Regression Models, final work

Alejandro Borges Sanchez

7/12/2021

Context

Its a journalistic endevour. We are to provide visualizations and insightful figures with statistically sound intervals on data that might explain MPG by a number of variables.

The questions

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

The data

32 cars from the mtcars dataset. I used that one because I think its what the course wants us to address on purpose: it is too small for any real inference, all the cars are old and a lot of confounders are just not included (city mpg is not the same as highway mpg, for one). The mpg dataset (type help("mpg") to see it), from the ggplot2 package, has much more comprehensive and modern data.

Having said that, car efficiency was not on my mind before attacking this problem It has taken me all week to reach the conclussion I have arrived to because, hell I knew nothing about cars and how they work. I now know a heck of a lot more. What follows is my "research" table of what each variable means and how I think they relate to each other. It really (really) helped:

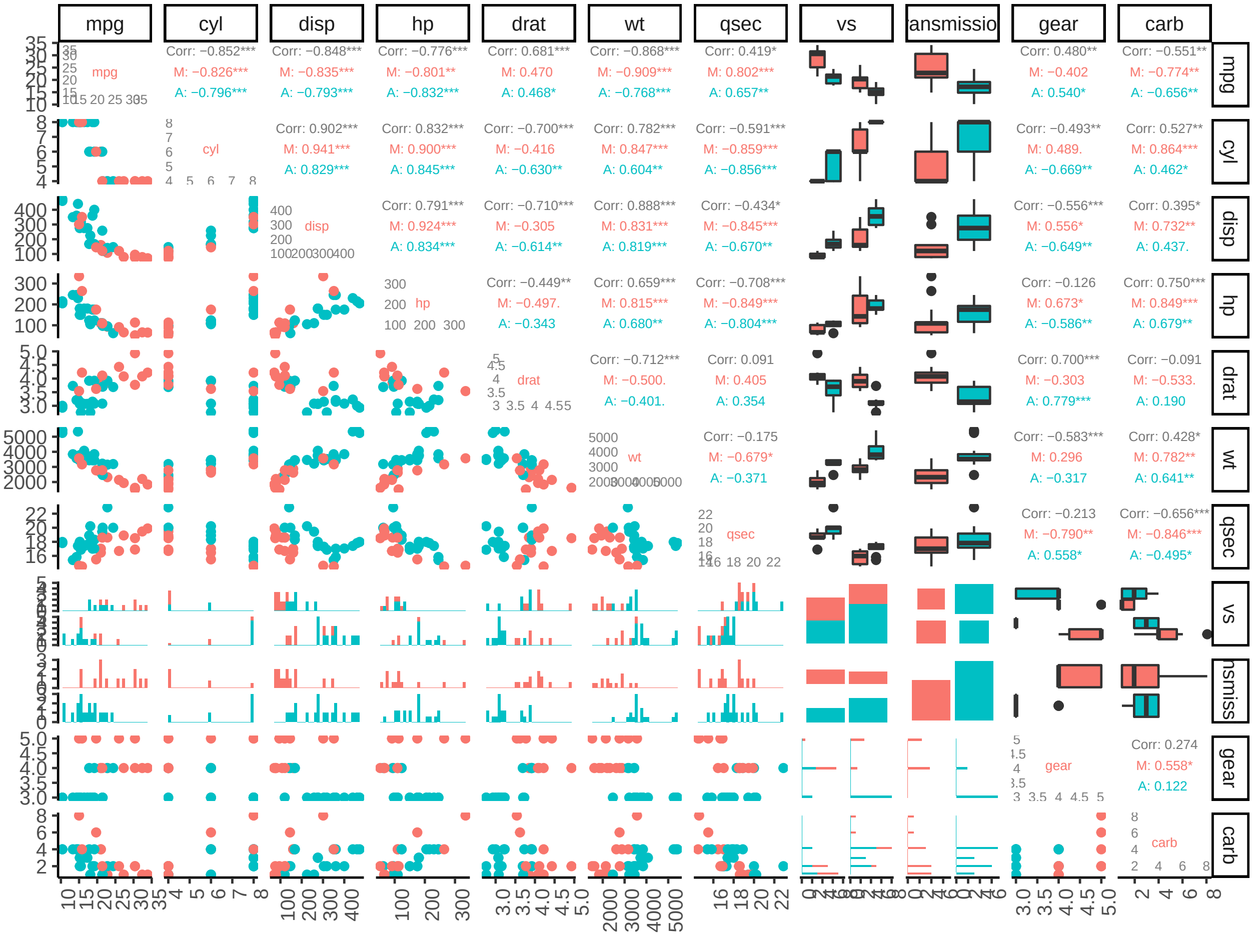
Var	Doc	Note	Source	f(?)
mpg	Miles/(US) gallon			
cyl	Number of cylinders			
disp	Displacement (cu.in.)	How much gas/air mix passes through the cylinders in total	https://www.yourmechanic.com/article/what-is-engine-displacement	f(cyl,volcyl)=cyl*VolCyl
hp	Gross horsepower		https://www.wikihow.com/Calculate-Horsepower	f(Torque,RPM)=T*RPM+e
drat	Rear axle ratio	Motor revs/rear axle revs (kind of a measure of efficiency)	https://www.autolist.com/guides/axle-ratio	f(RPMaxlefront,RPMaxleback)=RPM
wt	Weight (1000 lbs)			f(car)

Var	Doc	Note	Source	f(?)
qsec	1/4 mile time			
vs	Engine (0 = V-shaped, 1 = straight)	V: classic, can have more cylinders, Straight: more efficient, normally less cylinders	http://www.whyhighend.com/inline-vs-v-engine.html	f(cyl)
am	Transmission (0 = automatic, 1 = manual)			
gear	Number of forward gears			
carb	Number of carburetors			
TORQUE				F(force,distance)=F(disp,drat)+e

