Analyzing Factors Affecting Electric Vehicle Prices Using the Weighted Least Squares (WLS) Model

Tubagus Zakki I.A
Computer Science and Statistics
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
tubagus.ahmad001@binus.ac.id

Joevan Alezka
Computer Science and Statistics
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
joevan.alezka@binus.ac.id

Farhan Armandy Rasyid
Computer Science and Statistics
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
farhan.rasyid@binus.ac.id

The automobile industry has become a major player in both the global economy and the world of Research and Development (R&D). With the constant advancement of technology, vehicles are now equipped with features that prioritize the safety of both passengers and pedestrians. This has led to an increase in the number of vehicles on the road, providing us with the convenience of quick and comfortable travel. However, this progress has come at a cost.

I. Background

In recent years, sales of electric vehicles (EVs) have increased rapidly worldwide. EV sales in 2022 totaled 10.52 million units, up 55.46% from 6.76 million units the year before, according to EV Volumes statistics. Stricter pollution rules and rising consumer awareness of environmental issues are driving the automobile industry's shift toward low-emission vehicles, which is reflected in this growth.

The use of electric vehicles has advanced significantly in Indonesia as well. EV sales in 2022 were 15,437 units, a startling 383.46% increase over the 3,193 units sold the year before. Rising environmental consciousness, government incentives, and the introduction of more sophisticated new models were the main drivers of this expansion.

Electric vehicle pricing is still difficult to determine, though, because of a number of impacting elements, including range, maximum speed, and energy efficiency. Non-normal distributions and non-constant variance are common in price analysis data, which can compromise the precision of prediction models. Weighted least squares (WLS), robust regression, and linear regression are some of the regression techniques whose performance is compared in this study in order to determine which model is best for assessing and forecasting the pricing of electric vehicles.

II. Problem Statement

1. How do energy efficiency, range, maximum speed, and acceleration influence the price of electric vehicles?
2. Which regression method, linear regression, robust regression, or weighted least squares, proves to be the most effective in predicting electric vehicle prices based on the dataset utilized?

III. Objectives

1. To analyze the relationship between energy efficiency, maximum speed, and range with the price of electric vehicles.
2. To compare the performance of linear regression, robust regression, and weighted least squares (WLS) in predicting electric vehicle prices and identify the most optimal method.

IV. Significance of the Study

This study aims to provide valuable insights for both manufacturers and consumers regarding the pricing of electric vehicles and the factors influencing their price and quality, based on accurate data analysis. By examining key determinants such as energy efficiency, range, and performance characteristics, the research seeks to offer data-driven recommendations that support the adoption of electric vehicles. These insights aim to enhance understanding of the critical factors shaping electric vehicle prices, thereby aiding stakeholders in making informed decisions to promote sustainable transportation solutions.

V. Dataset

The dataset utilized in this study consists of variables categorized into two groups: independent variables and a dependent variable. The independent variables represent the key features expected to influence the price of electric vehicles, while the dependent variable corresponds to the target outcome—the price of the electric vehicle. Below is an explanation of each variable included in the analysis:

A. Independent Variable

Energy Efficiency (kWh/100 km): This variable measures the energy consumption of the vehicle, indicating its efficiency over a distance of 100 kilometers. Lower values signify higher energy efficiency.

Battery & Fast-Charge Capability: This variable captures the specifications of the vehicle's battery system, including its capacity and the availability of fast-charging technology, which directly affects usability and convenience.

Maximum Speed (km/h): This represents the highest speed the vehicle can achieve, reflecting its performance potential.

Acceleration (seconds from 0–100 km/h): This variable measures the time taken by the vehicle to accelerate from 0 to 100 kilometers per hour, serving as an indicator of its power and performance.

Range (km): This variable denotes the maximum distance the vehicle can travel on a full charge, which is a critical factor for consumers considering electric vehicles.

B. Dependent Variable

Electric Vehicle Price (in German currency): This represents the monetary cost of the electric vehicle, which serves as the target variable in this study.

VI. Method

In this section, we outline the methodology employed in this study to analyze the factors influencing electric vehicle prices. The analysis involves comparing three regression methods: Ordinary Least Squares (OLS), Robust Regression, and Weighted Least Squares (WLS). Each method is evaluated based on its predictive performance to determine the most suitable approach for the dataset utilized. Additionally, feature scaling is applied to ensure that the input variables are standardized, thereby improving the comparability and performance of the regression models.

A. Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS) is one of the most widely used methods for estimating the parameters of a linear regression model. It works by minimizing the sum of the squared residuals, the differences between observed and predicted values. OLS assumes that the errors in the model satisfy three key assumptions: normality, homoscedasticity, and no multicollinearity. These assumptions are fundamental to the reliability of the OLS estimates.

To ensure the validity of the OLS regression results, it is important to check the assumptions of **identically**, **independence**, **normality**, and **multicollinearity**. These can be assessed through separate tests, which evaluates the following:

a. Identically

The variance of the residuals should be constant across all levels of the independent variables.

b. Independence

Each observation in the dataset should be independent of the others. This means there should be no autocorrelation (i.e., no relationship between the residuals of one observation and another). Autocorrelation can lead to biased estimates of the regression coefficients and affect the reliability of the model.

c. Normality

The residuals should follow a normal distribution.

d. No Multicollinearity

The independent variables should not be highly correlated with each other, as multicollinearity can distort the estimates and reduce the model's reliability.

If these assumptions hold, OLS provides unbiased and efficient estimators. However, if the assumptions are violated, the reliability of OLS estimates may be compromised.

B. Robust

Robust regression is an alternative to OLS that aims to provide reliable estimates even when the assumptions of OLS are violated. It is particularly useful when there are outliers or when the data violates the assumption of homoscedasticity or normality. Robust regression uses a weighting mechanism to reduce the influence of outliers, making it more robust to deviations from the classical assumptions of linear regression.

Robust regression is used when there are violations in the OLS assumptions, particularly when the data includes outliers or heteroscedasticity (non-constant variance of errors), or when the residuals fail the normality test. Unlike OLS, which can be heavily influenced by outliers, robust regression minimizes their effect and provides more reliable estimates.

The main advantage of using robust regression is its ability to handle datasets where the residuals do not meet the i.i.d. assumptions, providing more accurate estimates in the presence of outliers or non-homoscedasticity.

C. Weighted Least Squares (WLS)

Weighted Least Squares (WLS) is a variation of OLS that accounts for heteroscedasticity by assigning weights to the observations. The main principle behind WLS is to give less weight to observations with higher variance, thereby reducing their influence on the regression model. WLS is particularly useful when the assumption of homoscedasticity (constant variance of residuals) is violated.

WLS is used when there is evidence of heteroscedasticity, as indicated by the failure of the normality, heteroscedasticity, or multicollinearity tests. If the errors have non-constant variance, OLS estimates may become inefficient and biased. In such cases, WLS allows for more accurate modeling by adjusting for the varying importance of each observation.

The main benefit of WLS is its ability to provide efficient and unbiased estimates in the presence of heteroscedasticity, making it a suitable method when the variance of the errors is not constant across observations.

VII. Result

A. Ordinary Least Squares (OLS)

Parameter	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.2e-10	3.068e-02	0.000	1.00000
Battery	0.1309	2.072e-01	0.632	0.52798
Efficiency	0.2894	1.081e-01	2.678	0.00781
Fast_charge	0.1245	5.453e-02	2.283	0.02315
Range	0.2838	2.091e-02	0.136	0.89211
Top_speed	0.7423	7.500e-02	9.897	< 2e-16
acceleration 0 to 100	0.1578	6.940e-02	2.274	0.02366
Residual Std. Error	0.5375	-	-	-
R-squared	0.7167	-	-	-
Adjusted R-squared	0.711	-	-	-

The output presented alongside shows the factors that influence the price of electric vehicles. Among these factors, efficiency, fast charge, and top speed are found to have a significant impact on the price, while battery and range do not appear to affect it.

Additionally, the influence of each variable can be assessed by looking at the estimated values for each factor. For instance, the fast charge variable has an impact of 1.245e^-01, meaning that an increase in the fast charge category would result in a corresponding change in the price, assuming all other categories remain constant. This estimation provides a quantitative understanding of how each variable influences the price of electric vehicles.

