RiESTech
Recent in Engineering Science and Technology

MBI

Article

Comparative Analysis of Regression Methods for Estimation of Remaining Useful Life of Lithium Ion Battery

Idrus Assagaf [1,*], Agus Sukandi [1], Sonki Prasetya [1], Asep Apriana [1], Ega Edistria [2], Nugroho [3], Raihan Kamil [3] Abdul Azis Abdillah [4]

[1] Mechanical Engineering Dept, Politeknik Negeri Jakarta, Depok, Indonesia
[2] Civil Engineering Dept, Politeknik Negeri Jakarta, Depok, Indonesia
[3] IT Department, Politeknik Negeri Jakarta, Depok, Indonesia
[4] Mechanical Engineering Dept, University of Birmingham, Birmingham, United Kingdom
* Correspondence: idrus.assagaf@mesin.pnj.ac.id

Abstract: Lithium batteries play a critical role in modern technological applications, including electric vehicles and portable electronic devices. Ensuring accurate estimation of their remaining useful life is essential to improve system efficiency and reliability. This study focuses on predicting the remaining useful life of lithium batteries using advanced regression methods. Data were collected from lithium battery charge-discharge cycles, encompassing key operational parameters such as voltage, current, and temperature. The analysis employed several regression models, including linear regression, lasso regression, and Ridge regression, to identify relationships between these parameters and battery life. The models were evaluated based on estimation accuracy, with Root Mean Square Error (RMSE) as the primary performance metric. The findings demonstrate that regression methods can effectively capture non-linear relationships between input variables and the remaining useful life, with lasso and Ridge regression showing superior performance in reducing prediction errors. These results underscore the potential of regression-based approaches in providing robust and reliable estimations of battery life. The conclusions highlight the importance of these models for developing predictive battery management systems, which can optimize battery performance and extend their operational lifespan across various applications. This research establishes a solid foundation for future studies on intelligent battery health monitoring and management.

Keywords: Lithium Battery; Remaining Useful Life; Regression Methods; Lasso Regression; Ridge Regression; Battery Management Systems.

Citation:. Assagaf, I.; Sukandi, A.; Prasetya, S.; Apriana, A.; Edistria, E.; Nugroho, N.; Kamil, R.; Abdillah, A.A.. (2025). Comparative Analysis of Regression Methods for Estimation of Remaining Useful Life of Lithium Ion Battery, 3(01), 9–18. Retrieved from https://www.mbi-journals.com/index.php/riestech/article/view/93

Academic Editor: Vika Rizkia

Received: 03 January 2025
Accepted: 16 January 2025
Published: 31 January 2025

Publisher's Note: MBI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

In the rapidly developing era of electric vehicles (EVs), lithium-ion batteries play a critical role in determining their performance and lifetime [1], [2]. Battery decreasing durability and capacity over time pose significant challenges, affecting EVs' efficiency and lifespan. Consequently, a reliable and accurate method for estimating the remaining useful life (RUL) of lithium-ion batteries is essential to predict performance degradation and support preventive maintenance systems [3], [4].

This article aims to conduct a comparative analysis of various regression methods for estimating the RUL of lithium-ion batteries in EV applications. Regression techniques such as linear regression, Lasso Regression, and Ridge Regression are selected because they can model linear relationships between key variables related to battery charge-dis-

charge characteristics and lifetime estimation [5], [6]. Each method offers unique advantages in terms of accuracy, simplicity, and regularization capabilities, essential to address data variability and multicollinearity.

In this study, the dataset includes a variety of battery variables, including charging time, maximum and minimum voltages, and charging duration at a given voltage. By comparing the performance of three regression methods using evaluation metrics such as Root Mean Square Error (RMSE) and R-squared (R^2), this article attempts to identify the optimal approach for battery RUL estimation. The findings are expected to contribute to developing more intelligent and sustainable battery management systems, which are crucial for optimizing the use of EVs in the future.

Previous studies have explored various regression-based approaches to modelling battery capacity degradation. For example, Kwon et al. [7] showed that multiple linear regression effectively estimates battery capacity degradation using data from the accelerated deterioration test. However, this method suffers from highly complex data and multicollinearity among variables, necessitating regularization techniques such as Deep Learning.