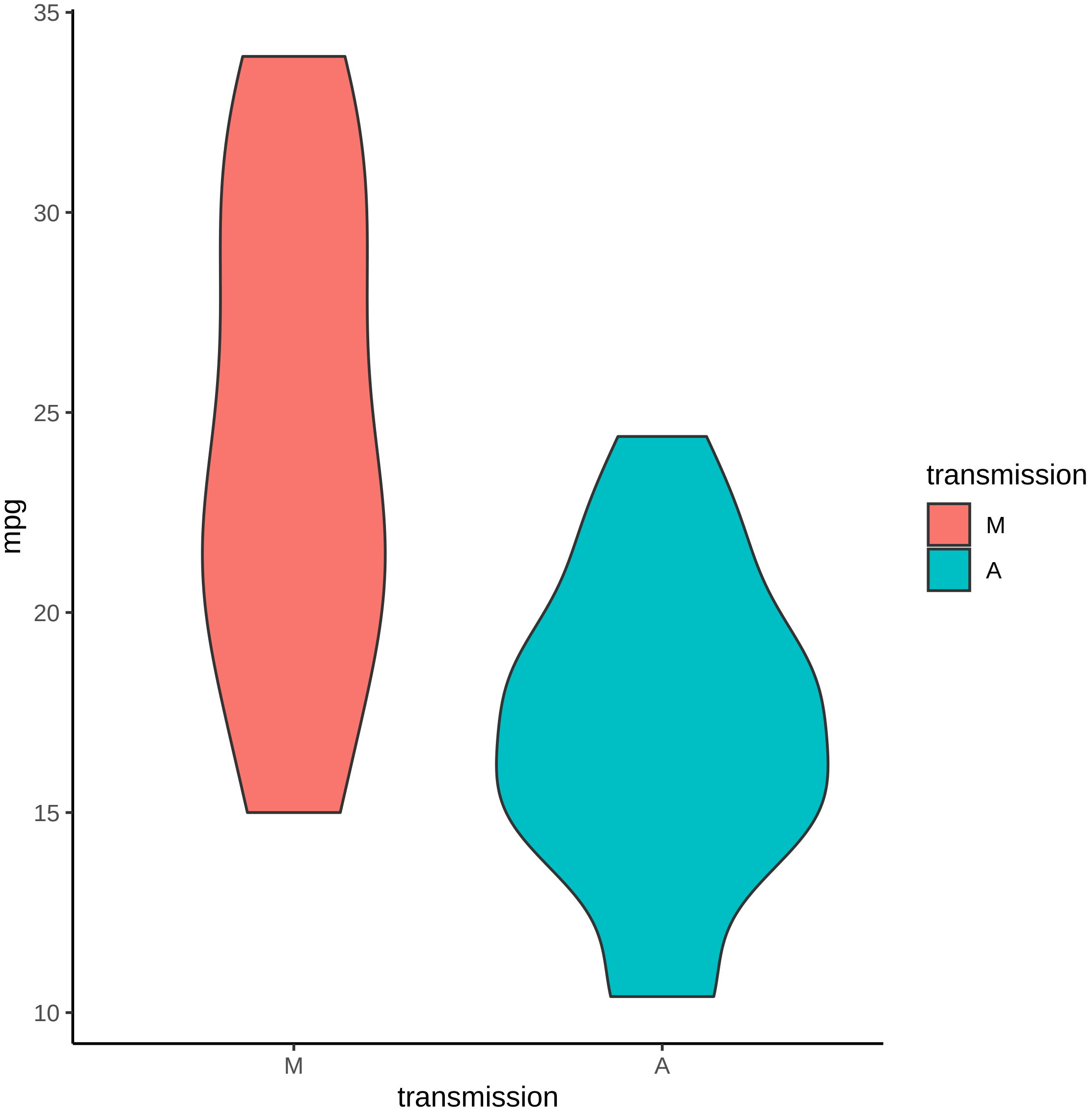
The way to the answer

We can address this problem with a linear regression model that has mpg as the dependent and at least the transmission (automatic or manual) as the independent.

The reference for all our work is still just a SPLOM, where we can see at a glance correlations between all variables and the independent variable, mpg:



So the base relationship and linear model would look like this:



The basic linear model would look like this (I use the broom package for summarizing and modeling, check the code out: its really cool in a tidy way):

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	24.392308	1.359578	17.941085	0.000000	21.61568	27.16894
transmissionA	-7.244939	1.764422	-4.106127	0.000285	-10.84837	-3.64151

So, as transmission is a binary factor with M=0 and A=1, we can report to our magazzine:

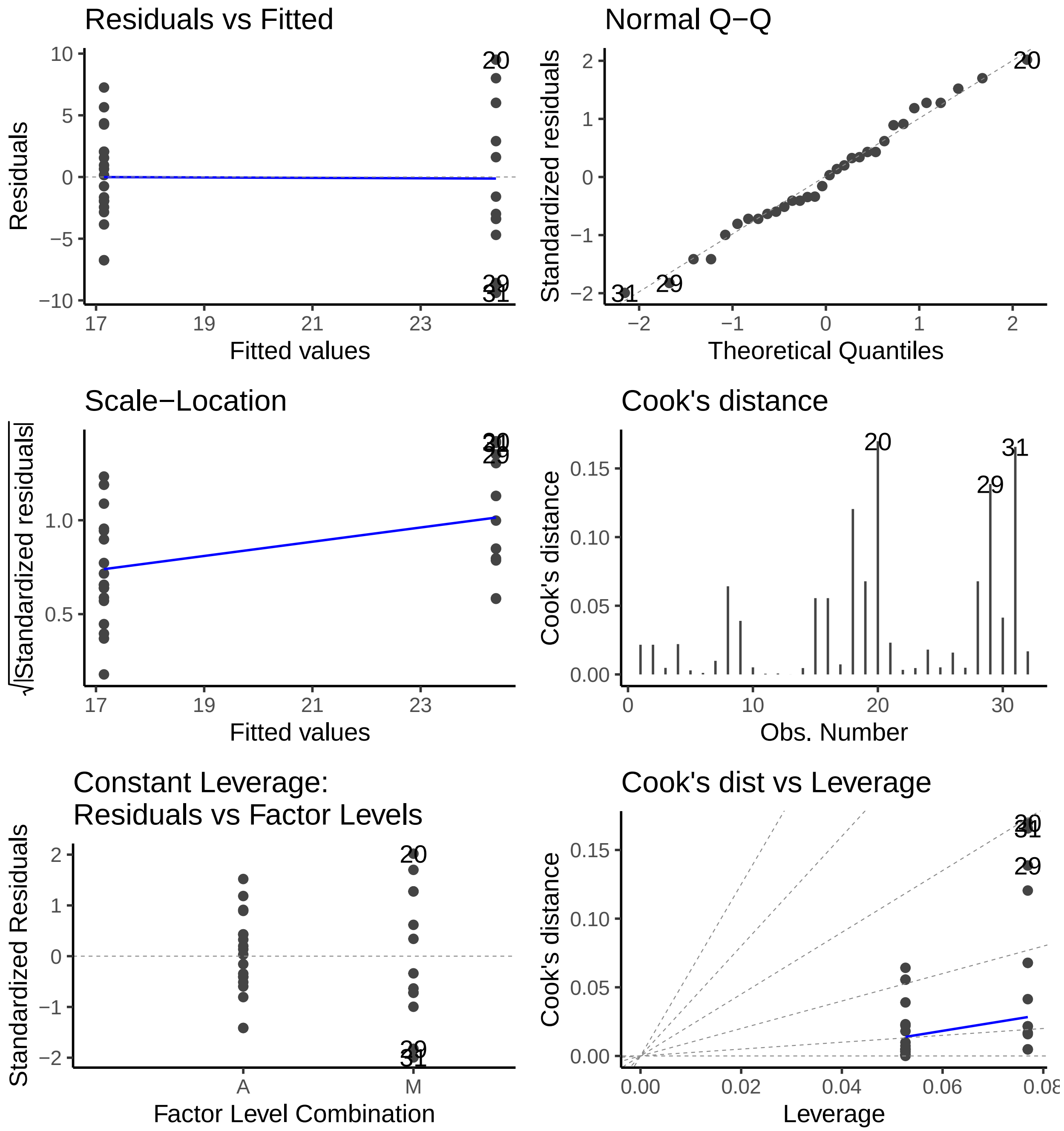
“Assuming the cars you gave us were picked at random and are representative of all the cars in the market, the average manual transmission cars consume from 21.6 to 27 MPG while the average automatic car consumes from 10.76 to 23.6 MPG”

This implies some cars, regardless of transmission, consume between 21.6 to 23.6 MPG. This rare cars are:

```
## # A tibble: 2 x 4
##   row_number Model         mpg transmission
##        <int> <chr>       <dbl> <fct>
## 1          3 Datsun 710  22.8 M
## 2          9 Merc 230    22.8 A
```

So yes, we can say from this data set “which is better on average, assuming this 32 cars are representative of the whole car market” and we can quantify it to a point (with the exception of this two cars…), but this is not, obviously, publication quality material.

