

Regression Models Course Project

Technophobe01

2015-05-20

1 Executive Summary - 1974 Motor Trend Data in Review

This report was requested to enable a review of the 1974 magazine data on past car mpg performance and changes in vehicle efficiency. It address to core quesitons:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions.

Based upon our analysis we conclude that in 1974:

1. ***Manual tranmisssions were better for MPG.***
2. Transmission, weight and quarter mile time significantly influenced mpg efficiency.
3. Manual transmissions were on average ***7.24 miles per gallon*** more efficient that automatics.
4. For the same weight and quarter mile time, manual transmission cars were between **1.41** and **2.23** miles per gallon more efficient than an automatic transmission (Depending on whether we account for transmission and weight interaction in the linear model).

Conclusion: In 1974, cars that were lighter in weight with a manual transmission and cars that were heavier in weight with an automatic transmission had higher MPG values. See: Appendix A, Figure 2.

Note: The report source code is available via [GitHub](#) for deeper review.

1.1 Data Set Description

The Motor Trend magazine data was extracted in 1974, it coverered gasoline mileage in miles per gallon (MPG), and ten aspects of vehicle design and performance for 32 vehicles (1973-74 models), the sample includes 32 vehicles and has a bias to exotic, non-US., automobiles: it includes seven Mercedes, a Porsche, a Ferrari, a Maserati and offers us us an amazing view of vehicle efficiency and performance from 1974.

1.2 Basic exploratory data analyses

Our initial goal was to test the hypothesis that cars in 1974 with an automatic transmission used more fuel than cars with a manual transmission. Whether cars have an automatic or manual transmission is found in the `am` (**Auto/Manual**) column of the `mtcars` dataset.

1.2.1 Is an automatic or manual transmission better for MPG? Answer: *"Manual"*

The box plot (Appendix A, Figure 1.) provides a simple yet clear indication that automatic cars from the 1974 data set had lower mpg efficiency (**Automatic Transmission Median: 17.15 mpg**) than the manual cars reviewed (**Manual Transmission Median: 24.39 mpg**). Based on figure 1 we appear to have a clear hypothesis that automatic cars had a lower miles per gallon, and therefore a lower fuel efficiency, than manual cars did. i.e (24.39mpg - 17.15mpg creates a 7.24mpg manual transmission advantage. *The median is depicted as a thick line in figure 1, the edges of the box show the 25th and 75th quantiles of the data respectively.*

Did lower automatic transmission mpg efficiency occur in the data by random chance? To try to confirm, we performed a `t.test()` - `t.test(mpg ~ am, data=mtcars)`. Our goal being to determine if MPG results relied on having an automatic or manual transmission. In essence, we wanted to try to refute that random chance created the inference depicted in figure 1.

	P.Value	Confidence	Lower.Bound	Upper.Bound
MPG ~ Transmission	0.001374	95%	-11.28	-3.21

The t-test **p-value** result of **0.0013736** in combination with a confidence value of **95%**, indicate that we can be confident that the probability of chance is low. We conclude that automatic transmissions **did** have lower **MPG** in 1974 than manuals, what we were unable to determine from this initial testing is what exactly was causing the correlation. We checked to see if automobile weight in 1974 **correlated** with **MPG** efficiency. **Appendix A, Figure 2** shows that in 1974 automobile weight did correlate with MPG for both manual and automatic cars. Here we used a linear model to depict the manual and automatic trend lines.

1.2.2 Answer: Quantify the MPG difference between automatic and manual transmissions.

Appendix A, Figure 3 shows our final exploratory chart which indicates visually that weight is a key driving factor. However, the predictors (Weight, Cylinder Size, and Displacement) could potentially be confounded. To determine which predictors were significant, and to select the preferred model to use, we performed an Akaike information criterion (AIC) model selection (*Ref: The R Book, Chapter 9.17*) against the baseline linear model, the smaller the AIC, the better the fit.

```
baseline <- lm(mpg ~ ., data=mtcars)
prefferedModel <- step(baseline, k=log(nrow(mtcars)))
```

we reviewed the `prefferedModel` results and picked **weight**, **quarter mile** and **transmission** as the predictors of most value based on the lowest **AIC of 67.17**. Our final check was to see if the predictors interact. We checked this with `aov` and noted that `wt` and `am` do appear to interact (**Sum Sq 52.0**), hence we specified to `lm()` that weight and transmission interact `wt:am`

```
selectedModel <- aov(mpg ~ wt*qsec*am, data=mtcars)
summary(selectedModel)
finalModel <- lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
summary(finalModel)$coef
```

