The relationship between fuel efficiency and transmission type in cars

Irena Papst

2019-07-08

Executive summary

Fuel efficiency is often an important factor considered when purchasing an automobile as fuel is one of the major operating costs of a car. If an engine is more fuel efficient than another, it is capable of converting the same amount of fuel into more distance traveled. In the US, fuel efficiency is measured in miles per gallon (mpg). It is commonly believed that cars with a manual transmission are more fuel efficient than those with an automatic transmission. We investigate this hypothesis using a linear regression model on the `mtcars` dataset available in `R`. We find **no statistically significant difference** in fuel efficiency between manual and automatic transmissions after controlling for car weight and number of cylinders (the two variables most highly correlated with fuel efficiency). We conclude with a discussion of the limitations of the data, and therefore of the study.

Research question

Is an automatic or manual transmission better for fuel efficiency, and can we quantify the fuel efficiency difference between transmission types in miles per gallon?

Looking at a box plot of the data split by transmission type (Figure 1), it appears plausible that manual transmissions are more fuel efficient than automatic ones.

Linear regression model

Regressor selection

Since we are interested in the relationship between transmission type (`am`) and fuel efficiency (`mpg`), we will certainly want to include transmission type as one of our regressors. To determine which other covariates to include in our models, we consider the correlation between fuel efficiency and all of the other variables in our dataset (Figure T1; Figure 2).

The most strongly correlated variables are all negatively correlated: weight in 1000 lbs (`wt`), number of cylinders (`cyl`), displacement in cubic inches (`disp`), and gross horsepower (`hp`). We restrict ourselves to this set of covariates and explore a series of nested linear regression models: model 1 (`mpg ~ am`), model 2 (`mpg ~ am + wt`), model 3 (`mpg ~ am + wt + cyl`), model 4 (`mpg ~ am + wt + cyl + disp`), model 5 (`mpg ~ am + wt + cyl + disp + hp`).

Model assumptions

Before selecting the best model to answer our research question, we first check that the standardized residuals are approximately normally distributed to ensure that one of the main assumptions of a linear model is satisfied in each case. Figure 3 shows that this assumption is satisfied for each model. We also verify that the means of the standardized residuals are near zero (Figure T2).

Model selection

To select a model, we perform an ANOVA on our set of nested models to test whether the addition of the next most correlated covariate offers a significant improvement over the immediately preceding model (Figure T3).

The ANOVA shows that model 2 is a highly significant improvement over model 1 ($p = 7.0166459 \times 10^{-9}$), and that model 3 is a further improvement over model 2 ($p = 9.4231666 \times 10^{-4}$). The subsequent models cannot be said to offer much further improvement, and overfitting will only serve to increase the actual error of regression coefficient estimation.

We select model 3, where fuel efficiency is predicted using transmission type, weight, and the number of cylinders (`mpg ~ am + wt + cyl`).

We check the variance inflation factors (VIFs) for each covariate in model 3 (Figure T4). Since the VIFs are moderate, we are not concerned about colinearity in this particular set of regressors.

Results and discussion

Full model results are presented in Figure T5. The regression coefficient of the transmission type is 0.1764932, which would indicate that a manual transmission offers a fuel efficiency increase of 0.1764932 miles per gallon over an automatic transmission (holding weight and number of cylinders constant). However, if we compute the 95% confidence interval around this regression coefficient (Figure T6), we get [-2.496, 2.849]. Since this confidence interval includes zero, we **cannot reject** the null hypothesis that there is no statistically significant difference in fuel efficiency between cars with manual and automatic transmission, after controlling for car weight and number of cylinders.

Data and study limitations

It is not known how fuel efficiency measures were gathered in this dataset; the original data source (Hocking 1976) simply states that "road tests were performed by 'Motor Trend' magazine in which gasoline mileage and ten physical characteristics of various types of automobiles were recorded". No data collection protocol is described, which raises several issues. For instance, were these data collected during a controlled experiment or using real-world data from various drivers? In the latter case, there may be bias comparing data from different drivers as driving style (*e.g.* aggressive, cautious) can affect fuel efficiency. Also, there is a known difference between fuel efficiency in "city" and "highway" driving conditions, so much so that the United States Environmental Protection Agency reports both figures on the fuel efficiency labels it issues ("Interactive Version of the Gasoline Vehicle Label," n.d.). It is not clear whether this factor has been either controlled or randomized for in the data collection protocol.

There is also a known bias toward exotic, non-US cars in this dataset, as noted by Henderson and Velleman (1981). It is possible that our finding of no significant difference in fuel efficiency between automatic and manual transmissions of exotic cars cannot be extrapolated to more common cars found in the US. The results of this report should be treated with caution as they are based on a biased data set.

References

Henderson, Harold V, and Paul F Velleman. 1981. "Building Multiple Regression Models Interactively." *Biometrics* 37 (2): 391–411.

Hocking, R. R. 1976. "A Biometrics Invited Paper. the Analysis and Selection of Variables in Linear Regression." *Biometrics* 32 (1). Wiley, International Biometric Society: 1–49. http://www.jstor.org/stable/2529336.