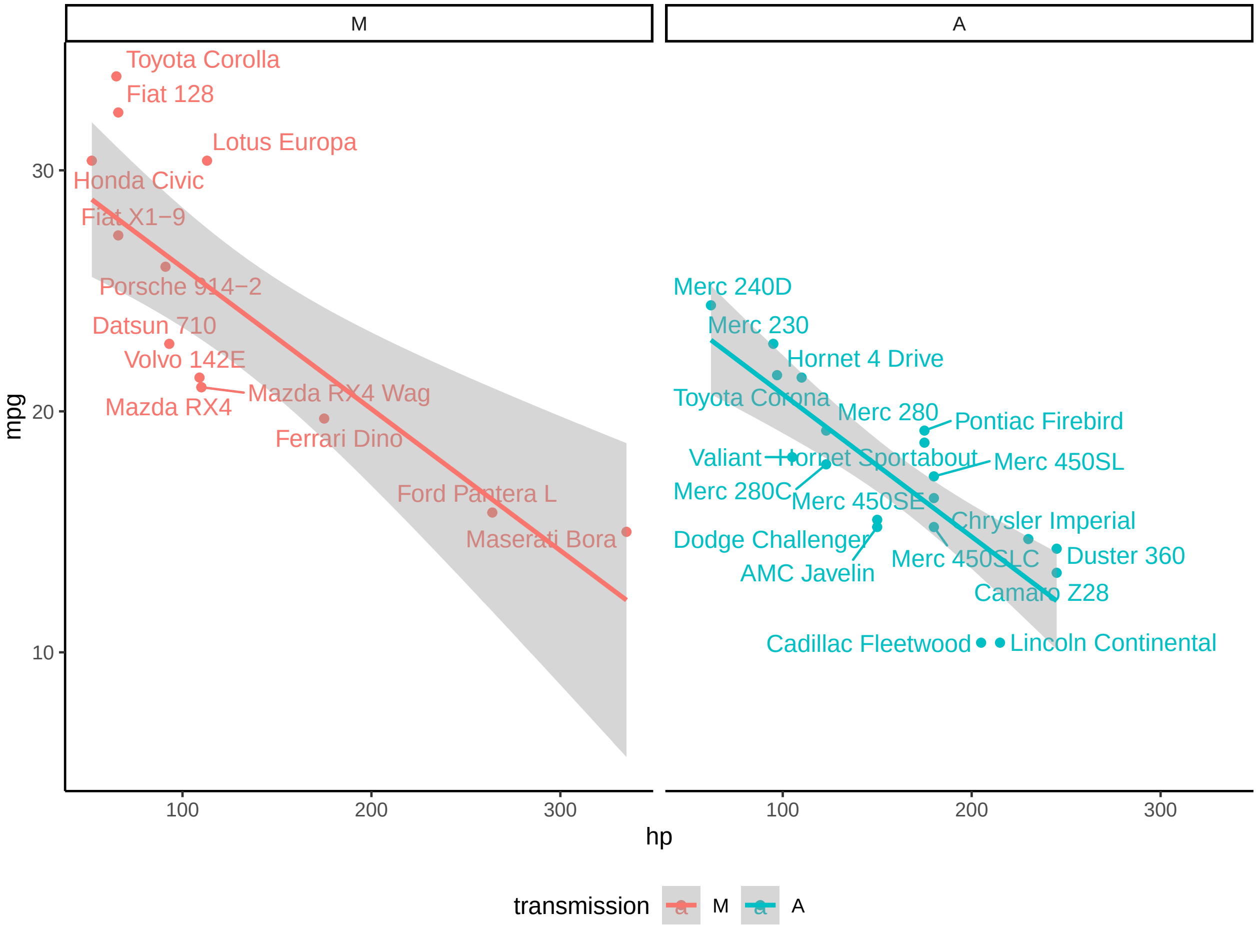
For one, variances (from the violin and splom plots) of mpg vs each transmission group is anything but similar. Its very possible we cannot trust the p-values or confidence interval for inference. Check the diagnostic plots:



You can see the heterosckedacity (I love that word) in scale-location: variances differ thus sd of the residuals vs the fitted values show a not-at-all horizontal line. It was evident before, but these are the diagnostic plots.

So if we hold other variables constant, that is, if we remove the effect of some other variable, we might offer the readers something a bit more interesting.

For example, a clear variable that tells us a lot about a car's performance is the horse power. It stands to reason that the more HP, the more we will be spending at the gas station. Lets explore HP vs mpg taking into account transmission :

