**ALY6020 Predictive Analytics**

Dr. Thomas Goulding

Module 2

Building the Car of the Future

Jeff Hackmeister

6/1/2025

**Abstract**

Working with a provided dataset, the goal of this study is to identify attributes that lead to higher gas mileage with the goal of producing a more efficient model in the future. A linear regression model was developed to produce a model and a stepwise selection process was used to determine the most influential variables in fuel efficiency in cars. In the end, a model was produced that achieved an R-Squared of 0.782 on test data with only three variables: Weight (strongest predictor with -0.832 correlation), Model Year, and US Made origin, reducing severe multicollinearity while maintaining predictive power.

**Introduction**

The goal of this study is to determine the most influential variables in determining fuel efficiency in automobiles. By focusing on these elements, manufacturers hope to produce even more fuel-efficient models in the future.

**Step One – Exploratory Data Analysis**

To begin the study, we first need to import the necessary Python packages to effectively clean, prepare, model, and visualize the data.

A screenshot of a computer program

AI-generated content may be incorrect.

With these packages installed, the data can be read in from the provided csv file and explored.

A screenshot of a computer program

AI-generated content may be incorrect.

For this study, our target variable is fuel efficiency, which is measured in miles per gallon, or MPG, in the dataset. As seen above, all but one of the explanatory variables are already in a numeric format. Looking into the Horsepower variable, it appears that while the data type is an object, the values are numeric with 6 null values.

A screenshot of a computer

AI-generated content may be incorrect.

After confirming that these 6 observations are not part of an obvious cluster, it is determined that they can be dropped from the analysis.

Next, plots were developed to see in the interaction of each explanatory variable and the target variable, MPG.

For discreet variable types, box plots were chosen

**A graph with blue and green lines

AI-generated content may be incorrect.**

While scatterplots were used for continuous variables.

**A group of blue and white graphs

AI-generated content may be incorrect.**

These diagnostic plots are used to check for distribution among each variable and to look for obvious outliers in the dataset. And while there are some observations near the extremes of each category, none are determined to be large enough to been considered data entry errors or requiring additional exploration before proceeding with the analysis.

**Step Two – Feature Selection**

To produce the most effective model possible, while avoiding overfitting to the dataset, a clear understanding of the correlation between each variable is needed. To that end, a correlation matrix was produced from the dataset. In this heatmap style visualization, each variable is mapped against all the variables in the dataset and their linear relationship. When comparing with the target variable, MPG, a negative score indicates that a higher value in that variable correlate to lower MPG scores. When comparing between two explanatory variables, a lower value indicates a low level of multicollinearity, meaning the variables tend to explain one another. This is vital in understanding which features to include in the final analysis.

**A red squares with white text

AI-generated content may be incorrect.**

The heatmap reveals critical multicollinearity issues among engine specifications:

* Cylinders and Displacement (0.951): Nearly redundant information
* Displacement and Weight (0.933): Larger engines require heavier vehicles
* All engine variables negatively correlate with MPG (-0.78 to -0.83)

This pattern suggests weight reduction offers the most direct path to efficiency gains. As these are all different measurements of engine capacity, this makes intuitive sense.

To further explore the issues of multicollinearity, the Variance Inflation Factor (VIF) for each variable will be calculated.

A screenshot of a computer

AI-generated content may be incorrect.

This also illustrates high levels of multicollinearity in the data. By eliminating some of the variables from the dataset, a model should be produced with less risk of overfitting the example data and therefore a more useful model in making real-world predictions and recommendations.

To better understand which features to include in the final analysis and to focus on for our manufacture, a stepwise function will be used to evaluate the effectiveness of each variable. This processes in an iterative exploration of each variable in the dataset where variables are added and removed from the analysis and gauged on the improvement or degradation of the model’s performance. The steps are continued until all possible combinations of variables have been tested, at which time, an ideal set of variables is suggested. The results of this process are shown below.

**A screenshot of a computer

AI-generated content may be incorrect.**

At the end of the stepwise process, a combination of 3 variables remained – Weight, Model Year, and US Made. Both the R-Squared and Adjusted R-Squared values for this model (0.819 and 0.818) indicate strong model performance along with the high F-statistic of 585.0. Also, each variable’s p-value came in at <0.001 indicating a high level of significance to the estimation of fuel efficiency. To further examine the results, a series of diagnostic plots were created.

**A graph with a red line

AI-generated content may be incorrect.**

Both the residuals versus fitted values and the residual distribution plot indicated strong performance of the model. The Q-Q Plot shows a right-side or positive skew in the results. This indicates that the model is performing well for most vehicles but may be underestimating the efficiency of a few of the highest efficiency models.

**Step Three – Testing**

With our explanatory variables chosen, the data set is split into a training and a testing set, with 80% of the data to be used for training and the remaining 20% for testing. Then a linear regression model is fit to the training set.

A screenshot of a computer program

AI-generated content may be incorrect.

With the model fit, predictions are made on both the training and testing data and the results are compared.

A computer screen shot of a computer program

AI-generated content may be incorrect.

**A screenshot of a computer program

AI-generated content may be incorrect.**

The test results indicate that the model continues to function well with unknown data. The R-Squared and Root Mean Squared Error (RMSE) values are nearly identical for both the training and test datasets.

When the results of the final test model are compared with the initial model fit of all available variables, the test results hold up favorably with a much lower risk of overfitting.

A graph of a model performance

AI-generated content may be incorrect.

**Conclusions**

By working through all available data in the dataset, the most influential variables were identified for improving fuel efficiency in future vehicles.

Weight, perhaps unsurprisingly, played the biggest role in determining the MPG rating for a vehicle, with every reduction of 1,000 pounds, MPG increased by 6. The other two variables identified in the stepwise selection process, at first glance, are seemingly difficult for manufacturers to incorporate into their design process. Model yeah and foreign versus domestic manufacturing were the other two variables included in the final model. The model year is a valuable indicator that newer technology and increasing consumer demand is leading to year-over-year increases in fuel efficiency in all automobiles. Therefore, a company looking to stand out in the marketplace must set aggressive goals to not just better that their competition today, but also with the understanding the entire industry is moving towards better fuel efficiency every year. Similarly, the 2.1 MPG penalty on US made vehicles is, at first, difficult to include in recommendations to a manufacturer. But it is an indication that vehicles produced outside of the United States tend to be smaller and lighter weight. Therefore, a manufacturer looking to produce more fuel-efficient cars should not simply cut weight from their design (although that will have a big impact on efficiency) but they should also learn from the internation automakers to find out how and where they are cutting weight from their vehicles, as well as in their marketing strategy for smaller vehicles to ensure solid sales performance.

**A graph with different colored squares

AI-generated content may be incorrect.**

**References**

[1] Detecting Multicollinearity with VIF - Python. (2020, August 14). GeeksforGeeks. https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. Springer New York.

[3] Kwok, R. (2021, March 9). *Stepwise Regression Tutorial in Python | Towards Data Science*. Towards Data Science. https://towardsdatascience.com/stepwise-regression-tutorial-in-python-ebf7c782c922/

‌[4] Linear Regression — statsmodels. (n.d.). [Www.statsmodels.org](http://Www.statsmodels.org). https://www.statsmodels.org/stable/regression.html

[5] OpenAI. (2024). ChatGPT (Claude 4) [Large language model]. https://claude.ai

‌[6] Steele, B., Chandler, J., & Reddy, S. (2016). *Algorithms for Data Science*. Springer.