Using Multiple Linear Regression (MLR) to predict the real cost of a car model

Tien Anh Nguyen

Abstract—In this paper, an analysis of basic car characteristics is taken into account to predict the real price of different automobile models. Multiple linear regression (MLR) analysis was performed on the data using structural equation modeling with JMP 17. My methodology is divided into three main steps: the first uses various statistical analysis techniques to evaluate and preprocess the data and collected variables; the second involves choosing the most significant variables using multiple methods. The final phase uses RMSE, AICc, BIC, Mallow's Cp, and Adjusted R^2 to compare the outcome of many MLR models built using the chosen variables. The collected findings indicate that the model produced with variables chosen using the Stepwise Selection approach performs better than the models utilizing other approaches, having the lowest AICc, RMSE, and highest Adjusted R^2. In the results, a reasonable regression model acquired a remarkable ability to predict the price of car models.

Keywords— Real Value Estimation, Statistics, Multiple Linear Regression, Economy

1. Introduction

In our constantly evolving world, owning a car can be important for many people, as it provides a convenient and flexible mode of transportation. However, an automobile is an investment that lasts a lifetime; therefore, consumers need to make rational financial decisions. This study will facilitate potential customers in making better educated decisions that fit their budget and lifestyle, by allowing them to take into account the potential purchase price. Estimating a car model's true cost is also crucial for manufacturers in identifying the precise cost, which makes it possible to identify cost drivers for optimization.

This study will build a linear regression using several independent variables to find important aspects influencing a car model's market value and predict the real cost of future car models. Linear regression analysis is a widely applied statistical method for modeling both time-series and cross-sectional data. It allows for the identification and characterization of relationships between different variables. [1] For this study, car models from various brands will be examined within their basic properties, and the data will be gathered from a dataset by Joan Pau Gutiérrez Pascual through Kaggle. [2] The dataset consists of 425 vehicles and their basic characteristics, including such: type of car (sport, SUV, wagon, minivan, pickup); horsepower; number of cylinders; miles per gallon; physical size (length, width, weight), and wheelbase.

This study is outlined as follows: Section II covers relevant works, Section III covers methodology, and Section IV includes findings and commentary. The paper's conclusion is provided in Section V, along with recommendations for more study and advancements. Finally, Section VI expresses my gratitude to the individuals who have assisted me in accomplishing this work.

2. Literature review

Noor and Jan [3] developed a multivariate linear regression model to forecast car prices. The authors were able to attain an R^2 of 98% using the configuration they were provided. Muti and Yildiz [4] also looked at the linear regression model's predictive power for used automobile prices. It was noted that the model's prediction success had an R^2 score of 73%.

While the above-mentioned authors only used linear regression as their main methodology, Pudaruth [5] attempts to achieve the same goal via different machine learning methods. The pricing from these approaches is quite comparable, as seen by the comparison of the forecast results from various methodologies. Nevertheless, it was discovered that the Naïve Bayes approach and the decision tree algorithm could not categorize and forecast numerical values.

Multiple Linear Regression has also been utilized for a long time by researchers for the analysis and prediction of real estate prices. Kaushal and Shankar's [6] paper's goal is to forecast home values using a variety of factors. They performed the prediction for this paper using the multivariate linear regression model. They evaluated their model's accuracy against that of other machine learning models, including decision tree regressors, Lasso, LassoCV, Ridge, and Ridge. With an R^2 value of 84.5%, multivariate linear regression outperforms the others, which means that using the multiple linear regression model to forecast house prices gets the most appropriate and reliable findings.

Another application of Multiple Linear Regression is in the field of education. Aissaoui et al. [7] presented an MLR application to construct a model to predict the performance of pupils. This study stands out because, while most of them create regression models using all of the features of the students, they have developed a multivariate linear regression model that only includes the most significant factors. The findings demonstrate that the model produced with variables chosen from the Multivariate Adaptive Regression Splines approach performs better than the other models.