The R-squared value of 0.7167 indicates that the predictor variables (x) in the dataset can explain 71.67% of the variation in the dependent variable (y), while the remaining 28.33% is attributed to other factors not included in the dataset. The p-value is 2.2e-16, which is smaller than the significance level of 0.05, suggesting a statistically significant relationship between the predictor variables (x) and the dependent variable (y).

a. Identically, independence, normality, and multicollinearity test.

i. Breusch-Pagan Test(Identically)

The hypothesis test for homoscedasticity is conducted by testing the null hypothesis that the errors have constant variance (homoscedasticity), against the alternative hypothesis that the errors do not have constant variance (heteroscedasticity). The rejection region for this test is defined as a p-value less than the significance level of 0.05. In this case, the obtained p-value is 4.838e-14, which is less than 0.05. Therefore, we reject the null hypothesis, indicating that the errors do not have constant variance. This suggests that the assumption of homoscedasticity is not satisfied in the model.

ii. Durbin-Watson Test(Independence)

The hypothesis test for autocorrelation of errors is conducted by testing the null hypothesis that the errors are not autocorrelated, against the alternative hypothesis that the errors are autocorrelated. The rejection region for this test is defined as a p-value less than the significance level of 0.05. In this case, the obtained p-value is 0.2289, which is greater than 0.05. Therefore, we fail to reject the null hypothesis, indicating that the errors are not autocorrelated. This suggests that the assumption of no autocorrelation is satisfied in the model.

iii. Kolmogorov-Smirnov Test (Normality)

The hypothesis test for normality of errors is conducted by testing the null hypothesis that the errors follow a normal distribution, against the alternative hypothesis that the errors do not follow a normal distribution. The rejection region for this test is defined as a p-value less than the significance level of 0.05. In this case, the obtained p-value is 0.0000002128, which is less than 0.05. Therefore, we reject the null hypothesis, indicating that the errors do not follow a normal distribution. This suggests that the assumption of normality is not satisfied in the model.

B. Robust

Parameter	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.23060	0.04356	-5.294	2.32e-07
Battery	0.38697	0.14293	2.707	0.00717
Efficiency	0.12756	0.05988	2.163	0.03136
Fast_charge	0.04101	0.03718	1.103	0.27088
Range	-0.16493	0.13993	-1.179	0.23945
Top_speed	0.22738	0.10459	2.174	0.03049*
acceleration 0 to 100	0.01522	0.04024	0.378	0.70554
Residual Std. Error	0.2387	-	-	-
R-squared	0.7512	-	-	-
Adjusted R-squared	0.7462	-	-	-

The information derived from the above output highlights the relationship between the predictors (X) and the dependent variable (Y), the prediction error compared to reality, the significance of variables, and the explanatory power of the model.

From the robust regression results, the prediction error (Residual Standard Error) is approximately **0.2**, indicating that the model predictions are reasonably close to the actual data. The **R-squared** value of **75%** implies that the model explains 75% of the variation in the dependent variable (Y) through the predictors included in the dataset. However, 25% of the variation remains unexplained, suggesting that other factors affecting Y are not captured in this dataset.

Regarding the predictors that significantly influence Y, the robust regression identifies **Battery** and **Fast Charge** as impactful variables. These predictors show statistically significant t-values and p-values, demonstrating their strong contribution to the model.

Robust regression is particularly beneficial when handling datasets that may contain outliers or data points with high leverage. Unlike traditional OLS regression, robust regression minimizes the influence of these outliers on the model, leading to more reliable estimates. This is why the residual error is lower in robust regression compared to OLS in many cases.

Moreover, this approach provides a clearer understanding of the relationship between variables under real-world data conditions, which often contain noise or non-normal distributions. By using robust regression, we can derive insights that are not overly affected by extreme values, ensuring that the conclusions about the significant variables are reliable and generalizable.

This model also emphasizes that **efficiency** and **top speed** exhibit borderline significance, which might suggest a partial influence under specific scenarios or contexts. Nonetheless, these variables may require further investigation to determine their exact relationship with Y.

C. WLS

Parameter	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.07440	0.02892	-2.573	0.010577
Battery	0.72040	0.14487	4.973	1.11e-06
Efficiency	-0.01276	0.05568	-0.229	0.818902
Fast_charge	0.15167	0.03976	3.815	0.000166
Range	-0.70951	0.12968	-5.471	9.44e-08
Top_speed	0.39512	0.05696	6.937	2.47e-11
acceleration 0 to 100	-0.03550	0.01503	-2.361	0.018852
Residual Std. Error	1.246	-	-	-
R-squared	0.841	-	-	-
Adjusted R-squared	0.8378	-	-	-

The output for the Weighted Least Squares (WLS) regression demonstrates a strong model fit, indicated by a low residual standard error, a high number of impactful predictors (X), and a high R-squared value.