"Interactive Version of the Gasoline Vehicle Label." n.d. United States Environmental Protection Agency. https://www.epa.gov/fueleconomy/interactive-version-gasoline-vehicle-label.

Appendix

Figure T1: Correlation between fuel efficiency (mpg) and other variables in the dataset.

wt	cyl	disp	hp	carb	qsec	gear	am	vs	drat	mpg
-0.868	-0.852	-0.848	-0.776	-0.551	0.419	0.48	0.6	0.664	0.681	1

Figure T2: Means of standardized residuals for each model.

Model 1	Model 2	Model 3	Model 4	Model 5
0	0.0071562	-0.0013121	-0.0011076	0.002755

Figure T3: ANOVA results for the series of nested linear regression models. Significance codes are given in the last column and are: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	Significance
1	30	720.8966	NA	NA	NA	NA	
2	29	278.3197	1	442.576902	70.5432193	0.0000000	***
3	28	191.0471	1	87.272637	13.9105605	0.0009423	***
4	27	188.4258	1	2.621246	0.4178057	0.5236992	
5	26	163.1199	1	25.305963	4.0335681	0.0550966	.

Figure T4: Variance inflation factors for the covariates in model 3.

am	wt	cyl
1.924955	3.609011	2.584066

Figure T5: Regression coefficients in model 3.

Intercept	am	wt	cyl
39.41793	0.1764932	-3.125142	-1.510246

Figure T6: Confidence intervals for regression coefficients in model 3.

	2.5 %	97.5 %
Intercept	34.007153	44.8287134
am	-2.495554	2.8485408
wt	-4.991001	-1.2592836
cyl	-2.375245	-0.6452459

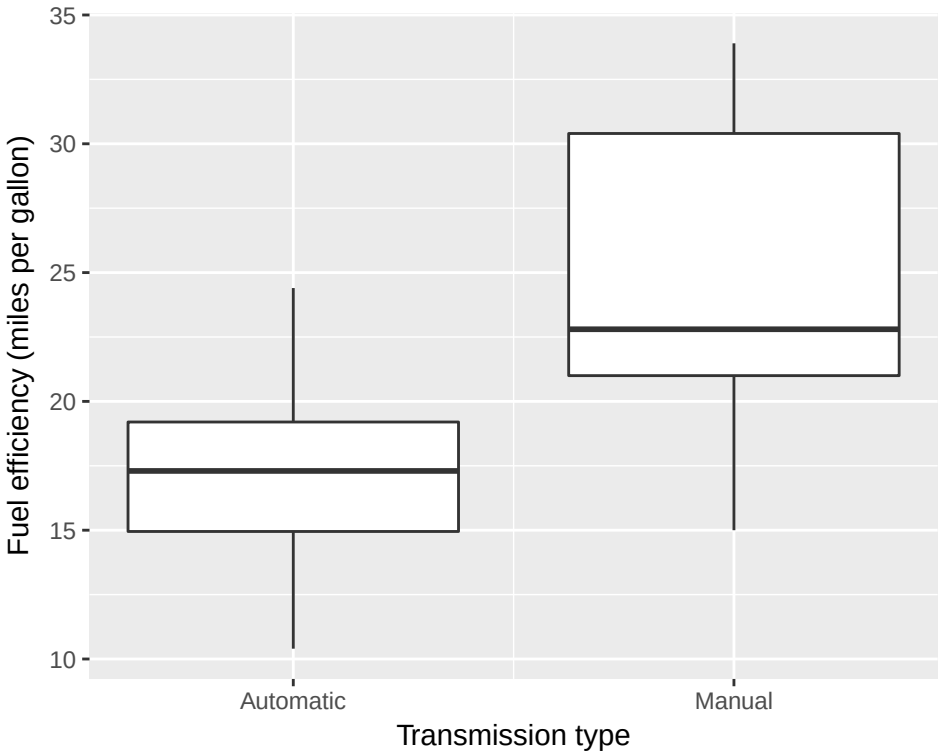

Figure 1: Boxplot of fuel efficiency for automatic and manual transmissions. The fuel efficiency of manual cars appears to be greater than that of automatic ones.

