Motor Trend Cars Regression Model

Maurício Collaça

Nov 22, 2017

Executive Summary

The Motor Trend magazine are interested whether an automatic or manual transmission is better for MPG and quantifying difference. The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance.

The following sections explain that average mpg for automatic transmission, average weight and average 1/4 mile time is 18.9 with a 95% confidence interval. Manual cars increases mpg because the manual transmission, holding fixed the weight and the 1/4 mile time, increases mpg by 2.94 miles with a 95% confidence interval.

Exploratory Data Analysis

The dataset contains 32 observations on 11 variables and row names for each car. A small preview:

	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

Legend: `mpg` Miles/gallon; `cyl` Number of cylinders; `disp` Displacement (cu.in.); `hp` horsepower; `drat` Rear axle ratio; `wt` Weight (1000 lbs); `qsec` 1/4 mile time; `vs` V/S; `am` Transmission (0=auto, 1=manual); `gear` Number of forward gears and `carb` Number of carburetors.

Figure 1 shows many variables are correlated with the `mpg` and `am`, characterizing a multicollinearity issue. Figure 2 suggests that manual transmission is better for `mpg` with a higher variance and that automatic cars are heavier than manual cars. The variables `wt`, `cyl`, `disp` and `hp` are the highest correlated to `mpg`. Figure 3 show these variables are somewhat measuring the same thing. It makes sense weight could be the main regressor of `mpg` as Newton's 2^{nd} law of motion tell us force is the product of mass and acceleration: $F = ma$. The variable directly related to acceleration is `qsec`. Its unit in seconds suggests the more seconds the less acceleration and somewhat less fuel consumption. It makes sense `qsec` is positively correlated to `mpg` and negatively correlated to `wt` but differently for each transmission condition as shown in figure 4.

Regression models

The strategy is perform a nested likelihood test of models by first adding the regressor of interest `am` and progressively the more relevant regressors according to the previous EDA which is also supported by the AIC stepwise model selection algorithm:

```
step(lm(mpg ~ ., mtcars), direction="both", trace=0)$call
```

```
lm(formula = mpg ~ wt + qsec + am, data = mtcars)
```

`wt` and `qsec` are centered for meaningful intercept. The tests show model improvements:

```
fit1 <- lm(mpg ~ am, data = mtcars)
fit2 <- update(fit1, mpg ~ am + I(wt-mean(wt)))
fit3 <- update(fit1, mpg ~ am + I(wt-mean(wt)) + I(qsec-mean(qsec)))
data.frame(anova(fit1, fit2, fit3), check.names = FALSE)
```

```
  Res.Df      RSS Df Sum of Sq        F       Pr(>F)
1     30 720.8966 NA        NA       NA           NA
2     29 278.3197  1  442.5769 73.20250 2.672820e-09
3     28 169.2859  1  109.0338 18.03425 2.161737e-04
```

The first and third models result significant transmission slope coefficients.

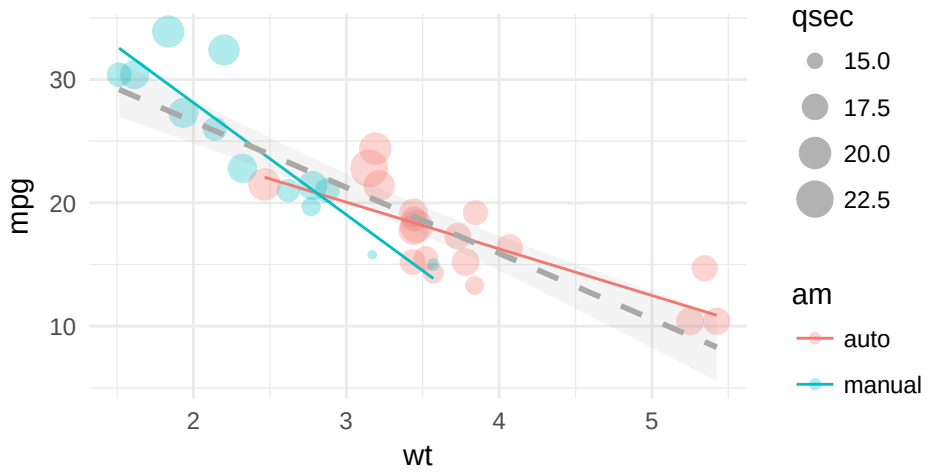
```
             Estimate Std. Error   t value     Pr(>|t|)
```

```
(Intercept) 17.147368   1.124603 15.247492 1.133983e-15
ammanual     7.244939   1.764422  4.106127 2.850207e-04
```

```
                   Estimate Std. Error     t value     Pr(>|t|)
(Intercept)      20.10021868  0.8331837 24.12459551 9.639957e-21
ammanual         -0.02361522  1.5456453 -0.01527855 9.879146e-01
I(wt - mean(wt)) -5.35281145  0.7882438 -6.79080719 1.867415e-07
```

```
                     Estimate Std. Error   t value     Pr(>|t|)
(Intercept)          18.897941  0.7193542 26.270704 2.855851e-21
ammanual              2.935837  1.4109045  2.080819 4.671551e-02
I(wt - mean(wt))     -3.916504  0.7112016 -5.506882 6.952711e-06
I(qsec - mean(qsec))  1.225886  0.2886696  4.246676 2.161737e-04
```

Assuming the third model as the best one, the following plot aims to show the relationship between selected terms.