Lasso Regression, introduced by Wang et al. [5], has been shown to effectively select the most influential features that contribute to battery degradation, especially in datasets with high variability. By applying L1 regularization, Lasso pushes the coefficients of insignificant variables to zero, improving model interpretability and reducing the risk of overfitting. In addition, the lasso algorithm has also proven to be easy to implement, especially in real-time battery health prediction.

Similarly, Ridge Regression, which uses L2 regularization, has been widely applied to handle datasets with high multicollinearity. Zequera et al. [8] found that Ridge provides fairly accurate battery RUL estimates when variables are highly correlated because the L2 penalty prevents coefficients from being too large. Ridge is very useful because it can prevent the model from overfitting and reduce errors.

Recent studies, such as those conducted by [9], have integrated regression approaches with other machine learning techniques, such as Random Forest [6], [9], Support Vector Machine (SVM)[9], and Deep Learning [3], [9], [10], to improve the accuracy of battery life prediction. However, tree-based methods, SVM, and deep learning often require higher computational resources and are less interpretable than regression models in explaining variable relationships.

Based on previous studies, this study focuses on a comparative analysis of Linear Regression, Lasso Regression, and Ridge Regression to estimate the RUL of lithium-ion batteries in electric vehicles. By evaluating the strengths and weaknesses of each method through metrics such as RMSE and R^2, this study aims to provide practical guidance for selecting the most appropriate regression approach in battery management system applications.

2. Materials and Experiment Methods

This research used experimental and machine-learning methods with several main stages. The first stage is dataset collection, where relevant data is collected to support the experiment while ensuring that the data obtained is by the research objectives. Next, the data understanding process and outlier removal are carried out. At this stage, the dataset is cleaned from noise, outliers are removed so the data is ready to be used in modelling, and the correlation between features is measured to understand the relationship between variables. After that, data preparation is carried out, which includes dataset division (splitting), data normalization to ensure uniform data distribution, and dimensionality reduction using the Principal Component Analysis (PCA) technique.

The next stage is modelling, where the model is built using machine learning algorithms such as Linear Regression, Lasso Regression and Ridge Regression. These various algorithms are compared to determine the most optimal model. After the model is developed, the evaluation and validation stages are carried out to test the performance of the model using test data. Validation is carried out using evaluation metrics such as RMSE and coefficient of determination (R^2) to ensure the accuracy and reliability of the predictions produced by the model. This stepwise approach is designed to ensure comprehensive and relevant research results. The detailed process for each step in this research is as follows:

2.1. Data Understanding

This experiment uses the HNEI Dataset [11], [12], [13] in an experiment to test the prediction model for the remaining battery life of an NMC-LCO (Nickel Manganese Cobalt Oxide - Lithium Cobalt Oxide)-based lithium battery with a nominal capacity of 2.8 Ah. The dataset includes data from 51 18650 battery cells, a typical lithium-ion battery type used in portable electronic devices and electric vehicles. The dataset contains 15,064 rows and 9 columns, with no missing or duplicate values. However, 643 outliers were removed and deleted using the Interquartile Range (IQR) method to ensure data quality.

This experiment uses several regression techniques, namely Linear Regression, Lasso Regression and Ridge Regression, to determine the best model for predicting battery life based on relevant variables. The model is selected based on its ability to handle linear relationships between variables and its similarity when combined with regularization to improve model performance.

The variables included in this dataset reflect important characteristics of battery charging and discharging, which are expected to affect the battery's remaining capacity over time. Here is a brief description of each variable:

1. Cycle Index: The cycle index indicates the number of charge and discharge cycles the battery has undergone.
2. Discharge Time (s): The duration of the battery discharge until it reaches a specific voltage, providing information about the battery’s durability.

3. Drop 3.6-3.4V(s): The time required for the battery to drop from 3.6V to 3.4V, indicating capacity degradation.
4. Max. Discharge Voltage (V): The maximum voltage reached during battery discharge is relevant to meeting voltage limits.
5. Min. Charge Voltage (V): The minimum voltage during charging represents the starting point of capacity replenishment.
6. Time at 4.15V (s): The duration the battery spends at 4.15V during charging, indicating stability in a certain phase.
7. Constant Current Time: Time in the constant current phase during charging, closely related to battery life.
8. Charging time (s): The total charging time from empty to full indicates charging efficiency.
9. RUL is the estimated number of cycles or time remaining before the battery loses significant performance.