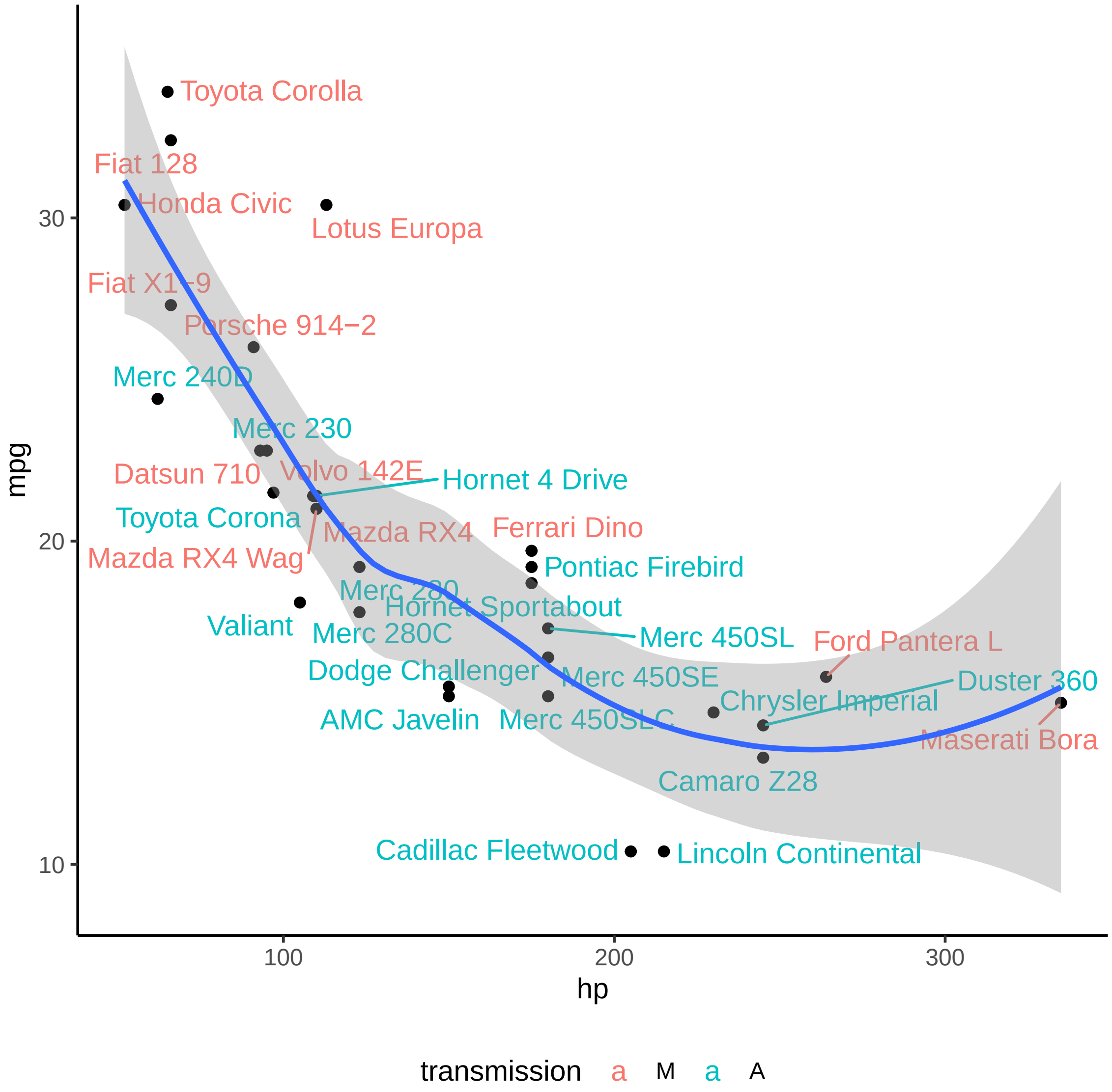


This would be an additive model, that is,

$$MPG(HP, transmission) = \beta_0 + \beta_1 transmision + \beta_2 HP$$

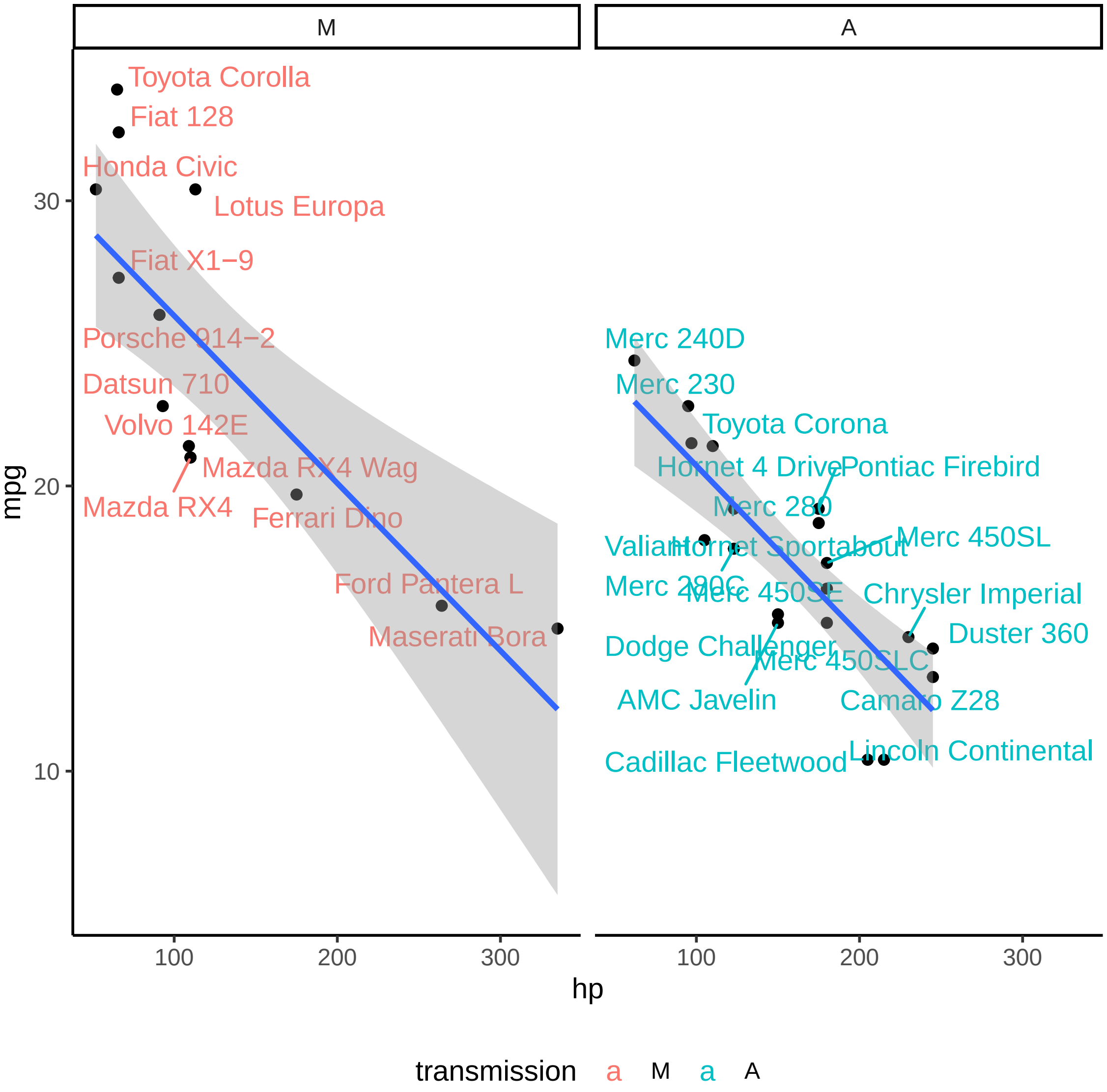
thus for manual transmission we would get the line on the left, for auto the one on the right.

And it does look like a pretty good fit especially for automatic transmission, but not for the manual cars. This HP vs MPG linear model was explored by the teacher in the course. To remind you:

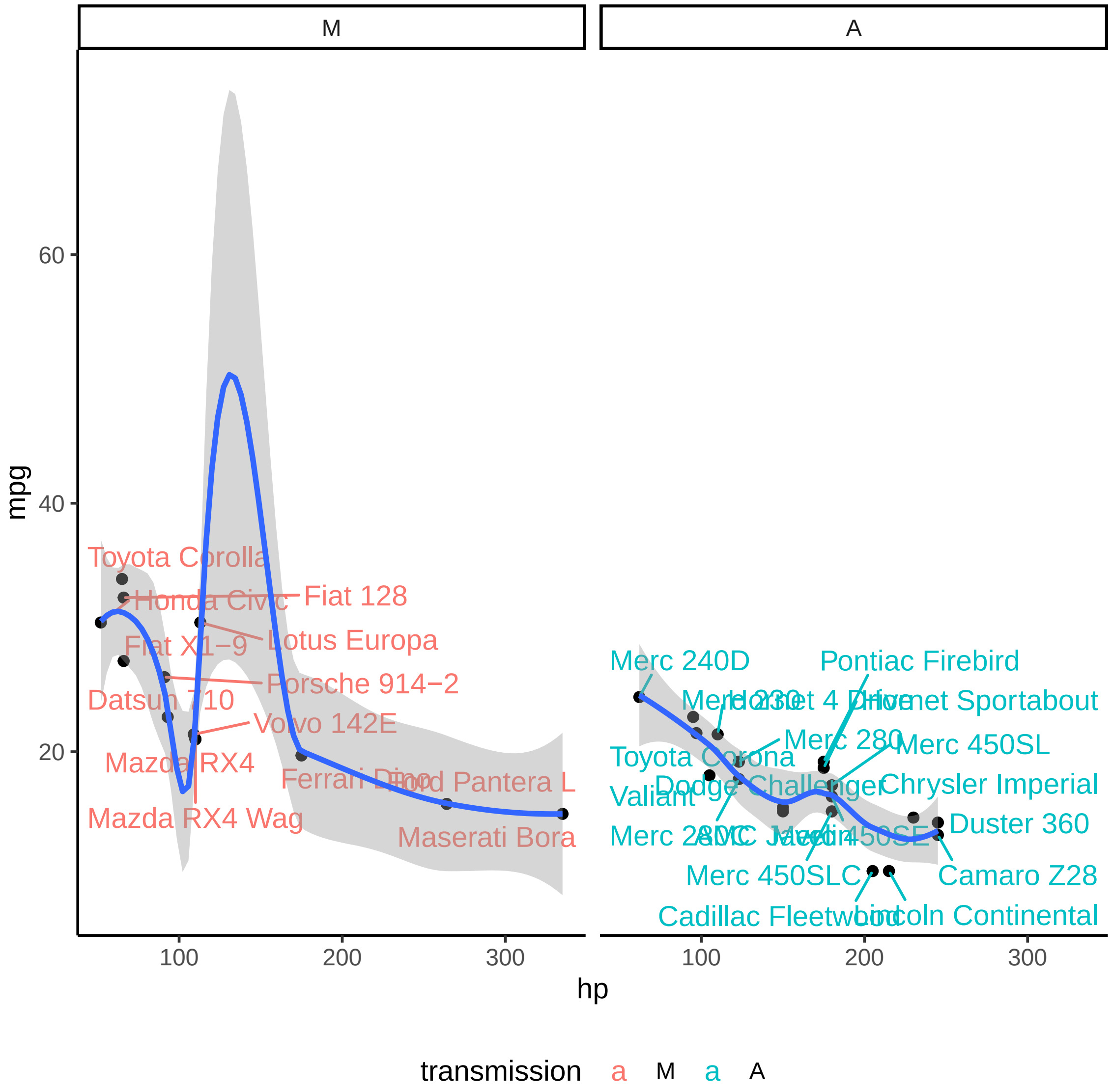


The teacher said: “maybe throw in some quadratic term and it’ll be better, although that cluster in the center looks … etc.”

So even in this one, with a loess approximation fit, is missing some things. If we look at it by transmission, and linear models:

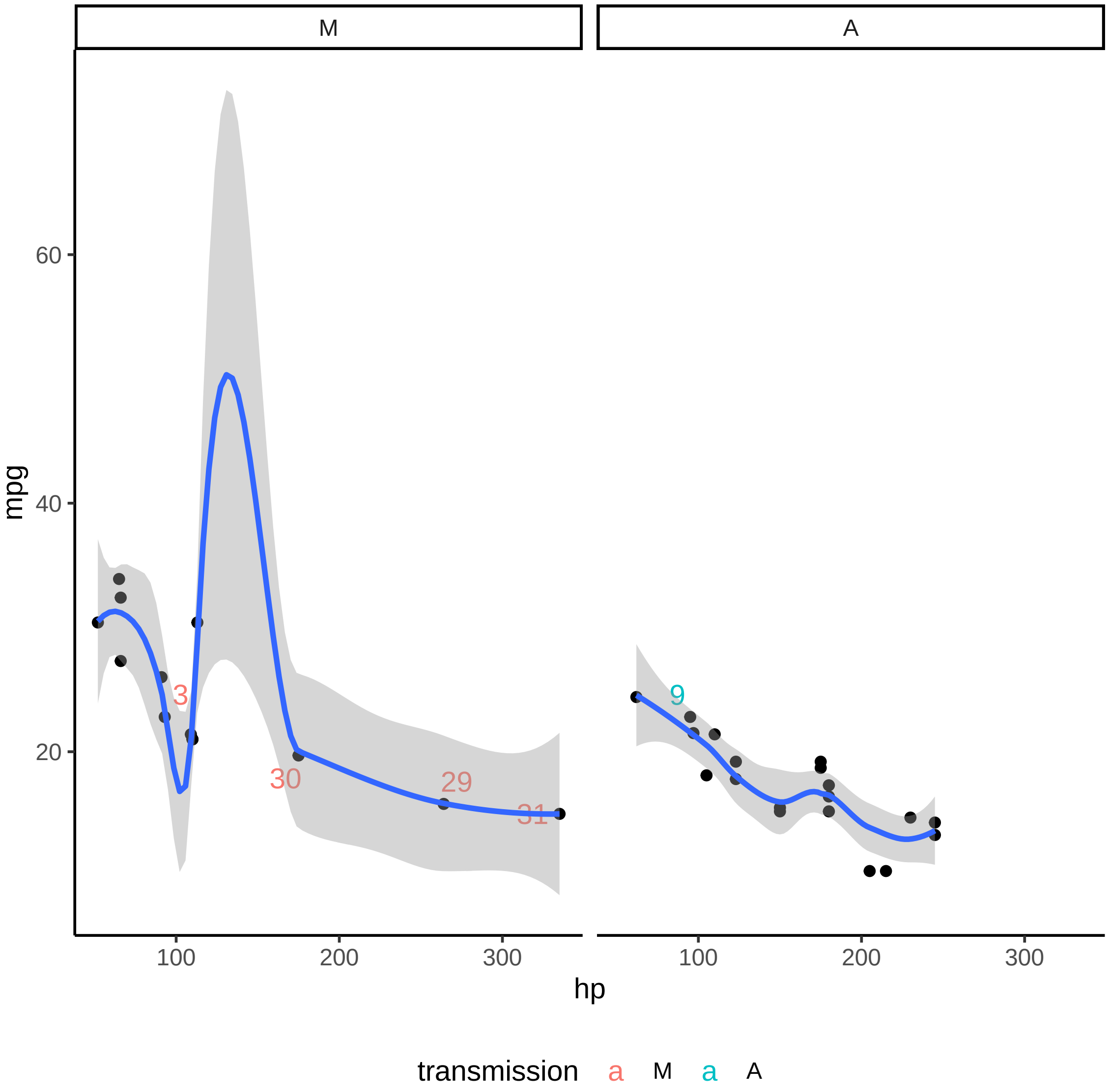


We can see in the automatic, HP explains the relationship *well-ish*, save for the cluster arround the pontiac firebird and three mercedes. The manuals though, look pretty bad. Check them out with loess:



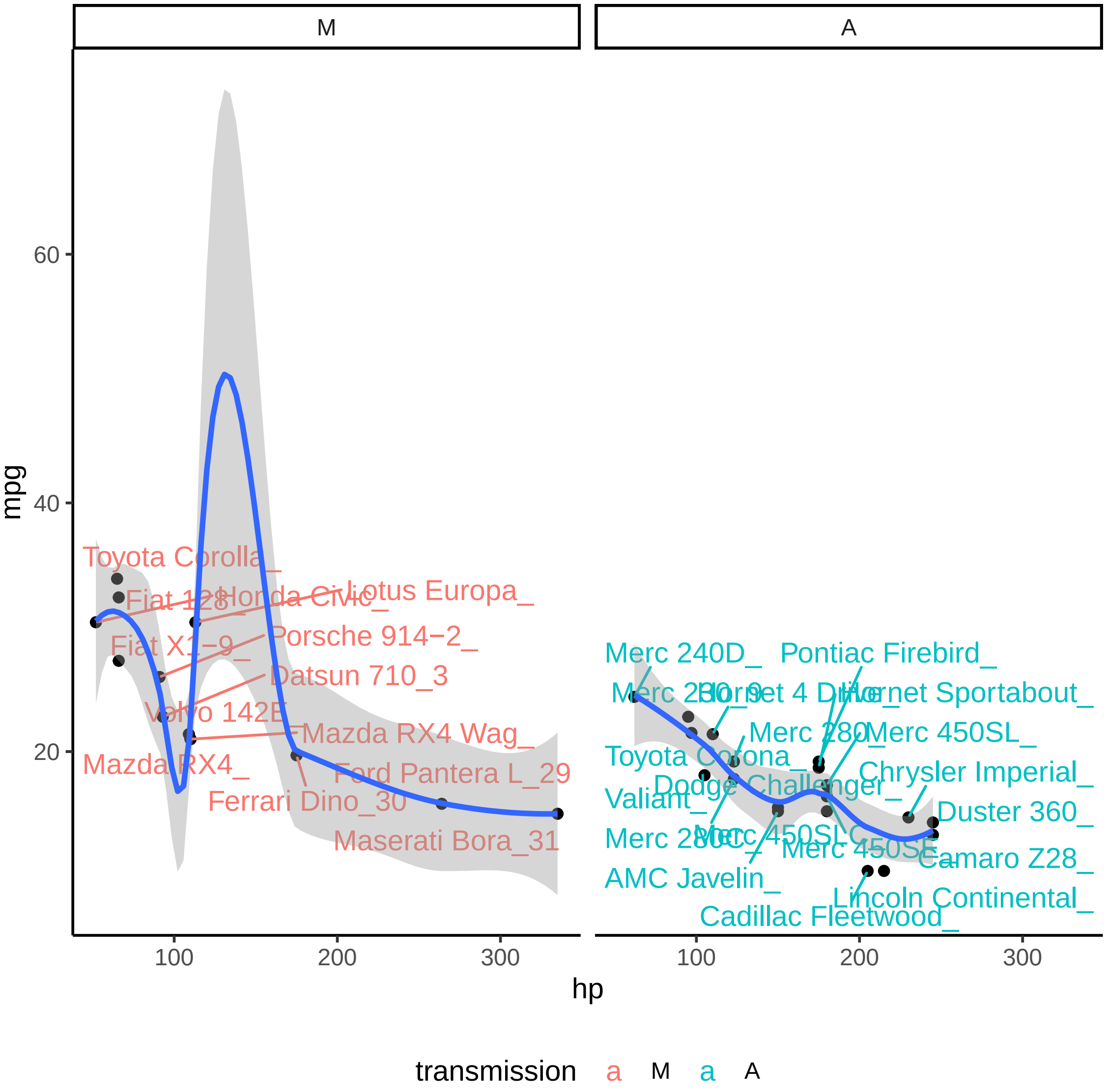
Now that is some freaky stuff in the manuals clearly because of the lotus europa and, save for the aforementioned cluster in the automatics, it looks well.

Go back to the diagnostic plots and verify that 29, 30 and 31 are the most extreme cars row numbers in all meassures. Also go back and see the table output from our very first try, where I show the cars that do not comply with the assumption that mpg depends solely on transmission, they are 3 and 9. Lets see where they lie:



All three extremes are manual cars. 3 and 9 have the exact same mpg and one is auto and the other is manual.

Ok, so far so bad, right? Lets pause and recognize:

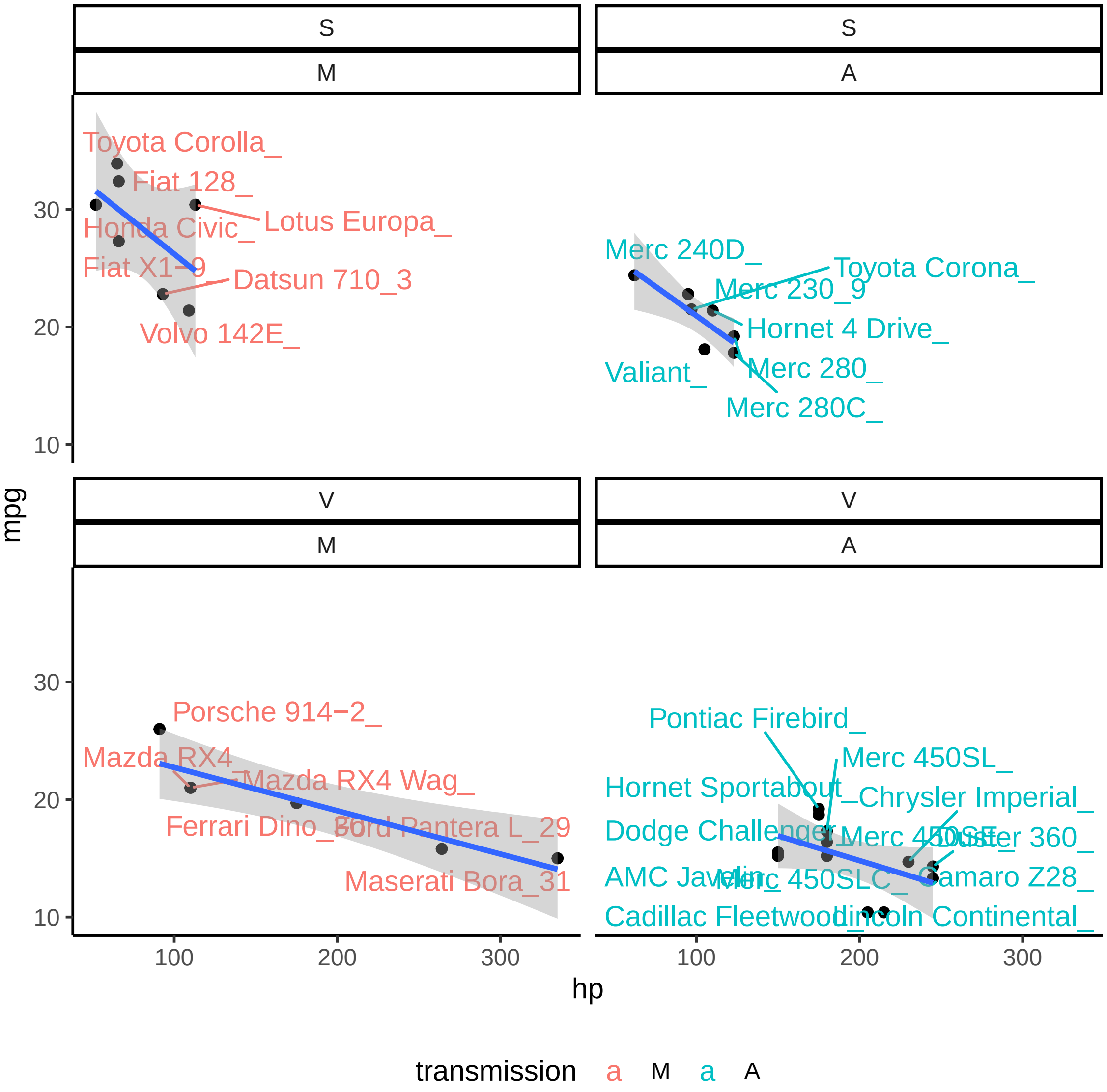


In the manual side, the toyota corola, fiat 128 and fiat X1-9 have almost the same HP but quite different MPG. Same goes, although graver, for the Volvo and the lotus europa. Then the relationship gets “inverted”: the 42E, the Ford Pantera, the ferrari dino and the Bora have much, much more horsepower than the rest but they dont diminish in MPG too much.

