Predicting Fuel Economy of Conventionally Powered 2015 Vehicles

Abstract:

Fuel economy, the measure of fuel efficiency in automobiles, has gained attention in recent years. Engineers and economists are working to produce novel ways to reduce fuel consumption, because our non-renewable supply of energy and fuel is limited, rapidly decreasing and damaging our planet. Using fuel efficiency testing data from the Environmental Protection Agency, we used backward selection to create a multivariable linear regression model for predicting fuel economy. We found six variables strongly associated with fuel economy and describe the relationships. Our model may be used to predict fuel economy for other vehicles and may be helpful for consumers preparing to purchase a new vehicle.

Introduction and Significance:

Fuel efficiency is an important metric to consider when shopping for a vehicle.  Drivers are highly conscious that fuel consumption has a large effect on the operating cost of a vehicle. The less fuel a consumer uses, the more funds are available for them to spend elsewhere. Greater fuel efficiency provides numerous benefits to both the individual consumer and society at large. Decreasing oil consumption extends the life of Earth’s finite hydrocarbon-based energy resources. Extracting fuel from the ground can cause additional pollution (Jaffe, 2013) and destruction of habitat (Hoffman, 2015). Reducing consumption reduces these indirect effects and “ultimately saves money, preserves the environment and increases energy sustainability” (EPA, “Why is Fuel Economy Important?”). This study examines and quantifies potential associations between fuel efficiency and a variety of measurements recorded in conventionally powered vehicles produced during the 2015 model year.

Methods:

This study examines data on fuel efficiency for vehicles produced in the 2015 model year collected by the Environmental Protection Agency (EPA) (EPA Green Vehicle Guide). In collaboration with the Department of Energy, the EPA collects fuel-efficiency information on all mass-produced vehicles commonly sold in the United States. The EPA states “vehicles are tested in the same manner to allow fair comparisons” (EPA Green Vehicle Guide). The outcome of interest was combined miles per gallon (mpg), defined as “a combination of miles driven per gallon of gasoline in the city (55%) and highway (45%)” (“Model Year 2015 Fuel Economy Guide”). Further details on data collection may be found in the 2015 Fuel Economy Guide.

Covariate measurements include engine displacement, number of cylinders, drive train, transmission type, carbon dioxide emission levels, and air pollution scores. Because many factor variables were listed by the EPA as a large number of very-specific but similar types, we grouped some of these into more general categories. For example, manual cars with six gears and manual cars with five gears were both grouped together as cars with manual transmissions. Supplementary Table 1 in the appendix describes how these variables were grouped.

Before modeling, we considered whether we had prior knowledge that could help prevent collinearity or other problems. Based on the documentation (“Model Year 2015 Fuel Economy Guide”), we decided to only use carbon dioxide emission levels and not to consider the strongly associated variables of Greenhouse gas score and the indicators of whether or not the vehicle was a standard or elite smart car (“Consider a Smartway Vehicle”, 2015). The decision was made because 1) the nature of the greenhouse gas score suggests that carbon dioxide levels are subject to less measurement error and misclassification than the Greenhouse gas score (“Model Year 2015 Fuel Economy Guide”) and 2) over 80% of cars were not considered smart cars compared to only 17.7% and 1.7% of cars considered Smart and Smart Elite. Upon examining pairwise correlations, we found that cylinders and engine displacement were highly correlated (R=0.93, p < 0.0001). We decided that conceptually displacement is more informative. Thus, we considered displacement as a candidate for modeling rather than number of cylinders. The final dataset had seven variables to consider for model selection.

A multivariable linear regression model was used to examine which of the seven variables are related to miles per gallon. An automated backwards selection process in *R* version 3.2.0 (R Core Team, 2014) was used to fit the model. The process works by starting with an initial model including all seven of the candidate predictors and computes its Akaike information criterion (AIC). Second, each model with one fewer predictor than in the previous step is fit and its AIC computed. The model with the lowest Akaike information criterion (AIC) is chosen from both steps.  The second step may be iterated based on each subsequently smaller model. The procedure stops when the model in the previous step has a lower AIC than all possible models with one fewer variable. The AIC includes a penalty for using more predictors, producing an optimal model that balances fit and number of terms (Weisberg, 2005).

Results:

The EPA tested 2682 vehicles and included 15 factors that may affect fuel economy. Upon dropping vehicles with missing values for at least one of the variables, our data included 2310 vehicles. As described in the Methods section, we considered seven candidates for model selection, because we did not believe the remaining eight would provide additional predictive information after including the other measurements. Descriptive statistics of the variables under consideration are included in *Table 1* below and *Supplementary Table 2* in the appendix.

