Data Science Coursera: Regression Models Project

Simon Keith

2016-11-29

Contents

Overview

Note

This work was done for the "Regression Models" course project, as part of the Data Science specialization on Coursera.
Disclaimer: for better readability I chose to keep plots and explanatory text together instead of adding an appendix. However, keeping the code would make the report far too long. You can get the *.Rmd* file on the github repository.

Executive summary

In this document, we explore the relationship between a set of variables and miles per gallon (mpg), using the mtcars dataset. We are particularly interested in the two following questions:

- Which transmission (automatic or manual) is better for mpg?
- What is the mpg difference between automatic and manual transmissions?

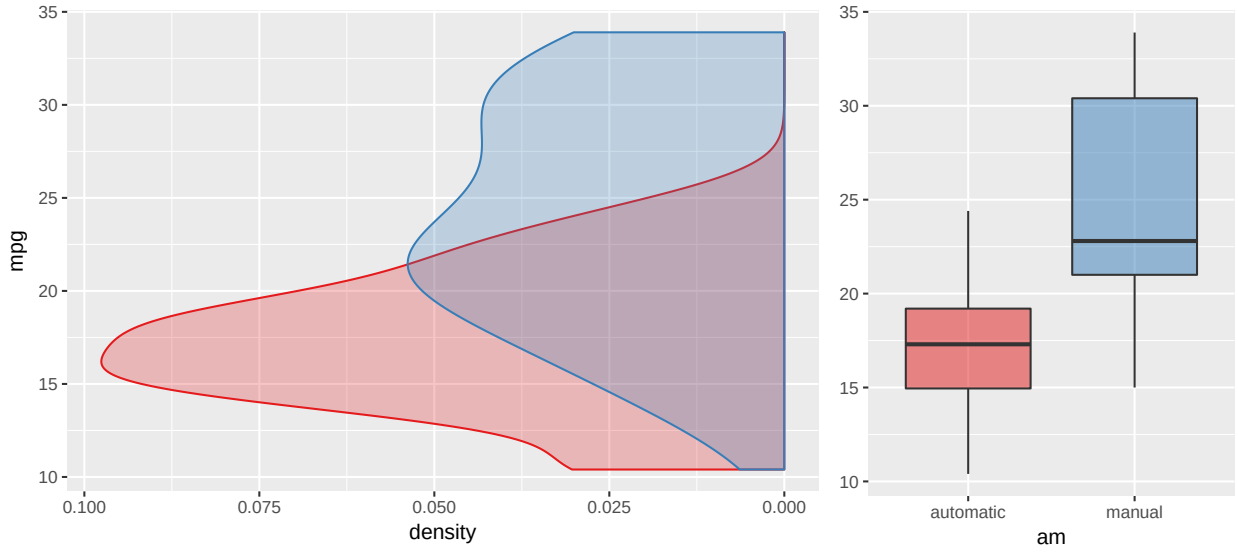
We use regression techniques in order to answer these questions.
We show that, at first glance, cars with manual **transmission** seem to have a better fuel efficiency (more **mpg**). However, when we apply the right transformations and select the right features, this relation seems to become insignificant. Instead, we choose a model explaining reduction in fuel efficiency by geometrical relations with **weight** and **gross horsepower**, both features having a negative impact on **mpg**.
Stepwise model selection by AIC was used to select the features, since many variables in the dataset are highly correlated.

Exploratory data analysis

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Exploring fuel efficiency (mpg) versus transmission (am)

Let's start by plotting **mpg** against the type of **transmission**. We first transform variables that have no more than 3 levels to factors. Also, for better readability we rename the levels of **transmission**.



This graph shows the density of **mpg** and a boxplot with **automatic transmissions** in red and **manual transmissions** in blue. We can already see that the medians are quite different for the two groups, and that the interquartile ranges don't overlap. From a visual point of view, it looks like automatic transmissions have a significantly lower fuel efficiency than manual transmissions. Let's see if we can confirm and quantify this with a regression model.