In the automatic side, things look better but for the “cluster” that plagues us. Both mercs 450s have almost identical HP but quite less MPG than the Hornet (up to confidence interval).

So something is going on and its worse in the manual side. This is why I compiled the table of variables: we need to know other variables somewhat unrelated to HP, but still related to mpg might explain all those differences.

Perhaps engine design?



Aha! This gets rid of problems in the manual side but really widens the interval for the Straight engine, Manual transmission (upper left) cars. Also, there is bound to be variance inflation.

We can see the model now:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	31.2915817	1.270609	24.627227	0.0000000	28.6888567	33.8943067
transmissionA	-5.2985368	1.037569	-5.106683	0.0000207	-7.4239009	-3.1731727
hp	-0.0447165	0.010776	-4.149623	0.0002810	-0.0667902	-0.0226428
vsV	-2.6588460	1.442471	-1.843258	0.0759010	-5.6136131	0.2959212

Notice the p-values: vsV, that is, automatic cars with V engine are not a significant group. Its closeish, but not there. From the plot, you can see that the “clusters” might still be messing things up. In particular, its notorious that so many v-shaped engine, automatic cars there have the same HP but very different mpg, save for the cadillac and the lincoln.... hum...

Also, three mercs 450 have (almost?) the exact same hp, but very different mpg.

So thinking about this I thought: what about the weight. I mean I remember cadillacs and licolns: those were big big cars (mtcars data set comes from the 70s). So maybe the interaction of hp with weight or…something… but we have been warned in the course that interactions are dangerous: we should very well prove that use of interaction terms is warranted.

I came up with a nice little measure called the power-to-weigt ratio. Therein, it is said:

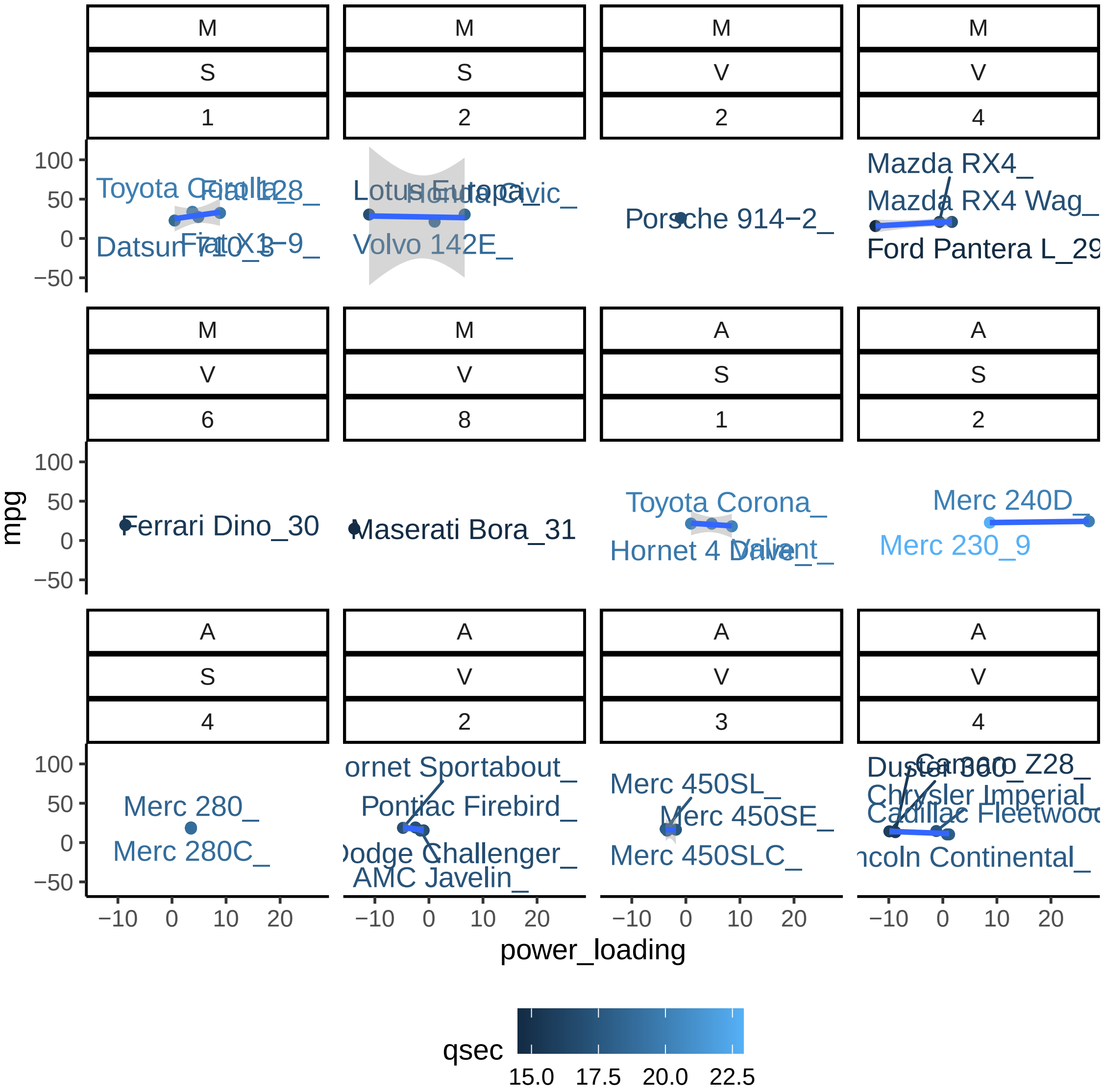
> **Power-to-weight ratio** (**PWR**) (also called **specific power**, or **power-to-mass ratio**) is a calculation commonly applied to engines and mobile power sources to enable the comparison of one unit or design to another. Power-to-weight ratio is a measurement of actual performance of any engine or power source. It is also used as a measurement of performance of a vehicle as a whole, with the engine's power output being divided by the weight (or mass) of the vehicle, to give a metric that is independent of the vehicle's size. Power-to-weight is often quoted by manufacturers at the peak value, but the actual value may vary in use and variations will affect performance.
>
> The inverse of power-to-weight, weight-to-power ratio (power loading) is a calculation commonly applied to aircraft, cars, and vehicles in general, to enable the comparison of one vehicle's performance to another. Power-to-weight ratio is equal to thrust per unit mass multiplied by the velocity of any vehicle.