Thus, the result shows that when `wt` (**weight/1000lbs**) and `qsec` (**1/4 mile time**) are fixed we can calculate the manual transmission advantage over an equivalent automatic transmission. Without interaction we see a **1.41mpg** advantage, if we remove interaction of weight and transmission we see a **2.23mpg** advantage.

```
newcar = data.frame(wt=2.5, qsec=16.5, am=1)
predict(finalModel, newcar)
```

That is, a manual transmission car that weighs 2500 lbs has, with a `qsec` of 16.5 has a predicted mpg of **22.9mpg**. We can also use the model to predict the mpg of all the vehicles in `mtcars`.

```
predict(finalModel)
```

1.3 Residual Analysis / Diagnostics

To wrap up, we ran a model fit, we can verify the following underlying assumptions:

1. The Residuals vs. Fitted plot shows no consistent pattern, in support of accuracy of the independence.
2. The Normal Q-Q plot indicates that the residuals are normally distributed to the line.
3. The Scale-Location plot confirms the constant variance assumption, points are randomly distributed.
4. The Residuals vs. Leverage implies no outliers are present, as all values fall well within the 0.5 bands.
5. We conclude that we do not have influential observations based on the `dfbetas` result of 0. `dfbetas` is a measure of influence of the observations on the regression coefficients. The thumb rule for the 'dfbetas' is that if their absolute value exceeds 1, the observations have significant influence on the covariates.

