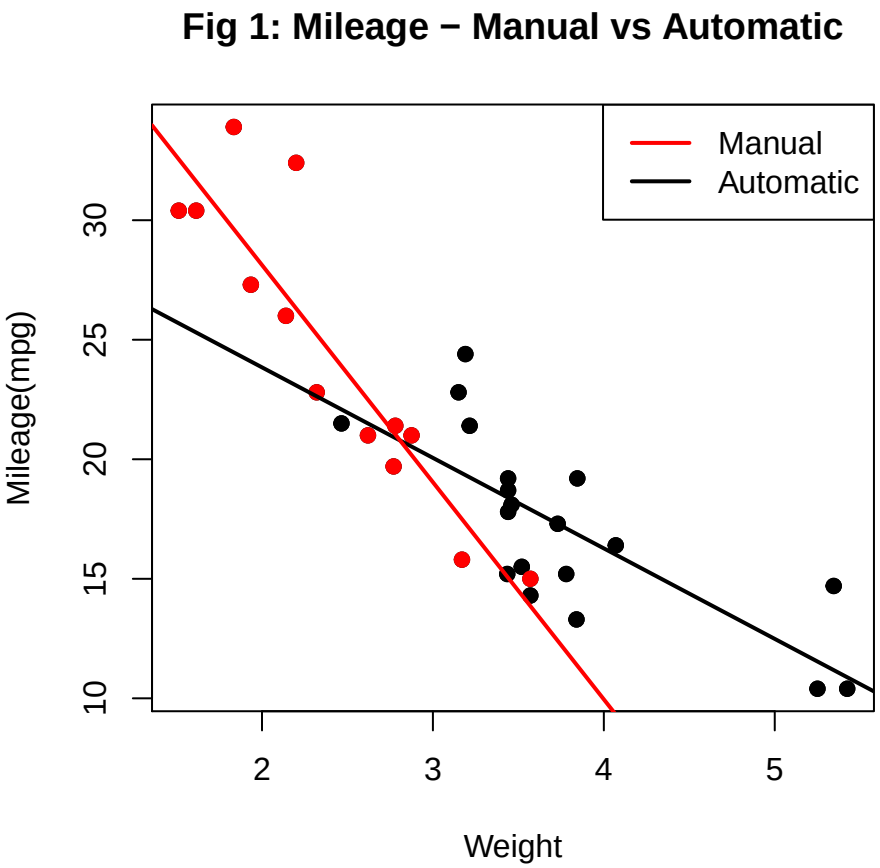

Fig 1 shows the clear interaction between weight (**wt**) and transmission type (**am**) in affecting mileage. If this interaction was not present, the red (manual) and black (automatic) fitted lines would have been parallel. For example, at a weight of 2000 lbs, manual has better mileage than automatic. But, at a weight of 3500 lbs, automatic has better mileage. Without interaction, the difference between manual and automatic would have been the same at 2000 lbs and 3500 lbs (and elsewhere), i.e. the lines would be parallel.

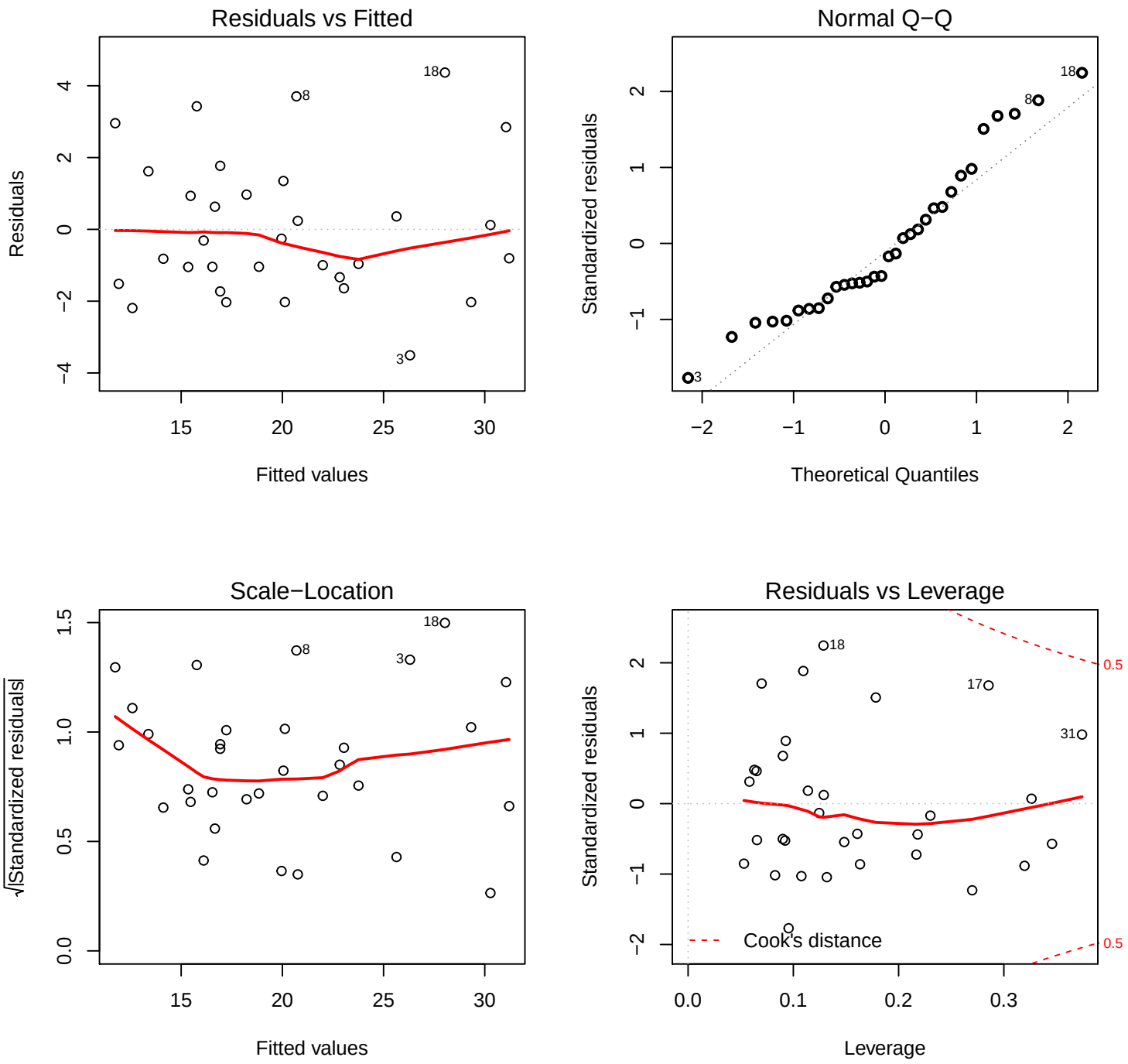

Fig 2: Residual plot of Model 2

NOTE: This report was authored in R Markdown and compiled to pdf using pdflatex (via knitr). To view the raw source, please visit the GitHub repo associated with this project.