3. Methodology

3.1 Description of the problem

In many cases, the objective of statistical analysis within research is to delineate the relationships between multiple variables [1] Through establishing a linear relationship based on the observed data, multiple linear regression "aims to model the relationship between two or more independent variables and a dependent variable". [7] The theoretical basis of MLR posits that each unit change in the independent variable leads to a consistent change in the dependent variable. This is how a multiple linear regression model may be expressed [8]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

where:

- Y: the predicted value of the dependent variable
- β_0: the y – intercept, also referred to as the value of y when all the other parameters are equal to 0
- X for i = 1,2,..,p, are the independent variables, also referred to as covariates
- B for i = 1,2,…,p are the regression coefficients, also known as the change in Y for a unit change in Xi (given all variables stay the same)
- ε is an error term that represents the difference in the model and an observed value of Y

Table 1 will list all the dataset variables included in this paper.

Table 1. Dataset variables

Attribute	Meaning	Attribute	Meaning
Name	Vehicle name and model	Rear_wheel	Is it a RWD? (binary: FALSE or TRUE)
Sports_car	Is it a sports car? (binary: FALSE or TRUE)	Msrp	Real cost (numeric)
Suv	Is it a SUV? (binary: FALSE or TRUE)	Eng_size	Engine size in Liters (numeric)
Wagon	Is it a wagon? (binary: FALSE or TRUE)	Ncyl	Number of cylinders (numeric: from 4 to 8)
Minivan	Is it a minivan? (binary: FALSE or TRUE)	Horsepwr	Horsepower of the car (numeric)
Pickup	Is it a pickup? (binary: FALSE or TRUE)	City_mpg	City consumption (in miles per gallon)
All_Wheel	Is it a 4x4? (binary: FALSE or TRUE)	Hwy_mpg	Highway consumption (in miles per gallon)
Weight	Weight (in pounds) (numeric)	Wheel_base	Wheel base (in inches) (numeric)
Length	Length (in inches) (numeric)	Width	Width (in inches) (numeric)

My methodology can be briefly explained via the following diagram (Fig. 1):