2 Appendix A

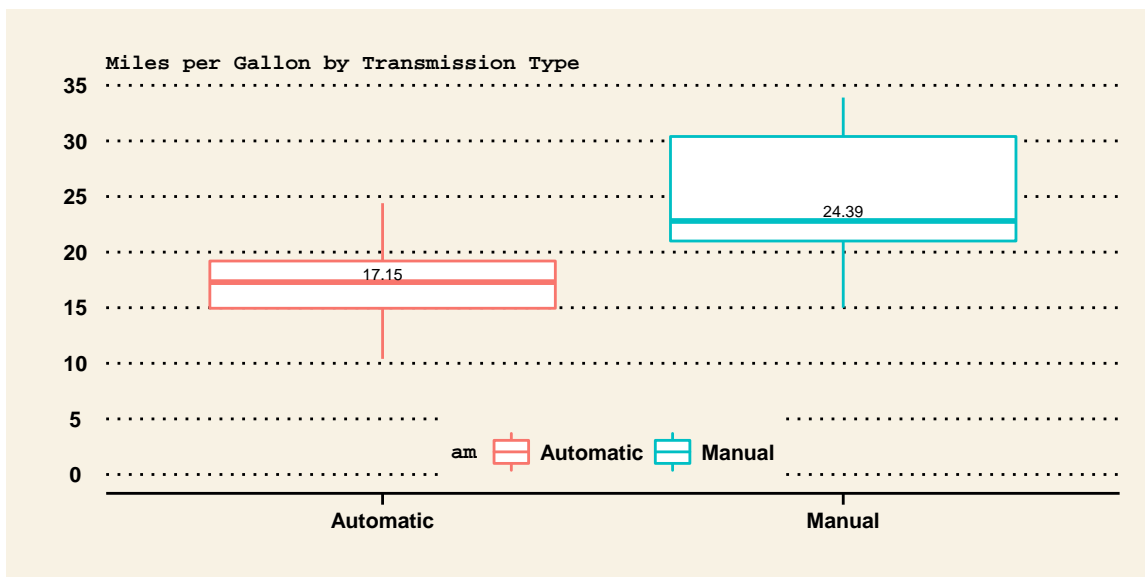


Figure 1: Exploratory BoxPlot

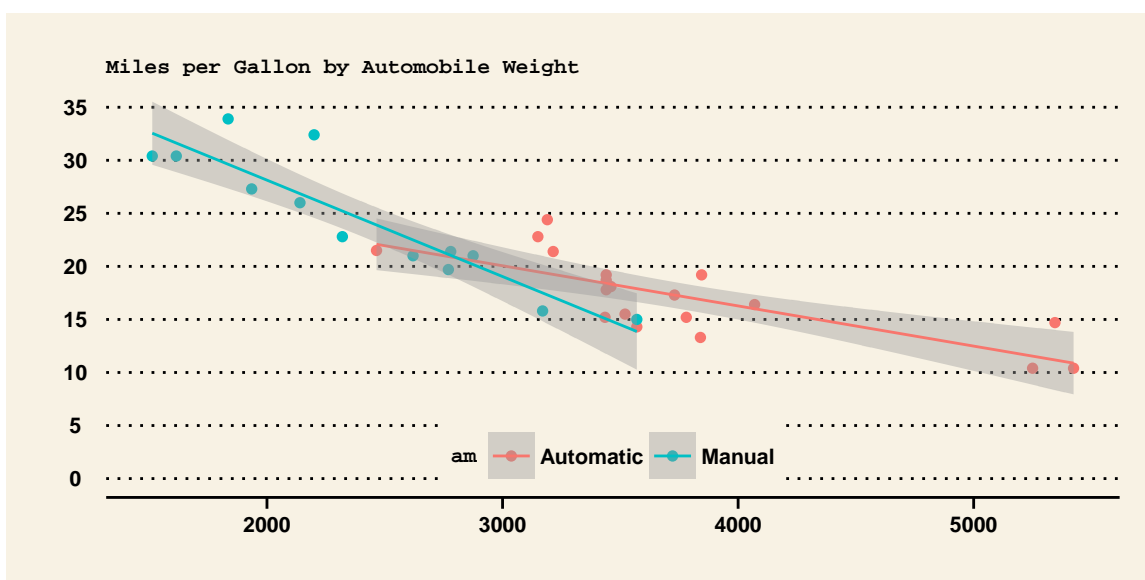


Figure 2: Exploratory Linear Plot

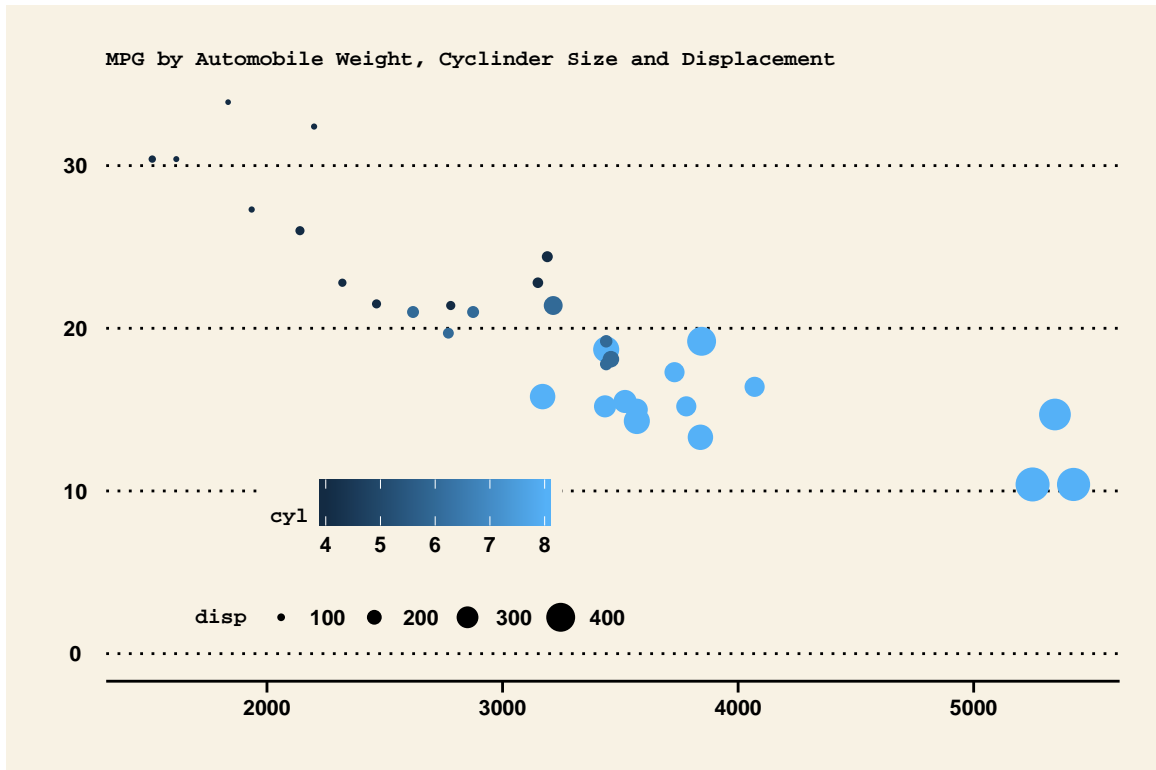


Figure 3: Exploratory Multiple Linear Regression Plot

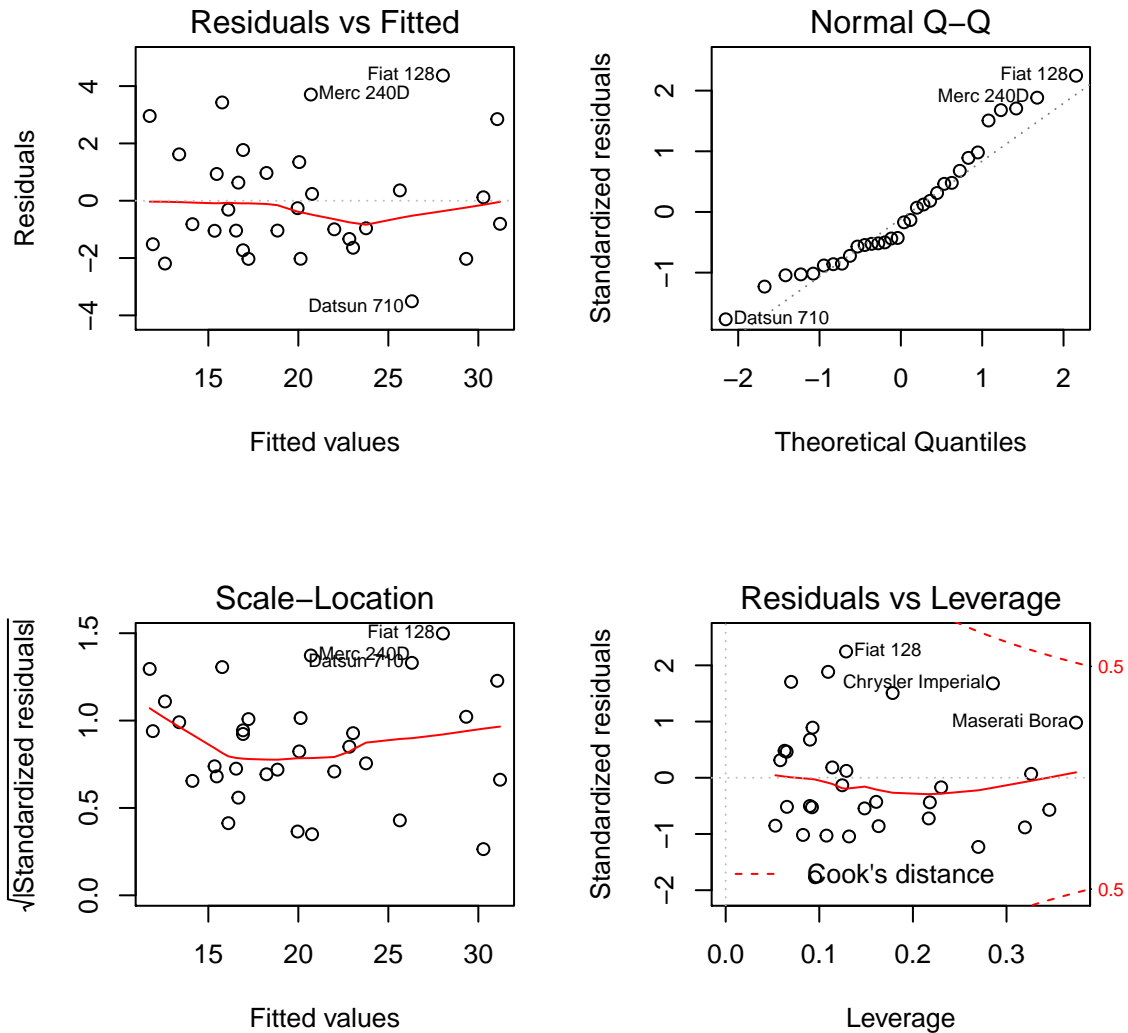


Figure 4: Exploratory Model Fit Plots

3 References

- *The R Book*
 - By: Michael J. Crawley, Publisher: John Wiley & Sons Pub. Date: December 26, 2012, eISBN: 978-1-118-44896-0
- *R in Action*
 - By: Robert I. Kabacoff Publisher: Manning Publications Pub. Date: August 24, 2011, ISBN-10: 1-935182-39-0
- *Mathematical Statistics with Resampling and R*
 - By: Laura Chihara; Tim Hesterberg Publisher: John Wiley & Sons Pub. Date: September 6, 2011 Print ISBN: 978-1-11-02985-5
- *Think Stats, 2nd Edition*
 - By: Allen B. Downey Publisher: O'Reilly Media, Inc. Pub. Date: October 28, 2014 Print ISBN-13: 978-1-4919-0733-7