**ALY6020 Predictive Analytics**

Dr. Thomas Goulding

Module 2

Regression Models

Jeff Hackmeister

5/31/2025

**Abstract**

This study utilizes data from the 1985 Ward’s Automotive Yearbook, hosted by the University of California Irvine’s Machine Learning Repository, to develop a linear regression model to predict car prices. The dataset contains various vehicle attributes that were cleaned and manipulated to develop a highly accurate prediction model.

**Introduction**

The dataset contains 26 different variables in relation to 205 different cars. An exploratory data analysis was conducted to understand the nature of the variables and a data cleaning was performed to prepare the data to be fit to the linear regression model. Once the data was appropriately prepared, a model was fit and analyzed for effectiveness.

**Methodology**

To begin, several Python libraries were imported into the project to perform the necessary analysis and visualization of the data.

A screenshot of a computer program

AI-generated content may be incorrect.

With the libraries imported, the csv file provided for this study was loaded. Looking at the structure of the data, there are 10 object variables along with 16 numeric variables. The “CarName” variable contains both manufacture and model name, the decision was made to split this to create a separate “make” variable for further analysis. After cleaning a few misspellings, there were 23 unique manufacturers, and it was decided to group this into 4 categories, luxury, premium, mainstream, and budget.

A screenshot of a computer code

AI-generated content may be incorrect.

The other categorical variables were converted to numeric variables for inclusion in the model, as a new dataframe was created for all numeric values.

With this many explanatory variables, multicollinearity is a major concern. To explore this, a correlation matrix was produced.

A screenshot of a graph

AI-generated content may be incorrect.

Clearly, this level of complexity is inappropriate for our regression model. There are several variables that point to similar characteristics of an automobile and can therefore be dropped. After this cleaning process, a new correlation matrix was produced.

A screenshot of a graph

AI-generated content may be incorrect.

This proved to be a vast improvement, there is still a high correlation between horsepower and curb weight (which was chosen as a representative of overall size) but it is within reason and as horsepower is often cited in retail descriptions of vehicles, it has been deemed important to keep in the model.

After partitioning the target variable and creating an 80/20 training and testing split of the data, a linear regression model was fit.

A screenshot of a computer code

AI-generated content may be incorrect.

The results of the model are as follows.

A screenshot of a computer

AI-generated content may be incorrect.

**Feature Analysis**

There were a few surprises in the results of the model. The strongest predictor in the model was curb weight (t=9.88, p < 0.001) meaning the heavier vehicles were more expensive, but at the same time fuel efficiency, as seen in citympg, was also significant (t=2.397, p=0.018). These would seem to be in contrast with one another. However, given the period of the data collection, not long after the gas shortages of the late 1970’s, it is perhaps less surprising to see fuel efficiency driving value even at a time when larger vehicles were still en vouge.

A bar plot was also produced to show the impact of each variable on price. The “mainstream” category was used as the baseline for make, so the impacts of premium, luxury and budget should be interpreted as their impact versus mainstream. The large impact of these categories bolsters the decision to create these variables from the original dataset.

A graph with blue squares

AI-generated content may be incorrect.

**Model Analysis**

The R-squared result of 0.858 indicates a very strong model in that 85.8% of price variation can be explained by this model. The lower adjusted R-squared value of 0.851. A plot of residuals vs fitted values was produced.

A diagram of a graph

AI-generated content may be incorrect.

This compares the predicted car values (fitted) vs the prediction errors (residuals). The plot illustrates the strength of the model, especially in the lower price points. As the price goes up, the predictive strength weakens, especially around the $30,000 price point. Overall, the results point to an effective model.

**Conclusion**

The results of the model show strong predictive capabilities from the data set. Next steps in exploration would be to obtain a similar dataset for different years, ideally 1995, 2005, 2015, and 2025 to measure the change in importance of these variables and their impact on car prices.

**References**

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

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

‌[3] Steele, B., Chandler, J., & Reddy, S. (2016). *Algorithms for Data Science*. Springer.

[4] *UCI Machine Learning Repository*. (n.d.). Archive.ics.uci.edu. https://archive.ics.uci.edu/dataset/10/automobile