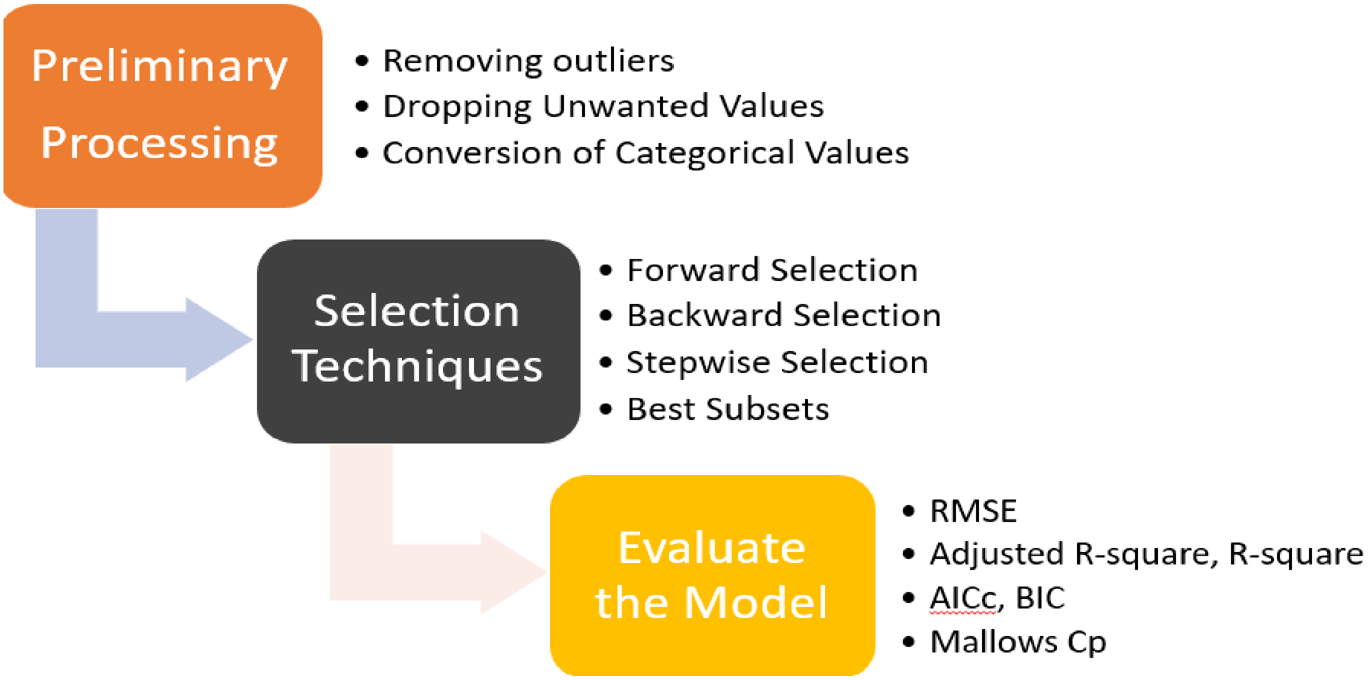

Fig. 1. My methodology

The following subsections will further elaborate on the different steps carried out in my analysis.

3.2 Preliminary Processing

Preprocessing the data and verifying the Multiple Linear Regression method's assumptions necessitate a preliminary analysis before identifying the most crucial variables. This step is of significant importance because if data is not cleaned thoroughly before analysis, the entire data flow becomes "garbage in, garbage out" [9] The preprocessed dataset is prepared for model construction. For training and testing, the division ratios in this paper are 80% and 20% of the total dataset, respectively. The values in the training and testing datasets are randomly selected. Qualitative data is turned into dummy variables, with the presence in a category taking a 1, and the absence is indicated by a 0.

Outliers Analysis and Removal

As outliers are generally regarded as values that are exceptionally far from the mass of the data, they can create many unfavorable consequences, such as distorting the outcome of the analysis. In this study, scatterplots are used to find highly skewed and uncommon values. Although outliers in this study are natural, as they reflect the true variability of the data, they still contribute to the distortion of the final result; hence, they must be removed. On top of that, in this particular dataset, there is some data not reported and without values; they must also be eliminated. Originally, the dataset had 425 unique values, but after cleaning, this paper will only analyze 380 values.

Testing the null hypothesis

Null hypothesis testing is "a formal approach to deciding between two interpretations of a statistical relationship in a sample". [10] The null hypothesis, often represented by the symbol Ho, is one explanation. According to the hypothesis, any association discovered in the sample is only the result of sampling error, and there is no correlation in the population. [10] For every variable, I have constructed a basic regression model with the real value of a car in this stage. To reduce the possibility of a Type I error, I decided to keep the variable if the p-value was less than 0.05.

3.3 Features Selection

Forward Selection

The variable that has the highest correlation with the dependent variable is chosen first to be incorporated into the model using a forward selection process. Following selection, the variable is evaluated using a predetermined set of criteria. If the first variable selected meets the inclusion criteria – that is, if the variables that are not part of the equation are selected based on their statistics – then the forward selection method proceeds. The procedure concludes when no more variables meet the requirements for entrance. [11] While Mallow's C_p or AICc is usually utilized as a stopping rule, in this paper, a p-value threshold of 0.05 will decide which variable can enter the model. However, this method is not without flaws. A difficulty with forward selection is that it has trouble with multiple testing, which can lead to the selection of a lot of irrelevant variables. Certain variables may have partial correlations with other variables, but not with the goal. This suppresses non-relevant errors in those variables, strengthening the model. [12]

Backward Selection

This technique is quite similar to the Forward Selection algorithm in principle, but it differs in that it removes variables one at a time depending on the rise in p-value after starting from the whole model (whenever it is possible to estimate the full model). [13] Again, a p-value of 0.05 acts as a hurdle the variable must overcome to not be rejected from the model. Backward Selection has two main drawbacks. Firstly, when additional variables are eliminated from the model, certain variables may no longer be able to minimize error due to overlap or interaction in their capacity to explain variation. Secondly, Backward elimination is rigid because once a feature variable leaves the model, it is never added back. A variable that was included at the beginning of the model may eventually surpass the stopping criterion as other variables are eliminated, but it will always remain outside of it. [14]

Stepwise Selection

The stepwise selection method differs from the methods above in that, following the entry of a variable, all previously entered variables are analyzed to see whether any should be eliminated based on the specified removal criteria. Every step of the stepwise process involves investigating the "least useful variable currently in the equation." One factor that could have been the best contender for admission early on in the process may become unnecessary later on due to the correlations it now has with other variables in the regression. [11] This method avoids the errors of the previous techniques as it may be applied to the evaluation of variables, enabling the investigation of the traits of variables serving as predictors in various models.