The variance-inflation factors aren't too high:

```
vif(fit3)
```

```
              am     I(wt - mean(wt)) I(qsec - mean(qsec))
        2.541437             2.482952             1.364339
```

The confidence interval considering average `wt` and `qsec`:

```
confint(fit3)
```

```
                         2.5 %    97.5 %
(Intercept)          17.42441087 20.371471
ammanual              0.04573031  5.825944
I(wt - mean(wt))     -5.37333423 -2.459673
I(qsec - mean(qsec))  0.63457320  1.817199
```

There are some potentially influential observations:

```
summary(influence.measures(fit3))
```

	dfb.1_	dfb.ammn	dfb.I(-mn(w))	dfb.I(-mn(q))	dffit	cov.r	cook.d	hat
15	0.03	-0.08	-0.15	-0.06	-0.17	1.47	0.01	0.23
16	-0.01	0.02	0.04	0.02	0.05	1.57	0.00	0.26
17	-0.17	0.56	1.09	0.34	1.27	0.72	0.35	0.23

The figure 5 show residual plots with some violation of the residuals normality.

Conclusions

For the selected regression model, the average `mpg` for automatic transmission, average `wt` and average `qsec` is 18.9 with a 95% confidence interval from 17.42 to 20.37. The manual transmission, holding fixed `wt` and `qsec`, increases `mpg` by 2.94 with a 95% confidence interval from 0.05 to 5.83.