*Table 1*: Summary statistics of numeric variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Min** | **Max** | **St. dev.** |
| MPG | 23.3 | 10.0 | 50.0 | 5.8 |
| Displacement (L) | 3.2 | 1.0 | 8.4 | 1.3 |
| Cylinders | 5.6 | 3.0 | 16.0 | 1.9 |
| Greenhouse Gas Score1 (1 - 10) | 5.4 | 1.0 | 10.0 | 1.6 |
| Air Pollution Score2 (1 -10) | 5.7 | 2.0 | 9.0 | 1.1 |
| Combined CO2 Score (g/mi)3 | 390.0 | 178.0 | 847.0 | 98.3 |

1 This metric reflects tailpipe emissions of greenhouse gasses. “Vehicles that score a 10 are the cleanest” (“Greenhouse Gas Emissions”).

2 This metric is a score determined from measurements of various gasses, including carbon monoxide, formaldehyde, non-methane organic gasses, and non-methane hydrocarbons (Parker). High scores indicate a low level of pollutants.

3 Indicates CO2 produced in grams/mile averaged from highway and city tests (“Greenhouse Gas Emissions”).

        Backward selection based on the AIC resulted in a model with 6 of our predictor variables. The adjusted coefficient of determination for the final model is approximately 0.92. Thus, our final model explains about 92% of the variability in fuel efficiency. The overall p-value is very low (< 0.0001). We can therefore reject the null hypothesis that our variables aren’t linearly related to fuel efficiency. Our model indicates that there is a linear relationship between fuel economy and the predictor variables.

Parameter estimates, shown in *Table 2*, measure the expected change in fuel efficiency per unit change in each predictor variable, holding all others constant. For example, if we hold all other variables constant, trucks are associated with 0.67 miles per gallon less efficiency than cars. Logically, this makes sense, as trucks are larger and therefore less fuel efficient. The p-value means that there is less than a 0.01% chance of observing this coefficient by chance if there is no true effect of CO2 emission level on fuel efficiency.

Conclusion:

We conclude that there is a strong linear relationship between fuel economy and the ten factors listed in the summary. According to a study done by Newman, “vehicle mass dominates the regression” far more than other variables (Newman, Alimoradian, & Lyons). Our study agrees with this finding when using vehicle class as a surrogate for mass.

*Table 2:* Regression results from the backwards selection process.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Linear Model** | | | | |
|  | **Estimate** | **Std. Error** | **T value** | **P-value** |
| **Intercept** | 41.59 | 0.31 | 135.45 | < 0.0001 |
| **CO2  Levels (g/mile)** | -0.05 | 0.00 | -114.29 | < 0.0001 |
| **4-Wheel Drive (1=Yes, 0=No)** | -0.67 | 0.08 | -8.77 | < 0.0001 |
| **Fuel: Diesel (1=Yes, 0=No)** | 4.65 | 0.19 | 23.98 | < 0.0001 |
| **Fuel: Ethanol (1=Yes, 0=No)** | -2.64 | 0.16 | -16.91 | < 0.0001 |
| **Vehicle Class: SUV (1=Yes, 0=No)** | -0.68 | 0.09 | -7.72 | < 0.0001 |
| **Vehicle Class: Truck (1=Yes, 0=No)** | -0.60 | 0.17 | -3.56 | 0.0004 |
| **Vehicle Class: Van (1=Yes, 0=No)** | 0.22 | 0.23 | 0.94 | 0.3467 |
| **Transmission: CVT 1 (1=Yes, 0=No)** | 4.12 | 0.18 | 22.89 | < 0.0001 |
| **Transmission: Manual (1=Yes, 0=No)** | 0.52 | 0.08 | 6.42 | < 0.0001 |
| **Transmission: SCV 2 (1=Yes, 0=No)** | 1.54 | 0.18 | 8.76 | < 0.0001 |
| **Air Pollution Score** | 0.22 | 0.04 | 6.11 | < 0.0001 |

1 Continuously Variable Transmission

2 Static Volt-ampere reactive (VAR) Compensator

Based on this model, one may predict fuel efficiency for a hypothetical new car. An Acura ILX, for example, would have automatic transmission, 2WD, gasoline powered engine, expected emission of 316 g/mile of CO2 and an air pollution score of 5. We would predict that the new car should achieve a combined 27.4 mpg and fall between the 95% prediction interval (24.2 mpg, 30.6 mpg). The observed value of an Acura ILX in the EPA dataset was 28 mpg.

A limitation for this study is that linear regression is not robust to outliers and influential points.  The linear regression method is also limited due to the fact that it assumes that a change of a certain amount in the predictor variables will result in a constant change in the response variable. Future research could examine polynomial effects and interactions between variables. A weakness of using backward selection based on the AIC is that the process only finds a relative minimum in AIC; consequently not all combinations of predictors are computed.