The **Residual Standard Error** is relatively low, which suggests that the model predictions closely align with the actual observed data. This means the differences between predicted and actual values are minimal, indicating that the dataset is accurate and well-suited for the regression analysis. A low residual error reflects the precision of the WLS model in capturing the relationship between the independent variables and the dependent variable.

The model also highlights a significant number of predictors with strong influence on the dependent variable (Y). This suggests that the variation in Y can be largely explained by the variables included in the dataset, providing a comprehensive understanding of the factors that impact Y.

The **R-squared value** of **0.84** further confirms the model's robustness, showing that 84% of the variance in Y is explained by the predictors in the dataset. The remaining 16% represents unexplained variance, which could be attributed to factors not included in the dataset. A high R-squared value like this signifies that the predictors have a strong explanatory power over Y, making the model highly effective.

WLS regression is particularly beneficial in datasets where heteroscedasticity (non-constant variance of residuals) is present. Unlike OLS regression, which assumes constant variance, WLS assigns weights to each observation based on the variance of its residuals. Observations with smaller residual variance are given more weight, leading to more reliable and efficient estimates.

Additionally, WLS helps to mitigate the impact of heteroscedasticity, ensuring that the model's estimates and predictions remain unbiased and accurate. This method is ideal for situations where the data points exhibit variability across different levels of predictors, making it a practical choice for real-world datasets.

From the WLS output, key predictors such as **Battery**, **Fast Charge**, **Top Speed**, and **Range** stand out as highly significant, showcasing their dominant role in explaining the dependent variable. This provides valuable insight for decision-making, as these variables can be targeted for optimization or improvement to influence Y effectively.

VIII. Conclusion

	R-squared	RMSE	MSE	AIC
OLS	0.7167	0.5313812	0.282366	499.0111
Robust	0.7512	0.7129383	0.508281	-193.7533
WLS	0.841	0.6133921	0.01383969	321.7693

From these results, we can conclude that the best model is the **WLS Regression**. This conclusion is based on several factors: WLS has a significantly higher **R-squared** value compared to Robust Regression, indicating that it explains a larger proportion of the variance in the dependent variable (Y). Additionally, the WLS model demonstrates a very low **Residual Standard Error**, suggesting that the predictions closely align with the actual data, ensuring higher accuracy.

The performance of the independent variables (X) in the WLS model is also excellent, allowing for a clear and comprehensive explanation of the dependent variable. Compared to the Robust model, WLS provides better predictive accuracy and a stronger relationship between the predictors and the dependent variable.

Based on the results, the independent variables that significantly influence the dependent variable—**electric vehicle**

price—are **Range**, **Fast Charge**, **Top Speed**, and **Acceleration**. These variables stand out as the most critical factors, highlighting their importance in shaping the pricing of electric vehicles.

Furthermore, the WLS model's ability to assign weights to observations effectively addresses issues like heteroscedasticity, ensuring that the model remains unbiased and robust in scenarios where variability in data points is present. This makes WLS not only the best-performing model among the three but also the most reliable for practical applications in analyzing and predicting electric vehicle prices.

IX. References

- A. Pramudito, "Global electric vehicle sales increase by 55%, reaching a record of 10.5 million units," *Katadata*, Jan. 9, 2025.

  (https://katadata.co.id/berita/industri/6435216b94681/penjualan-mobil-listrik-global-naik-55-sentuh-rekor-10-5-juta-unit/)

- "Electric vehicle sales report in Indonesia skyrocketed by 383% in 2022," *Bisnis.com*, Jan. 13, 2023.

  (https://otomotif.bisnis.com/read/20230113/46/1617910/rapor-penjualan-mobil-listrik-ri-pada-2022-meroket-383-persen)

- "Enhanced weighted least squares regression: A robust approach for addressing heteroscedasticity," *International Journal of Science and Technology Research Archive*, vol. 7, no. 2, pp. 096-106, 2024.

  (https://sciresjournals.com/ijstra/sites/default/files/IJSTRA-2024-0064.pdf)

- H. White, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, vol. 48, no. 4, pp. 817–838, 1980.

  (https://www.jstor.org/stable/1912934)

- P. J. Huber, "Robust estimation of a location parameter," The Annals of Mathematical Statistics, vol. 35, no. 1, pp. 73–101, 1964.

  (https://projecteuclid.org/euclid.aoms/1177703732)

- R. Koenker and G. Bassett, "Regression quantiles," Econometrica, vol. 46, no. 1, pp. 33–50, 1978.

  (https://www.jstor.org/stable/1913643)