	Estimate	Std. Error	t value	Pr(>|t|)	
ammanual	7.245	1.764	4.106	0.000285	* * *
(Intercept)	17.15	1.125	15.25	1.134e-15	* * *

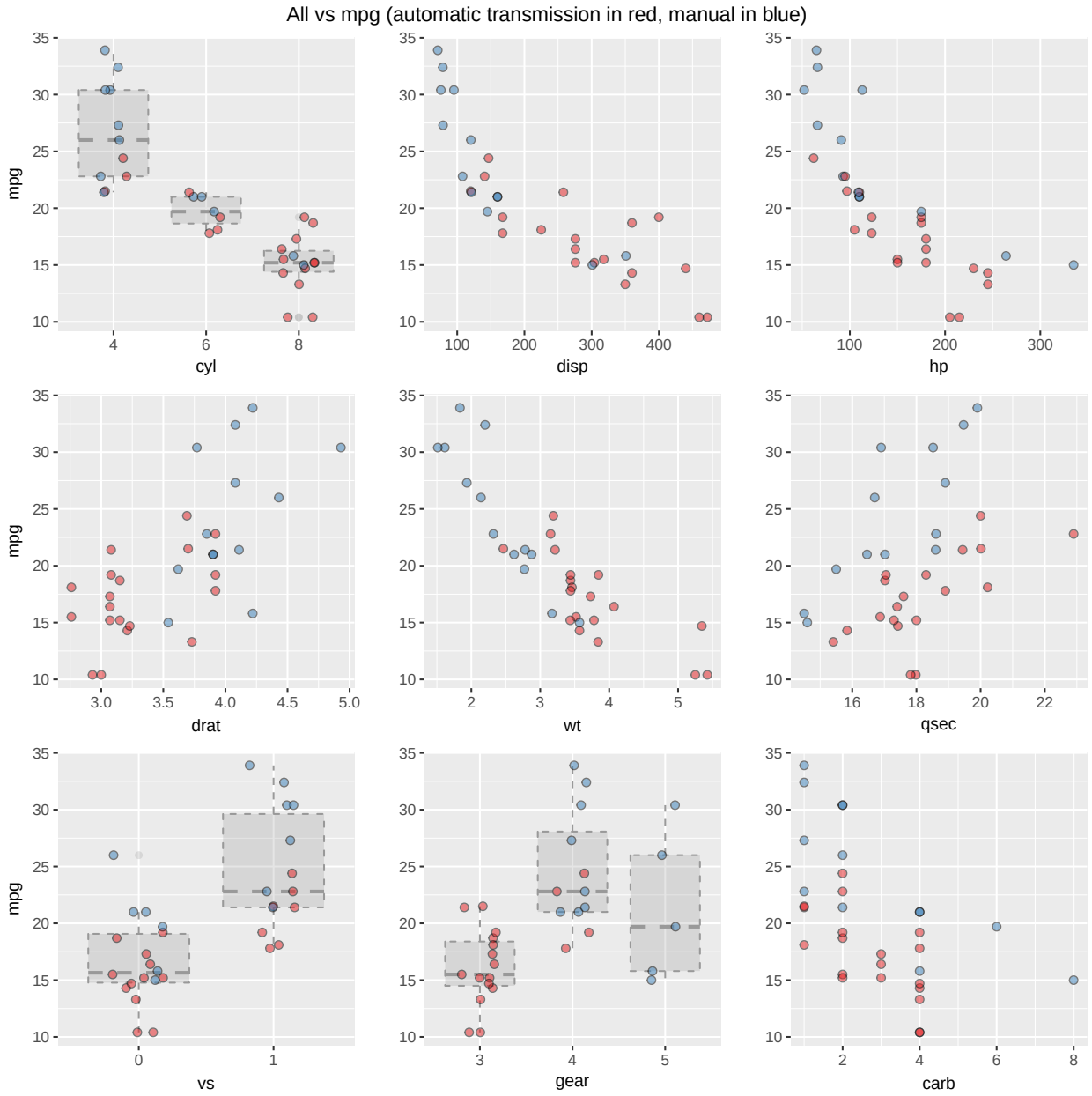
Table 2: mpg ~ am

Observations	Residual Std. Error	R^2	Adjusted R^2
32	4.902	0.3598	0.3385

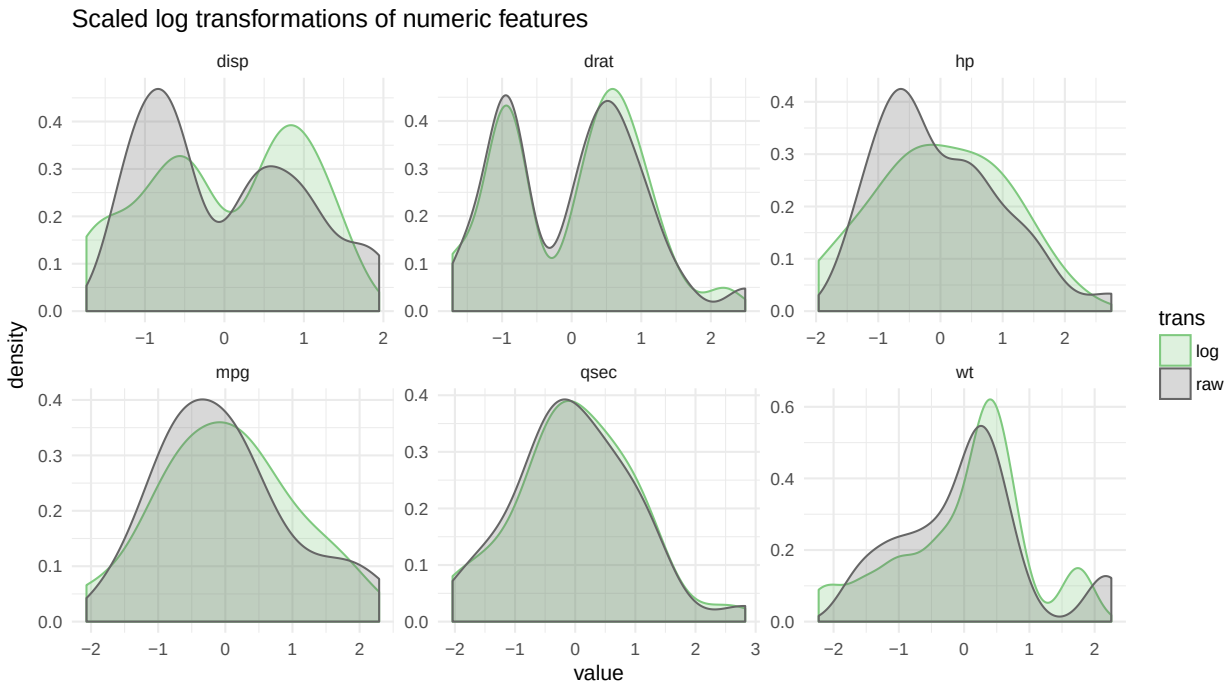
As we can see, there is an average **7.24** increase in fuel efficiency (**miles per gallon**) for **manual transmissions** as compared to **automatic transmissions**. A t-test confirms the significance of this difference in the mean efficiency of the two groups.

Exploring fuel efficiency (mpg) versus all features

Let's investigate the relation of **mpg** with other features. Below we plot **mpg** against all other features except **transmission** which is color coded (same colors as before).



We see some interesting things on these plots. First, we observe some obvious imbalance between the two types of **transmissions**. For example, cars with **automatic transmissions** are much heavier (**wt** feature) than their **manual** counterpart. We observe similar cases for **gear** , **disp** and **drat**.
Then, we see a clear pattern between **mpg** one one side and **disp** and **hp** on the other side, although the relations are not quite linear. It seems that a log transformation could make it more linear. Below, we take the logs of all numeric features (except **carb**) and then plot their density against a scaled axis.



Based on the shape of their distribution, **mpg** and **hp** seem to be good candidates for log transformation. We plot the transformed features below, with the same color code as before.