These variables provide quantitative insight into the condition and remaining capacity of the battery. The developed model will utilize this information to predict the remaining battery life accurately. The experimental results will show how effective each regression method is in modelling the relationship between these variables and the remaining life of lithium-ion batteries.

2.2. Data Preparation

The data preparation stage is carried out systematically to ensure that the data used in modelling is of optimal quality. This process includes data cleaning, feature correlation analysis against targets (RUL), normalization, dataset separation, and dimensionality reduction using PCA method. Each step in this process is designed to increase the effectiveness and efficiency of the prediction model to be developed.

1. Data Cleaning: The first step is to examine the structure of the dataset, including dimensions, data types, and number of entries, to gain an initial understanding. Summary statistics are used to analyze the distribution of numeric data and detect outliers. Boxplot and histogram mapping are used to visually identify extreme values, while pairplot helps analyze the relationship between features in the dataset. Correlations between features are also calculated and visualized using a heatmap to assess the dependency between variables. The plot of RUL Correlation Heatmap can be seen in figure 2. It can be seen that all of the features highly correlated with RUL with at least 0.82 for the minimum correlation. Detected outlier values are removed using IQR method to ensure clean and consistent data.

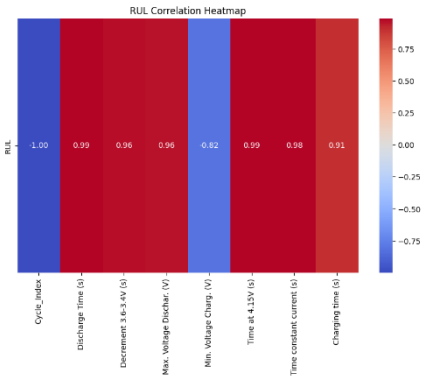

Figure 1. Correlation Heatmap Between RUL and Other Features

2. Dataset Splitting: The data is split into a training set and a testing set to ensure the objective model evaluation. The training set is used to train the model, while the testing set is used to test the model's performance on new, never-before-seen data. This process prevents overfitting, where the model focuses too much on the training data and cannot handle new data well. Typically, the dataset is split in a 50:50 ratio, with the additional option of using a validation set in parameter tuning.
3. Normalization: After cleaning, the data is normalized to align the scale of each variable. This step is important so that large-scale variables (e.g., time in seconds) do not dominate the analysis results compared to small-scale variables (e.g., voltage in volts). Normalization is done by transforming the data into a specific range, such as 0 to 1, or with a standard deviation scale. This process improves consistency between variables and makes further analysis easier.
4. Dimensionality Reduction with PCA: Dimensionality reduction is performed using PCA to simplify the data without sacrificing important information. PCA transforms the dataset into several principal components representing most data variability. This process helps reduce data complexity and improves the computational efficiency of the model. The PCA results are visualized in the principal component space (e.g., PC1 and PC2) to reveal patterns or clusters that may be hidden in the initial data space.

Each step in this data preparation process is designed to ensure the data is ready for modelling and prediction. The overall visualization of this step illustrates the effectiveness of the preprocessing method in simplifying and improving the data quality, thereby supporting the development of reliable and accurate predictive models.

2.3. Model Development

This study uses three main regression models: Linear Regression, Lasso Regression, and Ridge Regression. *Linear Regression* is an algorithm that models the linear relationship between independent and target variables. The linear regression formula is given by

$$Y = a + bX + e,$$

where Y is the dependent variable, X is the independent variable, a is the y-intercept, b is the slope, and e is the error term. This model finds the best line that minimizes the error between predicted and actual values using the Ordinary Least Squares (OLS)

method. Linear Regression does not use regularization, so it is prone to overfitting, especially on datasets with many features. This model is implemented using the `LinearRegression()` function from scikit-learn without additional parameter adjustments. The training process is carried out with `fit()`, and predictions are made using the `predict()` function. Performance evaluation uses RMSE metric to assess model accuracy.