Best Subsets

Regression analysis using best subsets regression is an experimental model development technique. It contrasts every model that might be developed using a given set of predictors. For k predictors, the complete set of models of any size sums up to 2^k -1, whereas the best-subset models total the same as the number of predictors. [15] For example, all subsets of four variables have fifteen possible regression models; however, only four optimal models – the best model with one variable, the best model with three variables, and the best model with all variables – are found for each subset based on certain criteria. Nevertheless, it is not always practical to employ best-subsets procedures, since they need a lot of computing power to produce all conceivable regressions, especially when the number of predictors is large. [15]

4. Results

4.1 Models building

I will build several linear regression models using the variables that were selected as the most important results of the methods in the preceding section. Next, I will use the following set of criteria to compare the models' results. The model that can choose the most correct variables has the highest performance accuracy. Table 2 shows the built model using each different method:

Table 2. Models' evaluation metrics

Features Selection	Variables chosen	The built model
Forward Selection	all_wheel, rear_wheel, horsepwr	-9026.47 + 4846.22*all_wheel + 7175.88*rear_wheel + 181.04*horsepwr

Backward Selection	sports_car, all_wheel, rear_wheel, horsepwr, hwy_mpg, width	28268.58 + 5624.77 * sports_car + 3259.85 * all_wheel + 7467.43 * rear_wheel + 163.89 * horsepwr + 849.69 * hwy_mpg + 13.76 * weight – 1480.74 * width
Stepwise Selection	sports_car, all_wheel, rear_wheel, eng_size, ncyl, horsepwr, hwy_mpg, weight, width	20977.41 + 6218.74 * sports_car + 3117.89 * all_wheel + 6476.23 * rear_wheel + 2820.25 * eng_size + 2368.30 * ncyl + 157.93 * horsepwr + 844.63 * hwy_mpg + 13.54 * weight – 1412.54 * width
Best Subsets	sports_car, all_wheel, rear_wheel, eng_size, ncyl, horsepwr, hwy_mpg, weight, width	30538.05 + 6209.08 * sports_car + 3418.84 * all_wheel + 6697.56 * rear_wheel + 1193.93 * ncyl + 152 * horsepwr + 881.98 * hwy_mpg + 13.18 * weight – 1555.52 * width

4.2 Evaluating the build models' performance

Akaike Information Criterion Corrected (AICc)

AIC model selection criterion attempts to quantify the relative quality of several models while penalizing for model complexity. [16] AIC achieves a balance between model fit and complexity by taking into account both the probability and the parameters' numbers. It promotes models that minimize complexity, avoiding overfitting and lowering the possibility of collecting noise or extraneous characteristics in the data while yet providing a good match for the data. However, there is a significant chance that AIC will choose models with excessive parameters when the sample size is small, or that AIC will overfit. Hence, AICc was created as a solution to this possible overfitting; it is "AIC with a correction for small sample sizes". The lower the AICc, the better fit the model is compared to others. [17]. AICc can be written as below:

$$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1}$$

where:

- N is the sample size
- K is the number of parameters

Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC), similar to AIC, is another model selection criterion that considers both model fit and complexity. In contrast to AIC, BIC offers a larger penalty for model complexity and is based on Bayesian principles. [18] Similar to the AICc, the lower the BIC, the better the model is. The following is the formula for BIC:

$$BIC = -2 * log\ (L)\ + k * log\ (n)$$

where:

- L represents the maximized likelihood of the model, which is commonly measured as the sum of squared errors (SSE)
- N is the sample size
- K is the number of parameters

Adjusted R^2

Adjusted R^2 is a corrected "goodness-of-fit" (or model correctness) measure. A problem with the R^2 is that it always rises as additional effects are added to it. The Adjusted R^2 considers this overestimation. If a variable has no beneficial effect on the model, the adjusted R^2 may go down [19] The more closely the model fits the target field's values, the closer its Adjusted R^2 is to 1. A number that is closer to 0, on the other hand, denotes a subpar model with no predictive value. Adjusted R^2 can be written as below:

$$Adj\ R^2 = 1 - \frac{(1-R^2)(n-1)}{(n-k-1)}$$

where:

- N is the sample size
- K is the number of parameters
- R^2 is the proportion of variation in the target variable that is explained by the model.

