Relationships between fuel efficiency and transmission type in cars

Irena Papst

2019-07-03

Executive summary

Fuel efficiency is often one of the factors considered when purchasing an automobile, as fuel is one of its 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 travelled. In the US, fuel efficiency measured in miles per gallon (MPG). It is commonly believed that cars with a manual transmission are generally 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 transmission 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 as well.

Data

There are 19 automatic cars and 13 manual cars in this sample, meaning that the automatic sample is about 46% larger than the manual sample. For such small sample sizes, this difference may present an issue, but we will not address it in this study.

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

Research questions

We seek to investigate the following research questions:

1. Is an automatic or manual transmission better for fuel efficiency?
2. Can we quantify the fuel efficiency difference between automatic and manual transmissions in miles per gallon (mpg)?

Model

We use a linear model to investigate our research questions.

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 2):

```
## [1] "Correlation between fuel efficiency (mpg) and other variables:"
```

```
##         wt        cyl       disp         hp       carb       qsec
## -0.8676594 -0.8521620 -0.8475514 -0.7761684 -0.5509251  0.4186840
##       gear         am         vs       drat        mpg
##  0.4802848  0.5998324  0.6640389  0.6811719  1.0000000
```

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 assumptions

Before selecting the best model for our purposes, 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 0:

```
## [1] "Means of standardized residuals for each model"
```

```
##         model_1     model_2      model_3     model_4     model_5
## 1 -1.238701e-17 0.007156206 -0.001312123 -0.00110763 0.002754995
```

Model selection

To select the best model to answer our research question, we perfom 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 model before it.

```
## Analysis of Variance Table
##
## 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
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 70.5432 7.017e-09 ***
## 3     28 191.05  1     87.27 13.9106 0.0009423 ***
## 4     27 188.43  1      2.62  0.4178 0.5236992
## 5     26 163.12  1     25.31  4.0336 0.0550966 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA shows that model 2 is a highly significant improvement over model 1 ($p = 7.0166459 \times 10^{-9}$), but 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 that is statistically significant, and such overfitting will only serve to increase the actual variance in error of regression coefficient estimation.

We select model 3, where fuel efficiency is predicted using transmission type, weight (in 1000 lbs), and the number of cylinders (`mpg ~ am + wt + cyl`).

We check the variance inflation factors (VIFs) for each covariate in our selected model.

```
##       am       wt      cyl
## 1.924955 3.609011 2.584066
```

Since the VIFs are moderate, we are not concerned about colinearity with this particular set of regressors.

Results and discussion

```
## (Intercept)          am          wt         cyl
##  39.4179334   0.1764932  -3.1251422  -1.5102457
```