Lasso Regression, on the other hand, introduces an L1 regularization penalty, which adds a penalty to the absolute value of the regression coefficients. Lasso regression, or L1 regularisation, adds a penalty term $\lambda\sum|\beta_i|$ to the ordinary least squares (OLS) formula, where λ is a tuning parameter and β_i are the coefficients. This regularization allows the model to perform feature selection by reducing some regression coefficients to zero, thus retaining only the most significant features. This makes Lasso Regression very suitable for high-dimensional datasets. The model uses the `Lasso()` function from scikit-learn, with the regularization parameter alpha tested through a grid search to determine the optimal value. The model is trained using the `fit()` function, and the prediction results are tested with RMSE to evaluate the balance between bias and variance.

As an alternative, Ridge Regression uses the L2 regularization penalty, which adds a penalty to the square of the regression coefficients. Ridge regression, or L2 regularisation, adds a penalty term $\lambda\sum\beta_i^2$ to the OLS formula. This regularization helps reduce overfitting by reducing large regression coefficient values, making the model more stable, mainly when multicollinearity exists between features. The Ridge Regression implementation uses the `Ridge()` function from scikit-learn, with the alpha parameter set to control the strength of the penalty. Similar to Lasso, the alpha value is tested through a grid search to obtain optimal performance.

These three models were chosen because they can handle the relationship between features in the dataset and RUL target. Each model is evaluated based on its performance on the test data using the metrics RMSE and R^2. The best model is selected based on the evaluation results to ensure the accuracy and reliability of the predictions.

2.3. Evaluation Metrics

This study evaluates model performance using two main metrics, namely RMSE and R^2. These two metrics provide a comprehensive understanding of the prediction accuracy and the model's ability to explain the variability of the target data.

RMSE measures the average error rate produced by the model in predicting the target value. This metric calculates the root mean of the square of the difference between the actual value and the predicted value. The smaller the RMSE value, the better the model produces predictions close to the target value. RMSE also provides a more significant penalty for large errors, making it sensitive to outliers in the data.

Meanwhile, the coefficient of determination or R^2 is used to evaluate the extent to which the model can explain the variability of the target data based on the input features. The R^2 value ranges from 0 to 1, where a value close to 1 indicates that the model can

explain most of the variability in the data. A high R^2 indicates that the relationship between the input and output variables can be modelled well. On the other hand, a low R^2 value indicates that the model cannot effectively capture patterns in the data.

In this study, the combination of RMSE and R^2 provides an in-depth evaluation of model performance. RMSE shows the absolute error in the same unit as the target (RUL), while R^2 measures the proportion of data variability that the model can explain. By using these two metrics, model performance can be analyzed comprehensively in terms of prediction accuracy and generalization ability.

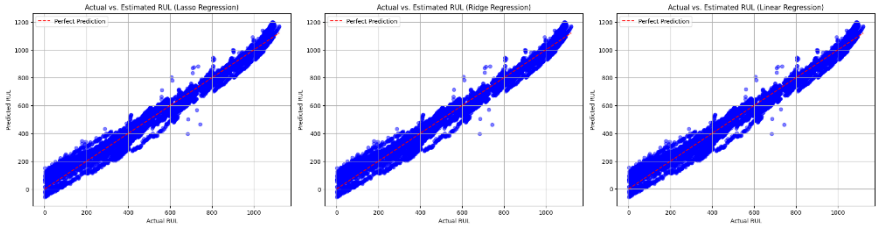

Figure 2. Plot of all models: Linear Regression (left), Lasso Regression (Midle), and Ridge Regression (Right)

3. Results and Discussion

3.1. Results

The experimental results show the performance of Linear Regression, Lasso Regression, and Ridge Regression models in predicting the remaining useful life of lithium batteries based on the main evaluation metrics, namely RMSE and R^2. The results experiment can be seen in table 1 and table 2. Meanwhile the plot of all estimation results can be seen in figure 2.

Table 1. The RMSE results for all model

Model	Training RMSE	Testing RMSE
Linear Regression	51.281033	51.879174
Lasso Regression	51.290612	51.881304
Ridge	51.282188	51.878767

Based on the RMSE value, the three models show similar performance. In the training data, linear Regression produces an RMSE of 51.281, while Lasso and Ridge have RMSE values of 51.291 and 51.282, respectively. On the testing data, the RMSE for linear Regression is 51.879, while Lasso reaches 51.881, and Ridge is slightly better with a value of 51.879. This slight difference indicates that Lasso and Ridge's regularization does not provide significant improvements compared to the simple linear regression model.