Finally, some variables, such as **disp**, **wt** or **hp**, seem to have a very similar relation with **mpg**. The table below show that several features are indeed highly correlated. This will be an issue for model selection.

var1	var2	corr
wt	disp	0.888
disp	cyl8	0.885
hp	cyl8	0.817
hp	disp	0.791
vs1	cyl8	-0.778
carb	hp	0.75

Final model

Feature selection

Based on the knowledge acquired from the exploratory data analysis, we transform **mpg** and **hp** to their log. Since the features are highly correlated, we also perform stepwise model selection by AIC on the resulting data.

	Estimate	Std. Error	t value	Pr(>|t|)	
wt	-0.1794	0.02742	-6.543	3.63e-07	* * *
log_hp	-0.2657	0.05646	-4.706	5.748e-05	* * *
(Intercept)	4.832	0.222	21.77	1.637e-19	* * *

Table 5: log_mpg ~ wt + log_hp

Observations	Residual Std. Error	R^2	Adjusted R^2
32	0.1043	0.8852	0.8773

The weight in 1000 lbs, **wt**, and the log of **hp** (gross horsepower) were selected. Both features have highly significant coefficients and we achieve a quite high **adjusted** R^2 of **0.8773**.
However, we note that **transmission** was not selected. Let's add it to the model just for reference:

	Estimate	Std. Error	t value	Pr(>|t|)	
wt	-0.1642	0.03778	-4.346	0.0001651	* * *
log_hp	-0.2762	0.05981	-4.618	7.869e-05	* * *
ammanual	0.03271	0.05513	0.5933	0.5578	
(Intercept)	4.821	0.2252	21.4	6.79e-19	* * *

Table 7: log_mpg ~ wt + log_hp + am

Observations	Residual Std. Error	R^2	Adjusted R^2
32	0.1055	0.8866	0.8745

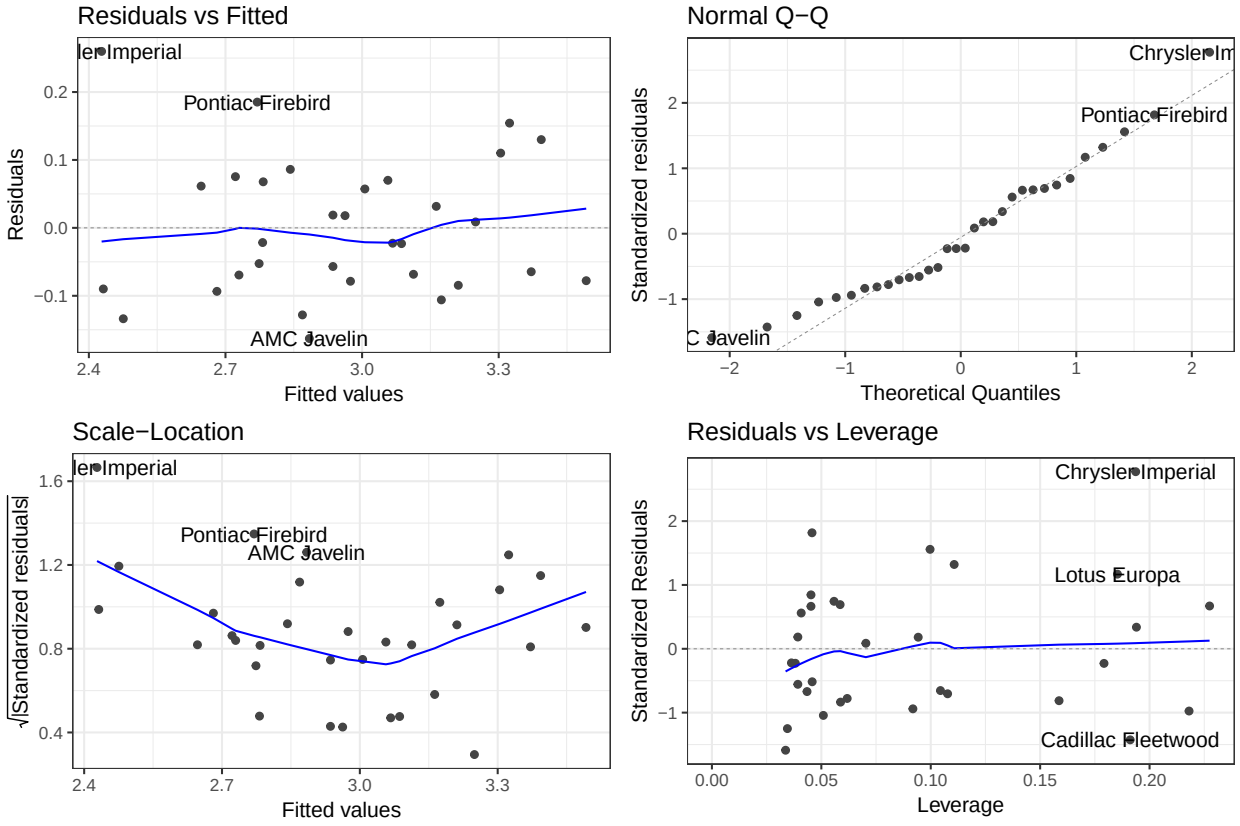
After transforming **mpg** and **hp** to their log, and accounting for **wt** and the log of **hp** in the model, **transmission** cease to have an impact on **mpg**. Indeed the estimate is not significantly different from zero and the **adjusted** R^2 dropped a little bit. Regarding the initial questions, we conclude that the type of **transmission** is not necessarily relevant when one wants to estimate the fuel efficiency of a car.