Mallow's Cp

Mallow's Cp assists in finding a crucial equilibrium regarding the quantity of the model's predictors. Mallows' Cp compares the model's overall accuracy and bias against models that just incorporate some of the predictors. The constant (p) plus the number of predictors in the model make up Mallows' Cp. The lower the Mallow's Cp is to the parameter, the more accurate the model is. When Mallow's Cp value is low, it means that the model has little variance and may be used to estimate the actual regression coefficients and forecast the responses in the future. [20] Mallow's Cp can be calculated as:

$$Cp = \frac{RSSp}{S^2} - N + 2(P + 1)$$

where:

- RSSp: The residual sum of squares for a model with *p* predictor variables
- S^2: The residual mean square for the model (estimated by MSE)
- N: The sample size
- P: The number of predictor variables

Root Mean Square Error (RMSE)

RMSE is usually understood as "the square root of the mean of the square of all errors".

Because it is thought to be a better "general-purpose error metric for numerical predictions", RMSE is commonly employed. [21] RMSE is a useful statistic for evaluating prediction errors between models or model setups since it is size-dependent. [21] The model is better if the RMSE is smaller. RMSE can be interpreted as below:

$$RMSE = \sqrt{\frac{\sum \ (y_x - \hat{y}_x)^2}{N - P}}$$

where:

- y_x is the actual value for the x^th observation.
- $\hat{y}_x$ is the predicted value for the x^th observation.
- N is the number of observations.
- P is the number of parameter estimates, including the constant.

Table 3 presents these values in each model built.

Table 3. Models' evaluation metrics

Features Selection	RMSE	R^2	Adj R^2	Mallows Cp	P	AICc	BIC
Best Subsets	8074.11	0.7398	0.7328	9	9	6406.638	6443.163
Backward Selection	8107.271	0.7367	0.7306	8	8	6408.045	6440.981
Forward Selection	8707.576	0.6923	0.6892	4	4	6447.578	6466.013
Stepwise Selection	8039.109	0.7429	0.7351	11.5245	10	6405.09	6445.19

The following table shows that the model conducted via Stepwise Selection has the lowest RMSE, AICc, and the highest adjusted R-squared and R-squared. Even though the BIC and Mallow's Cp isn't the lowest, it is still in an acceptance range compared to other models; hence, this model is the most performant in predicting the cost of a certain automobile model. It can be concluded that the best method for building a regression model is the Stepwise Selection method, and the attributes that affect the price more than others are: sports_car, all_wheel, rear_wheel, eng_size, ncyl, horsepwr, hwy_mpg, weight, and width. Except for the car's width, all the factors exhibit positive correlation coefficients, indicating that a unit increase in the width of a car will decrease the final car, keeping all the other explanatory variables unchanged in the model. This model has achieved great complexity while still managing to avoid overfitting.

Figure 2 conveys the relations between the actual vs prediction price of the car models on test data of the model conducted via Stepwise Selection. The blue label indicates the predicted values of the cars, and the orange label indicates the real price of the cars. The graph of the two values reveals that the values are closely related to each other, indicating that the model is a relatively great example of modeling the price of a vehicle.