Table 2. The R2 results for all model

Model	Training R2	Testing R2
Linear Regression	0.974206	0.973200
Lasso Regression	0.974196	0.973198
Ridge	0.974205	0.973201

For the R^2 value, the three models also show almost identical results. On the training data, Linear Regression, Lasso and Ridge achieved R^2 of 0.974206, 0.974196, and 0.974205, respectively. Similar results were also seen in the testing data, with R^2 values of 0.973200, 0.973198, and 0.973201, respectively. The high R^2 values of these three models indicate that the models can explain data variability in training and testing data.

Although all three models performed very well predicting the remaining battery life, the performance differences between linear Regression, Lasso Regression, and Ridge Regression were minimal. Thus, linear Regression can be considered the most efficient model, considering its simplicity and the absence of significant differences in the evaluation results. However, if there is a need for feature selection or handling multicollinearity in the future, Lasso and Ridge remain alternatives that can be considered.

3.2. Discussion

The *experimental* results show that the three regression models, linear Regression, Lasso Regression, and Ridge Regression, perform almost equally well in predicting the remaining life of lithium batteries. The high RMSE and R^2 values on both training and testing data indicate that these three algorithms can very well model the relationship between input features and the target (RUL).

Linear Regression gives excellent results with an RMSE value of 51.281 on the training data and 51.879 on the testing data and an R^2 of 0.974206 and 0.973200, respectively, for the training and testing data. The absence of regularization in Linear Regression does not significantly affect performance, indicating that the dataset used has a strong linear relationship between input and output variables and minimal multicollinearity problems or irrelevant features.

Lasso Regression, which uses an L1 regularization penalty for feature selection, produces almost identical RMSE values to linear regression. However, the feature selection properties of Lasso did not significantly improve performance, possibly because the dataset was already clean of less relevant features.

Ridge Regression, with an L2 regularization penalty, showed the best RMSE result on the test data, which was 51,878. However, the difference was very small compared to linear regression, so the effect of Ridge regularization in improving model generalization was also not significant on this dataset.

These results provide several important insights. First, a simple linear regression model can be effectively used for battery life prediction, especially when a strong linear relationship is present in the data. Second, Lasso and Ridge can still be considered in other scenarios, such as when the dataset has many potentially redundant features or multicollinearity.

This discussion also emphasizes the importance of comprehensive evaluation in selecting a model. Although all three models performed almost identically, linear regression was found to be the most efficient because it did not require additional parameter tuning, such as alpha values in Lasso and Ridge. In the future, experiments can be extended by using non-linear methods, such as tree-based algorithms or neural networks, to examine whether non-linear models can outperform complex relationships in larger or heterogeneous datasets.

4. Conclusions

This study successfully developed a prediction model for lithium batteries' RUL based on Regression, including Linear Regression, Lasso Regression, and Ridge Regression. The evaluation results show that the three models have similar performance, with an RMSE value on the test data of around 51.88 and a very high R^2 value, approaching 0.973. This indicates that the model can predict RUL with a good level of accuracy and can explain most of the variability in the data.

Linear Regression performs slightly better than Lasso and Ridge Regression based on the RMSE and R^2 values. However, the difference in performance between the three models is minimal, indicating that regularization in Lasso and Ridge does not provide significant benefits in this case. This indicates that the data used does not have overfitting or too many variables for the simple Linear Regression model.

These findings confirm that simple linear Regression can be an effective choice for predicting the RUL of lithium batteries under clean and structured data conditions. In the future, this approach can be used as a basis for the development of more sophisticated battery management systems, including the application of other machine learning techniques, such as non-linear or deep learning-based models, to capture more complex patterns.

Overall, this study contributes to understanding lithium battery RUL prediction and opens up opportunities for practical applications in battery management systems in various electronic devices and electric vehicles.

Acknowledgments: The authors would like to express their gratitude to Politeknik Negeri Jakarta for providing financial support for this research. This funding has been instrumental in facilitating the completion of this study. We deeply appreciate the institution's commitment to fostering research and innovation.