```
##                 2.5 %     97.5 %
## (Intercept) 34.007153 44.8287134
## am          -2.495555  2.8485408
## wt          -4.991001 -1.2592836
## cyl         -2.375245 -0.6452459
```

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/gallon over an automatic transmission (holding weight and number of cylinders constant). However, if we compute the 95% confidence interval around this regression coefficient, we get [-2.496, 2.849]. Since this confident interval includes zero, we conclude 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 the 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 phyiscal characteristics of various types of automobiles were recorded". No data collection protocol is desribed, 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). 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 on cars ("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 any differences observed in fuel efficiency between automatic and manual transmissions of exotic cars cannot be extrapolated to more common cards 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.

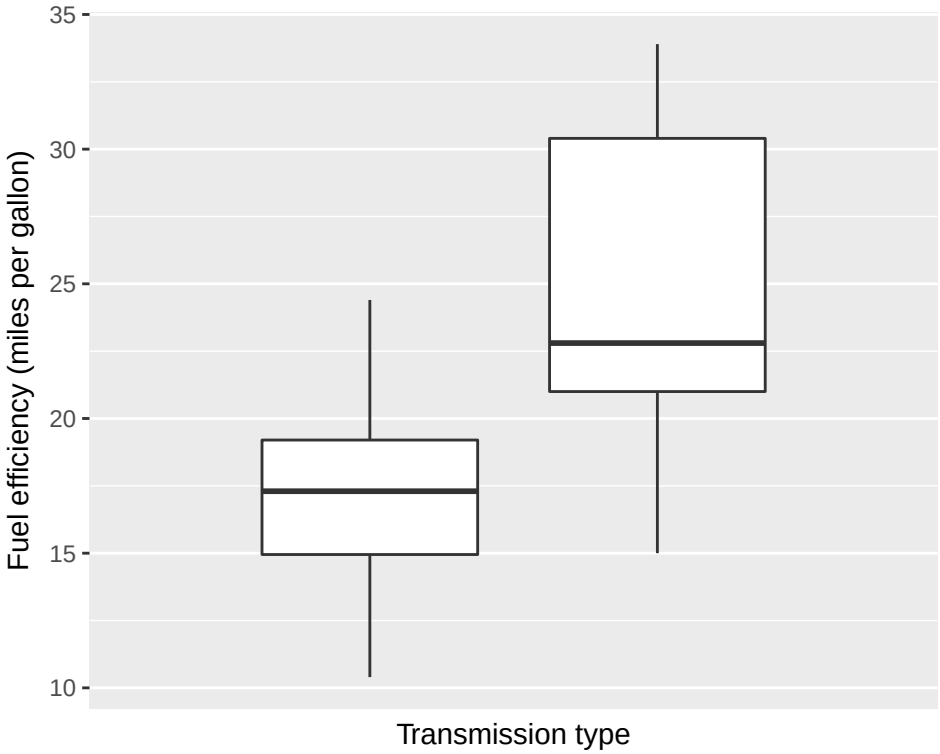

Figure 1: Boxplot of fuel efficiency for automatic and manual transmission. The fuel efficiency of manual cars appears to be larger than automatic ones.

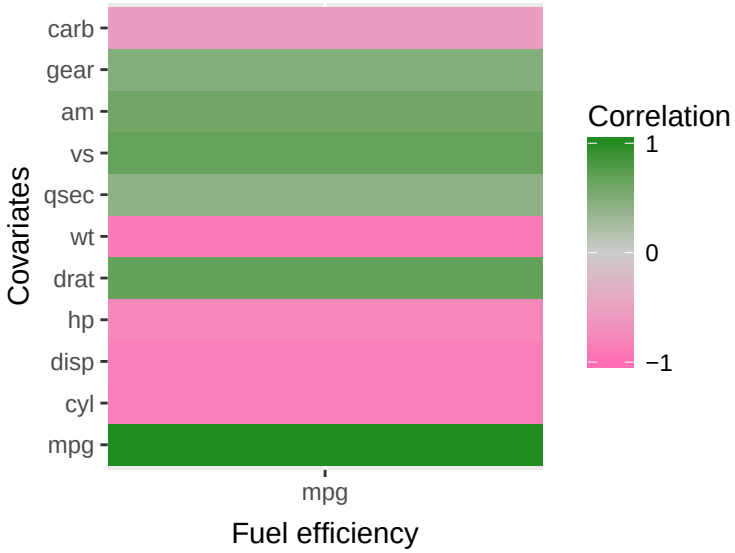

Figure 2: Correlation heatmap between fuel efficiency (mpg) and other variables available in the dataset. The most strongly correlated variables are all negatively correlated: weight in 1000 lbs (`wt`), number of cylinders (`cyl`), displacement in cubic inches (`disp`), and gross horsepower (`hp`).

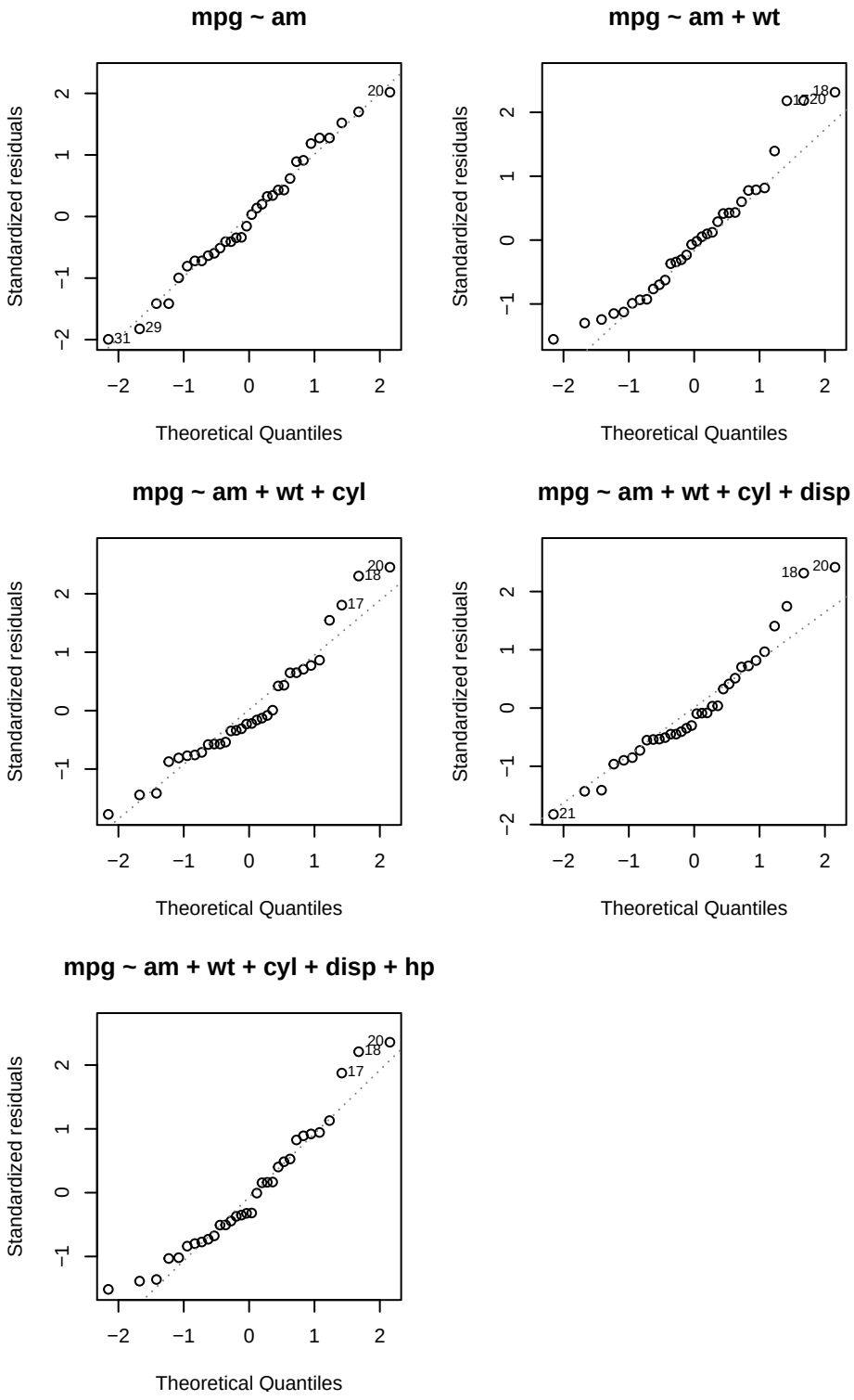

Figure 3: Normal q-q plots for the standardized residuals of the nested linear models, all of which have transmission type as a predictor and fuel efficiency as the response. The plots shows that the residuals appear to be reasonably normally distributed, satisfying one of the assumptions of a linear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