Interpretation

Since we made some transformations to the data, let's interpret the coefficients of our selected model. Holding all other variables:

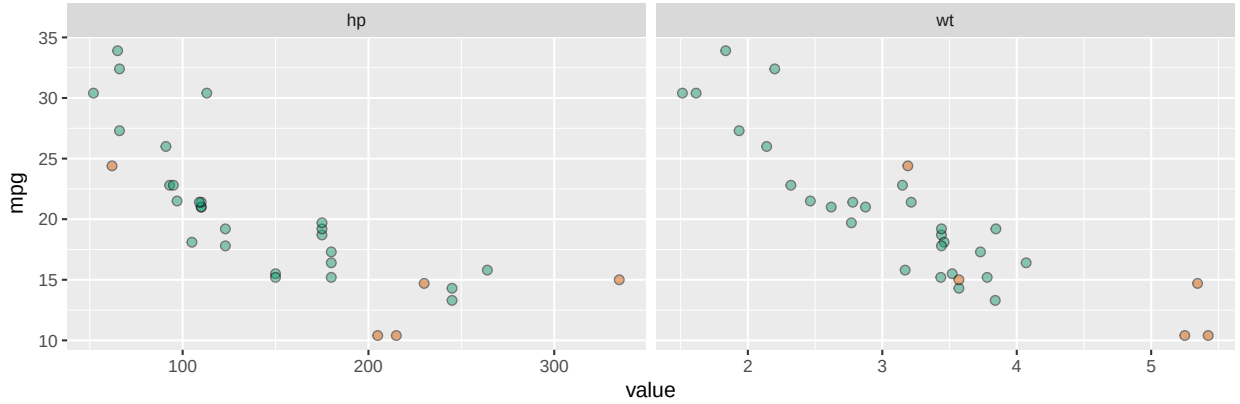
- **wt**: for a **1000 lbs increase** in weight, we expect to see a **17.94 % decrease** in **mpg**
- **hp**: for a **1 % increase** in **hp**, we expect to see a **26.57 % decrease** in **mpg**.

Residual diagnostics



From the residual diagnostics plots, we observe no particular pattern in the residuals versus fitted values. However, from the QQ-plot it seems that the residual is a little bit right skewed. The scale-location plot seem to show some heteroscedasticity, with wider spread residuals around the extremes of the fitted values. This might be due to the fact that we have less observations on the extremes, but also to the presence of some outliers, as we can see on the residual versus leverage plot.

Let's finish by checking hatvalues to identify the cars that have high leverage:



In this plot we show in orange the 5 cars with the highest hatvalues. Indeed, most of them have low **mpg** and high values for **wt** and **hp**.

Overall, we conclude that our model fits the data quite well. We may use this model for prediction as long as we pay attention to particularly low or high fitted values, as well as extreme values in the predictors.