Appendix

Figure 1: Generalized pairs plots

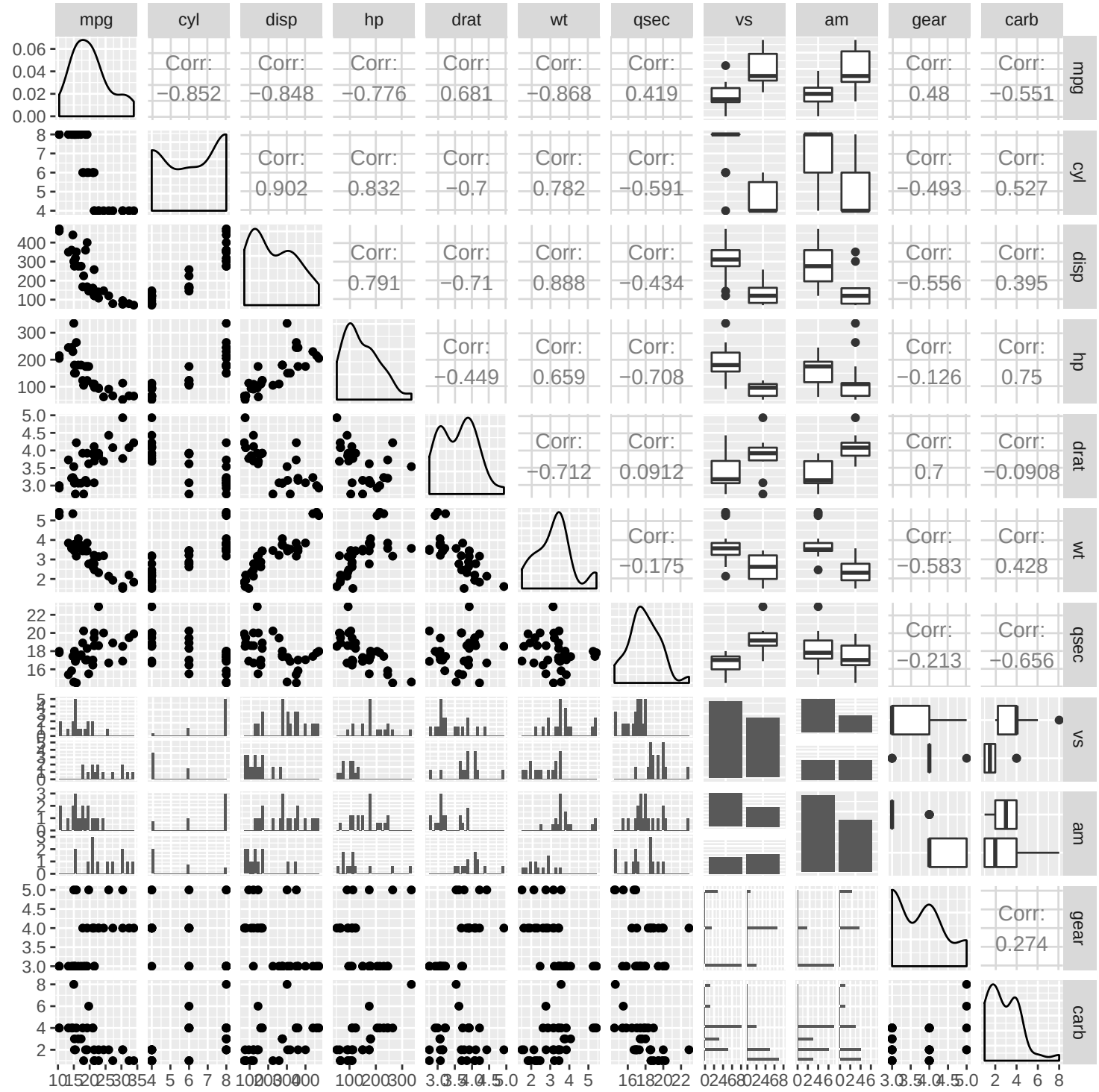


Figure 2: Miles per galon and Weight by transmission

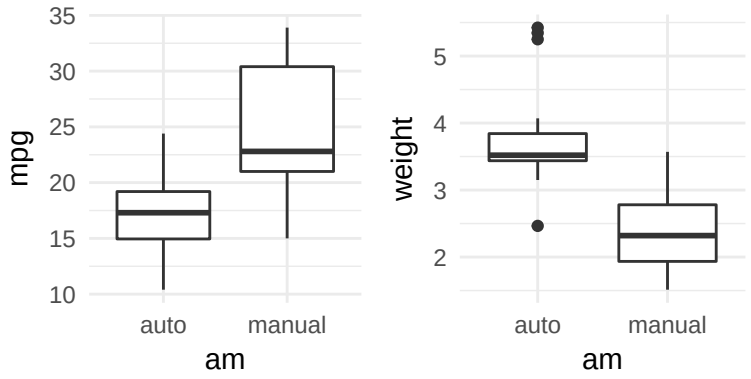


Figure 3: Miles per weight, number of cylinders and displacement volume.

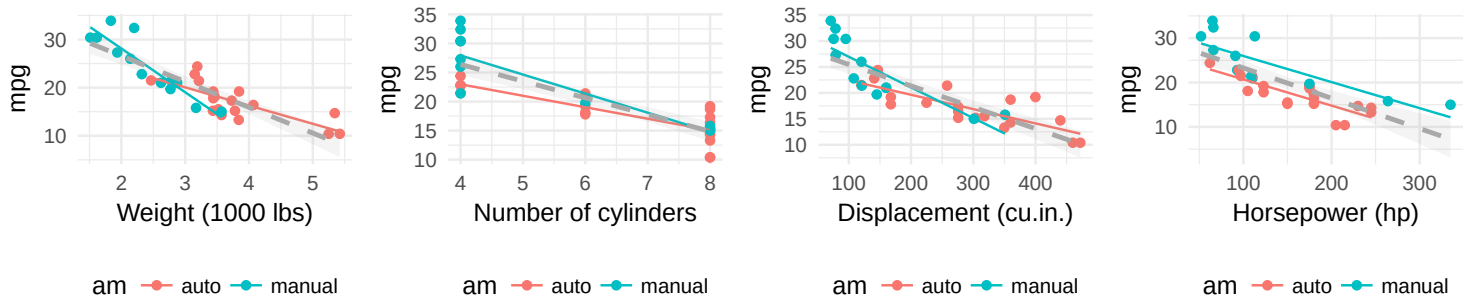


Figure 4: Miles per gallon and Weight by 1/4 Mile Time.

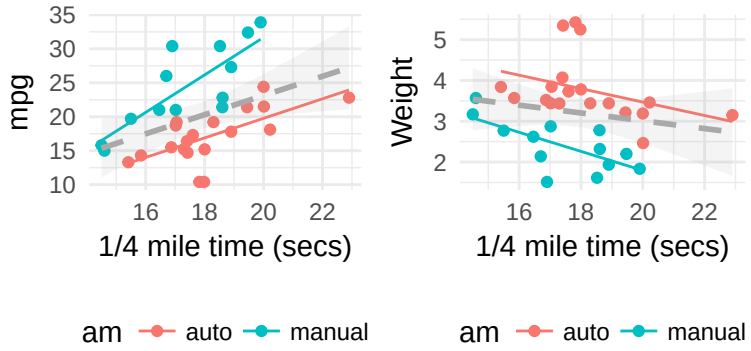


Figure 5: Residual plots