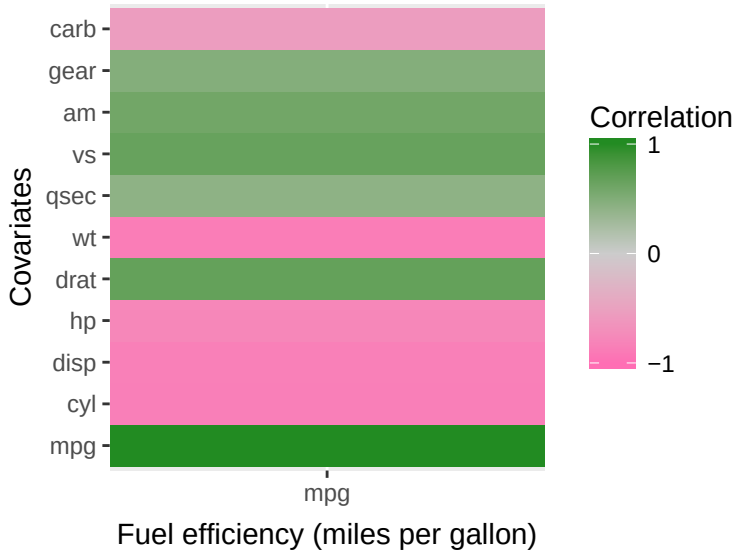

Figure 2: Correlation heat map between fuel efficiency (mpg) and other variables available in the dataset. The most strongly correlated variables are all negatively correlated: weight in 1000 lbs (`wt`), number of cylinders (`cyl`), displacement in cubic inches (`disp`), and gross horsepower (`hp`).

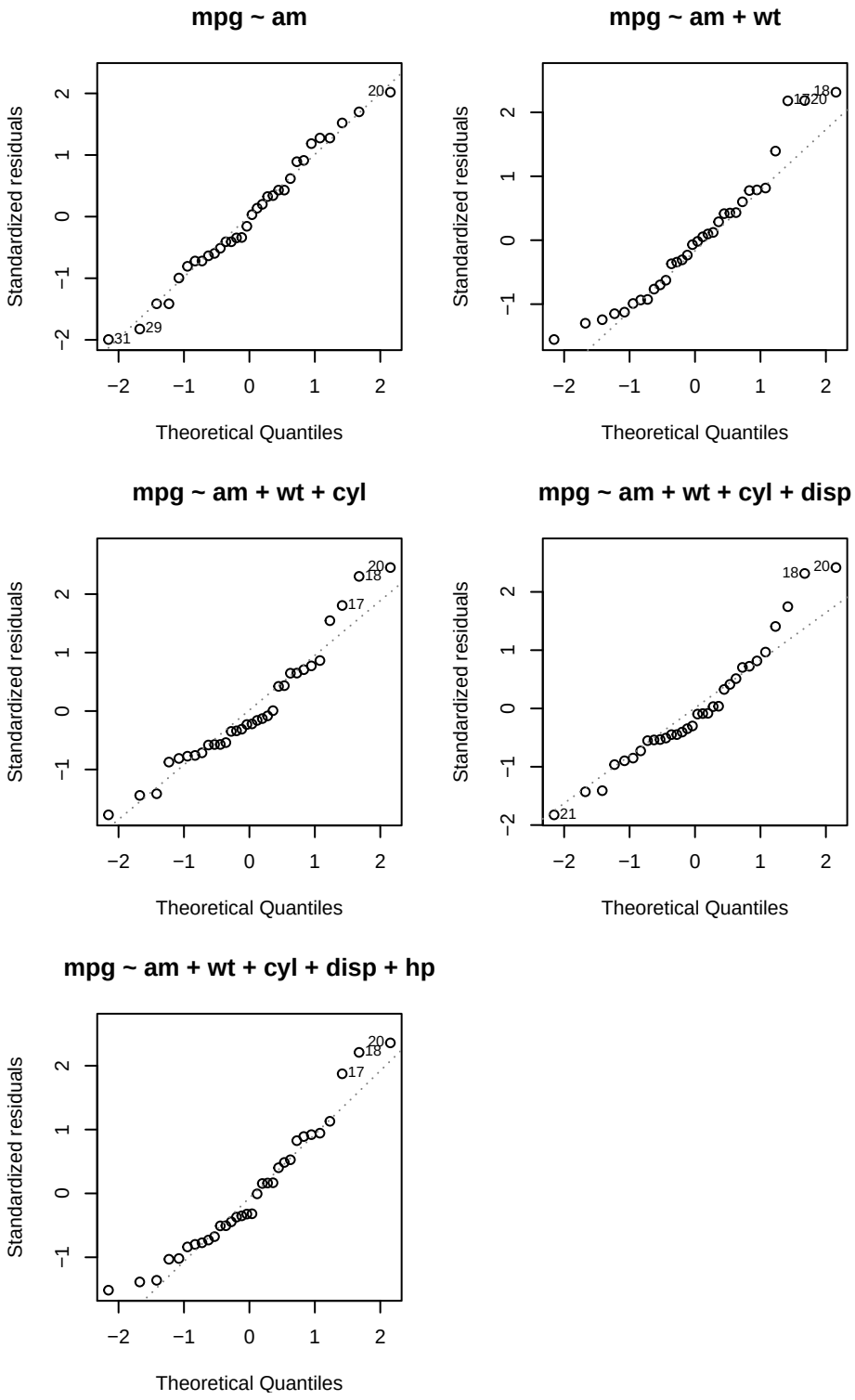

Figure 3: Normal q-q plots for the standardized residuals of the nested linear models, all of which have transmission type as a predictor and fuel efficiency as the response. The plots shows that the residuals appear to be reasonably normally distributed, satisfying a key assumption of a linear model.