Our analysis illuminates factors related to fuel economy. Consumers may use the results for a more educated vehicle purchase. Vehicle manufacturers may consider the findings in production of future vehicles. One may hope that these findings will lead to a cleaner and efficient future.

Sources Cited:

Consider a SmartWay Vehicle. (2015, March 3). US Environmental Protection Agency. Retrieved May 29, 2015, from http://www.epa.gov/greenvehicles/you/smartway.htm

EPA Green Vehicle Guide. (2015, March 25). US Environmental Protection Agency. Retrieved May 23, 2015, from <http://www.fueleconomy.gov/feg/download.shtml>

Greenhouse Gas Rating. (2015, March 3). US Environmental Protection Agency. Retrieved May 29, 2015, from http://www.epa.gov/greenvehicles/you/ghg.htm

Hoffman, J. (2015, May 20). Potential Health and Environmental Effects of Hydrofracking in the Williston Basin, Montana. Retrieved May 29, 2015, from https://serc.carleton.edu/NAGTWorkshops/health/case\_studies/hydrofracking\_w.html

Jaffe, M. (2013, Febuary 19). Study Finds Oil and Gas Drilling Caused Air Pollution in West. *Denver Post.* Retrieved May 27, 2013, from<http://www.denverpost.com/ci_22623664/study-finds-oil-and-gas-drilling-caused-air>

Model Year 2015 Fuel Economy Guide. (2015, May 21). US Environmental Protection Agency. Retrieved May 23, 2015, from <http://www.fueleconomy.gov/feg/pdfs/guides/FEG2015.pdf>

Newman, P., Alimoradian, B., & Lyons, T. (1989, February). Estimating Fleet Fuel Consumption for Vans and Small Trucks. Transportation Science, 23(1), 46-50.

Parker, A. (n.d.). What is my vehicle's air pollution score? Retrieved May 29, 2015, from <http://auto.howstuffworks.com/fuel-efficiency/hybrid-technology/air-pollution-score.htm>

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Weisberg, S. (2005). Logistic Regression. In Applied Linear Regression (3rd ed., pp. 210-220). Hoboken, New Jersey: Wiley.

"Why Is Fuel Economy Important?" *Why Is Fuel Economy Important*. U.S Environmental Protection Agency, n.d. Web. 29 May 2015.

Appendix:

*Supplementary Table 1:* Grouped Factor Variables. The EPA’s descriptions of specific categories were combined into broader categories.

|  |  |
| --- | --- |
| **Vehicle Class** |  |
| Car | small car, midsize car, large car, station wagon |
| Van | minivan, van, specific “special purpose vehicle”, done on a case by case basis |
| Truck | pickup, specific “special purpose vehicles” done on a case by case basis |
| SUV | small SUV, standard SUV |
| **Transmission** |  |
| Manual | Man-5, Man-6, Man-7 |
| Automatic | AMS-6, AMS-7 AMS-8, Auto-1, Auto-4, Auto-5, Auto-6, Auto-7, Auto-8, Auto-9, AutoMan-5, AutoMan-6, AutoMan-7, AutoMan-8, SemiAuto-4, SemiAuto-5, SemiAuto-6, SemiAuto-7, SemiAuto-8, SemiAuto-9 |
| SCV | SCV- SCV-6 SCV-7, SCV-8 |

*Supplementary Table 2*: Summary statistics of the factors. The count for each category is displayed for each factor variable followed by a percentage of the whole.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Transmission Type** | | | | | | | | |
| Automatic | Manual | | | Continuously Variable | | | | Static VAR Compensator |
| 1791 | 684 | | | 109 | | | | 98 |
| 63.0% | 28.6% | | | 4.2% | | | | 4.2% |
| **Fuel** | | | | | | | | |
| Gasoline | | | Diesel | | | Ethanol/Gasoline | | |
| 2112 | | | 76 | | | 122 | | |
| 91.4% | | | 32.9% | | | 5.3% | | |
| **Drive** | | | | | | | | |
| 2wd | | | | | 4wd | | | |
| 1463 | | | | | 847 | | | |
| 63.3% | | | | | 36.7% | | | |
| **Vehicle Class** | | | | | | | | |
| Car | | SUV | | | Truck | | | Van |
| 1579 | | 563 | | | 111 | | | 57 |
| 68.4% | | 24.4% | | | 4.8% | | | 2.5% |
| **SmartWay** | | | | | | | | |
| Yes | | No | | | | | Elite | |
| 476 | | 2168 | | | | | 476 | |
| 17.7% | | 80.1% | | | | | 1.4% | |