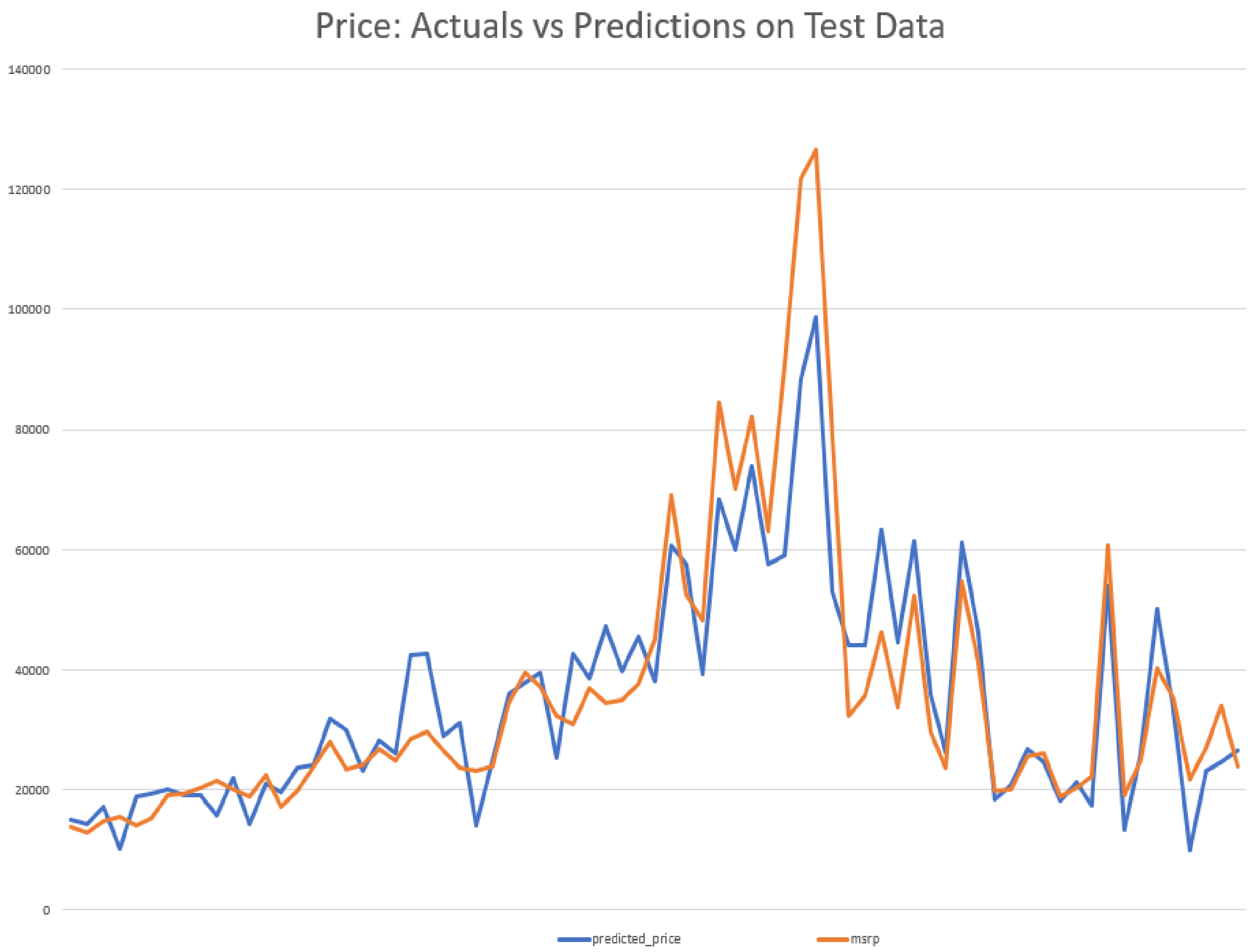

Fig. 2. Relation of Actuals vs Predictions in Stepwise Selection

5. Conclusion

Finding the elements that influence an automobile's real price is an exciting undertaking since it will enable both consumers and manufacturers to make the most rational decision. Within this framework, I have put forth a process that thoroughly assesses the features of a car and determines which are most crucial for constructing a prediction model. Our approach entails employing several techniques to identify the most significant variables, which are then used to construct various multiple linear regression models. I have discovered that the model developed utilizing the Stepwise Selection approach is the most performant after evaluating the performances of the other models. This study has two main limitations. First, the reliance on a community-produced dataset makes the study's quality heavily correlated with the quality and accuracy of the data. Second, the dataset size is not insufficient but can include more car models to be a more accurate representative sample of the population.

In future research, I would like to employ a dataset that documents even more automobile models and with more comprehensive data. Additionally, I would also conduct other regression and classification approaches besides Multiple Linear Regression.

6. Acknowledgement

Without the encouragement and assistance of Yixi Kennedy, my mentor from Polygence, this paper could not have been finished. Her unconditional guidance, encouragement, and critique played a vital role in the completion of this research project. She also gave me invaluable instructions in my discovery of the vast knowledge of data science and machine learning. I would also like to thank Brandon Foltz for his abundantly clear lessons about Linear Regression and Multiple Linear Regression uploaded on his YouTube channel. Lastly, I want to express my gratitude to my family for their care and support during my study.