Now this is okay, but im no mechanic. What does "power" mean? Well, I found this other resource and in that one:

> It's very easy to calculate a power-to-weight ratio. Simply divide the power output of a vehicle by its weight. For example, if you have a car that weights 2000 pounds and has 250 hp, the PWR will be as follows:
>
> 250 / 2000 = 0.125 hp for every pound of car.
>
> https://goodcalculators.com/power-to-weight-ratio-calculator/
> © 2015-2021 goodcalculators.com

So instead of using an interaction term in the linear model, and because Im a newb, ill just calculate a new variable p2w:

And now we are going to use that instead of hp, disregarding engine design (V or S):



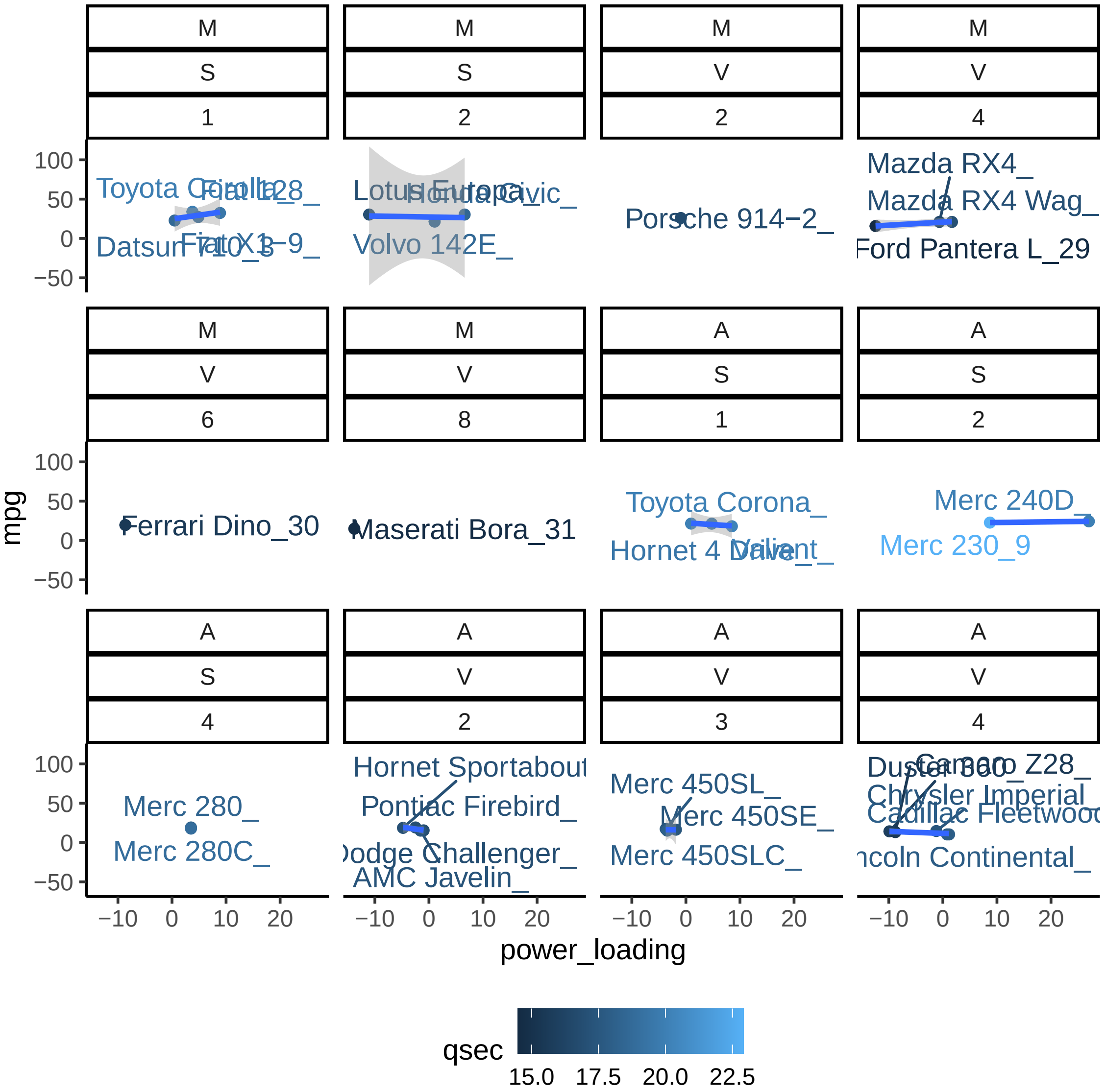
This took care of most clustery thingies and now we are ready to answer the questions in a more effective manner

The model becomes:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	27.0939592	1.1415569	23.734216	0.0000000	24.7555858	29.4323326
transmissionA	-6.7813019	1.3166603	-5.150381	0.0000184	-9.4783582	-4.0842455
vsV	-5.2923311	1.5904555	-3.327557	0.0024607	-8.5502314	-2.0344307
power_loading	0.1670455	0.1031113	1.620051	0.1164311	-0.0441684	0.3782594

Still not good that the power loading seems insignifficant. It most certaintly is not. But it might also be that we simply have not enough evidence within all those groups (transmissionXvs) for us to declare power loading a determining factor.

One more grouping variable is the number of carburators. This certaintly would be a part of whatever gobbles up gas, depending on engine design and transmission. Check the graphic:



Here you can see that we cant even fit a line for some groups as they identify one exact car. In the end what we have is a flimsy group. However, this does give us a window to our data:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	27.7006499	1.4021459	19.7558970	0.0000000	24.8067630	30.5945368
transmissionA	-7.6793232	1.3481247	-5.6963003	0.0000072	-10.4617159	-4.8969306
factor(carb)2	-1.1293749	1.6612838	-0.6798206	0.5031232	-4.5580961	2.2993463
factor(carb)3	-3.1114602	2.5424484	-1.2238047	0.2329051	-8.3588158	2.1358954
factor(carb)4	-6.0827941	1.8042261	-3.3714145	0.0025290	-9.8065339	-2.3590544
factor(carb)6	-6.2527975	3.7328532	-1.6750719	0.1069029	-13.9570279	1.4514329
factor(carb)8	-9.9047862	3.8997164	-2.5398734	0.0179798	-17.9534054	-1.8561670
power_loading	0.2026374	0.0919255	2.2043656	0.0373314	0.0129125	0.3923623

So, transmission is relevant, so is power loading, so is carburators at the 4 and 8 levels. This is enough: its clear to me that if a I knew more about cars I could create a grouping of interaction between transmission and carburators to declare cars more or less “muscled” or “big”, and that would yield real similar cars for which to fit lines of power_loading vs mpg, and it would tell us the whole story.

The Answer

But I dont so… im keeping this and asserting to my magazzine bosses:

> The car with average weight to power ratio, manual transmission and 1 carburator gives between 25 and 30.6 MPG. The car with the same loading and carburator but on an automatic transmission, spends almost 5 to 10 MPG more.
>
> If the car has 4 carburators instead of 1, it spends 2 to 10 MPG more. If it has 8, it can spend from 2 to 18 MPG extra.