References

1. B. E. Lebrouhi, Y. Khattari, B. Lamrani, M. Maaroufi, Y. Zeraouli, and T. Kousksou, "Key challenges for a large-scale development of battery electric vehicles: A comprehensive review," *J Energy Storage*, vol. 44, p. 103273, Dec. 2021, doi: 10.1016/J.EST.2021.103273.
2. X. Lai *et al.*, "Critical review of life cycle assessment of lithium-ion batteries for electric vehicles: A lifespan perspective," *eTransportation*, vol. 12, p. 100169, May 2022, doi: 10.1016/J.ETRAN.2022.100169.
3. M. Catelani, L. Ciani, R. Fantacci, G. Patrizi, and B. Picano, "Remaining Useful Life Estimation for Prognostics of Lithium-Ion Batteries Based on Recurrent Neural Network," *IEEE Trans Instrum Meas*, vol. 70, 2021, doi: 10.1109/TIM.2021.3111009.
4. H. Rauf, M. Khalid, and N. Arshad, "Machine learning in state of health and remaining useful life estimation: Theoretical and technological development in battery degradation modelling," *Renewable and Sustainable Energy Reviews*, vol. 156, p. 111903, Mar. 2022, doi: 10.1016/J.RSER.2021.111903.
5. X. Wang, J. Li, B. C. Shia, Y. W. Kao, C. W. Ho, and M. C. Chen, "A Novel Prediction Process of the Remaining Useful Life of Electric Vehicle Battery Using Real-World Data," *Processes 2021, Vol. 9, Page 2174*, vol. 9, no. 12, p. 2174, Dec. 2021, doi: 10.3390/PR9122174.
6. H. Rauf, M. Khalid, and N. Arshad, "A novel smart feature selection strategy of lithium-ion battery degradation modelling for electric vehicles based on modern machine learning algorithms," *J Energy Storage*, vol. 68, p. 107577, Sep. 2023, doi: 10.1016/J.EST.2023.107577.
7. S. J. Kwon, D. Han, J. H. Choi, J. H. Lim, S. E. Lee, and J. Kim, "Remaining-useful-life prediction via multiple linear regression and recurrent neural network reflecting degradation information of 20Ah LiNixMnyCo1−x−yO2 pouch cell," *Journal of Electroanalytical Chemistry*, vol. 858, p. 113729, Feb. 2020, doi: 10.1016/J.JELECHEM.2019.113729.
8. R. Gilbert Zequera, V. Rjabtšikov, A. Rassõlkin, T. Vaimann, and A. Kallaste, "Modeling Battery Energy Storage Systems Based on Remaining Useful Lifetime through Regression Algorithms and Binary Classifiers," *Applied Sciences 2023, Vol. 13, Page 7597*, vol. 13, no. 13, p. 7597, Jun. 2023, doi: 10.3390/APP13137597.
9. G. Naresh and P. Thangavelu, "Integrating machine learning for health prediction and control in over-discharged Li-NMC battery systems," *Ionics (Kiel)*, pp. 1–18, Sep. 2024, doi: 10.1007/S11581-024-05834-5/METRICS.
10. A. A. Abdillah, C. Zhang, Z. Sun, J. Li, H. Xu, and Q. Zhou, "Data-driven Modelling for EV Battery State of Health Estimation using SFS-PCA Learning," *Proceedings of the 2023 7th CAA International Conference on Vehicular Control and Intelligence, CVCI 2023*, 2023, doi: 10.1109/CVCI59596.2023.10397248.
11. Z. Li, Q. Shi, J. Xia, K. Wang, and K. Jiang, "A Novel Method Based on Stacking Model for Remaining Useful Life Prediction of Lithium-ion Batteries," *2023 26th International Conference on Electrical Machines and Systems, ICEMS 2023*, pp. 974–978, 2023, doi: 10.1109/ICEMS59686.2023.10345332.
12. J. N. C. Sekhar, B. Domathoti, and E. D. R. Santibanez Gonzalez, "Prediction of Battery Remaining Useful Life Using Machine Learning Algorithms," *Sustainability 2023, Vol. 15, Page 15283*, vol. 15, no. 21, p. 15283, Oct. 2023, doi: 10.3390/SU152115283.
13. "BatteryArchive.org." Accessed: Dec. 09, 2024. [Online]. Available: https://www.batteryarchive.org/list.html