References

[1] Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis: part 14 of a

series on evaluation of scientific publications. *Deutsches Arzteblatt international*, *107*(44), 776–782. https://doi.org/10.3238/arztebl.2010.0776

[2] Pascual, J. (2022, December 8). *Basic car characteristics*. Kaggle. https://www.kaggle.com/datasets/joanpau/cars-df

[3] Noor, Kanwal & Jan, Sadaqat. (2017). Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications. 167. 27-31. 10.5120/ijca2017914373.

[4] MUTİ, S., & YILDIZ, K. (2023). Using linear regression for used car price prediction. *International Journal of Computational and Experimental Science and Engineering*, *9*(1), 11–16. https://doi.org/10.22399/ijcesen.1070505

[5] Pudaruth, Sameerchand. (2014). Predicting the Price of Used Cars using Machine Learning Techniques. International Journal of Information & Computation Technology. 4. 753-764.

[6] Kaushal, Anirudh and Shankar, Achyut, House Price Prediction Using Multiple Linear Regression (April 25, 2021). Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021, Available at SSRN: https://ssrn.com/abstract=3833734 or http://dx.doi.org/10.2139/ssrn.3833734

[7] Aissaoui, Ouafae & Madani, Yasser & Oughdir, Lahcen & Dakkak, Ahmed & EL ALLIOUI, Youssouf. (2020). A Multiple Linear Regression-Based Approach to Predict Student Performance. 10.1007/978-3-030-36653-7_2.

[8] Dietrich, D., Heller, R., Yang, B., EMC Education Services: Data science and big data analytics: discovering, analyzing, visualizing and presenting data.

[9] Christine P. Chai (2020) The Importance of Data Cleaning: Three Visualization Examples, CHANCE, 33:1, 4-9, DOI: 10.1080/09332480.2020.1726112

[10] Paul C. Price, R. S. J. (2017, August 21). *13.1 UNDERSTANDING NULL HYPOTHESIS TESTING*. Research methods in psychology. https://opentext.wsu.edu/carriecuttler/chapter/13-1-understanding-null-hypothesis-testing/

[11] Walczak, Beata & Massart, D.. (2000). Chapter 15 Calibration in wavelet domain. Data Handling in Science and Technology. 22. 323-349. 10.1016/S0922-3487(00)80040-4.

[12] Borboudakis, Giorgos & Tsamardinos, Ioannis. (2017). Forward-Backward Selection with Early Dropping.

[13] Narisetty, Naveen. (2020). Bayesian model selection for high-dimensional data. 10.1016/bs.host.2019.08.001.

[14] Brandon Foltz. (2022). *Statistics 101: Multiple Regression, Backward Elimination* [Video]. YouTube. https://www.youtube.com/watch?v=pv4SBxyynxc

[15] Brooks, G. P., & Ruengvirayudh, P. (2016). Best-subset selection criteria for multiple linear regression. General Linear Model Journal, 42(2), 14-25.

[16] Akakike, H. (1974). “A new look at the statistical model identification”. IEEE Transactions on Automatic Control. 19 (6): 716 – 723,

[17] *McQuarrie, A. D. R.; Tsai, C.-L. (1998), Regression and Time Series Model Selection, World Scientific*.

[18] Gideon Schwarz. "Estimating the Dimension of a Model." Ann. Statist. 6 (2) 461 - 464, March, 1978. https://doi.org/10.1214/aos/1176344136

[19] *Adjusted R squared*. IBM. (2024, January 18). https://www.ibm.com/docs/en/cognos-analytics/12.0.0?topic=terms-adjusted-r-squared

[20] *What is Mallows’ cp?* Minitab. (n.d.). https://support.minitab.com/en-us/minitab/help-and-how-to/statistical-modeling/regression/supporting-topics/goodness-of-fit-statistics/what-is-

mallows-cp/

[21] David Christie, Simon P. Neill, 8.09 - Measuring and Observing the Ocean Renewable Energy Resource, Editor(s): Trevor M. Letcher, Comprehensive Renewable Energy (Second Edition), Elsevier, 2022, Pages 149-175, ISBN 9780128197349, https://doi.org/10.1016/B978-0-12-819727-1.00083-2.
(https://www.sciencedirect.com/science/article/pii/B978012819727100