


Paper Type: Original Article



## Predicting the Claim Amount from Car Insurance Using Multiple Linear Regression: a Case Study of Iran Insurance

Reza Rasinojehdehi<sup>1,\*</sup> , Soheil Azizi<sup>2</sup>

<sup>1</sup> Department of Industrial Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran; reza.rasi1515@gmail.com.

<sup>2</sup> Department of Accounting and Management, Allameh Tabatabai University, Tehran, Iran; soheilazizi1@gmail.com.

Citation:



Rasinojehdehi, R., & Azizi, S. (2023). Predicting the claim amount from car insurance using multiple linear regression: a case study of Iran insurance. *Big data and computing visions*, 3(3), 125-136.

Received: 10/02/2023

Reviewed: 12/03/2023

Revised: 09/05/2023

Accept: 01/06/2023


### Abstract

The escalating annual insurance costs nationwide have sparked a growing interest among insurance industry managers and policymakers in analyzing insurance data to forecast future costs. Accurately predicting the number of claims and implementing appropriate policies can help mitigate potential losses for insurance companies and customers. This study focuses on predicting the amount of customer claims and utilizes data from 128 individuals insured by Iran insurance company. The dataset includes various attributes such as the age of the vehicle owner, type of car, age of the car itself, number of claims, and the corresponding claim amounts (measured in 10,000 Tomans) recorded in the year 1400. All features, except the claim amount (the target variable), were discretized into ordinal variables to ensure accurate analysis and address any outliers or data inconsistencies. Multiple linear regression was employed to predict the target variable, enabling an investigation into the influence of each feature on estimating the claim amount. The data analysis was conducted using IBM SPSS MODELER software, allowing for a comprehensive examination of the assumptions associated with the regression model. By leveraging this approach, insurance industry stakeholders can gain valuable insights into predicting claim amounts and make informed decisions to optimize their operations and minimize potential financial risks.

**Keywords:** Data analysis, Prediction, Regression, Insurance claim amount.

## 1 | Introduction

In today's rapidly evolving business landscape, science has become a critical and highly sought-after resource for organizations aiming for growth and prosperity. This trend holds across various industries, including financial and insurance institutions, which have recognized the immense value of predicting customer behavior. Organizations lacking a scientific perspective in their vision are disadvantaged in an increasingly competitive environment, as science is now considered a strategic asset. Customer behavior prediction has gained significant importance for financial and insurance institutions. Economists now view customer behavior as a predictable process, prompting businesses to understand and influence consumer actions throughout their purchasing journey proactively. By comprehending customer behavior, insurance brokers can anticipate future conduct and reactions, enabling them to adopt appropriate sales strategies based on these forecasts. One of the most valuable sources for gathering customer information is customer communication software, which records and

 Licensee Big Data and Computing Visions. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

 Corresponding Author: reza.rasi1515@gmail.com

 <https://doi.org/10.22105/bdcv.2024.424709.1169>

stores essential data for analyzing customer behavior. Businesses can use mathematical, statistical, and artificial intelligence techniques to effectively identify behavioral patterns and analyze raw data. This analytical approach helps reduce risk and increases profitability by enabling organizations to make data-driven decisions and optimize their operations. The primary objective of this study is to identify patterns in the estimation of customer-requested damages by utilizing the data recorded in customer communication software used by insurance brokers. Uncovering these patterns allows organizations to gain valuable insights, improve their claims processes, and enhance customer satisfaction. This research aims to give insurance institutions a strategic advantage in understanding and predicting customer behavior, ultimately leading to improved business outcomes.

## 2 | Literature Review

Predicting the claim amount from car insurance is critical to the insurance business. It is essential for insurance companies to accurately estimate potential risks and losses to set appropriate premiums and reserves. This prediction allows companies to manage their financial resources effectively and ensure they have enough funds to cover potential claims. Because of the importance of predicting claim amounts in the insurance industry, many scholars and researchers have contributed their models and solutions to improve predictive accuracy. These models often utilize advanced statistical techniques, machine learning algorithms, and big data analytics to understand better and predict claim amounts. By continuously refining and developing new predictive models, scholars aim to provide insurance companies with more accurate and reliable tools for assessing risk and setting premiums.

Resource allocation is critical to any business, and insurance companies are no exception. Effective resource allocation is essential for ensuring that companies have enough funds to cover potential claims, manage risks, and provide quality services to their customers. Many scholars have worked on developing models and solutions to improve resource allocation [1]–[7].

Predicting claim amounts is closely related to resource allocation in insurance companies. Insurance companies can better allocate their resources by accurately predicting claim amounts to ensure they have enough funds to cover potential claims. This, in turn, enables companies to operate more efficiently and effectively, ultimately improving their profitability and ability to provide quality services to their customers. In recent years, predictive modeling, such as multiple linear regression, has become increasingly popular in the insurance industry for predicting claim amounts. These models allow insurance companies to analyze various factors that may impact claim amounts, such as the type of car, driver's age, driving history, and geographical location. Using these models, insurance companies can identify the most influential factors and create a model that can accurately predict claim amounts.

The study by Ye et al. [8] contributes to the literature on auto insurance claims prediction by addressing the challenge of highly skewed data. The authors analyze the Kangaroo Auto Insurance company data and investigate the impact of combining different prediction methods on accuracy. Their findings highlight the importance of selecting appropriate prediction accuracy measures and the potential benefits of model combination methods, such as ARM-Tweedie, for improving forecast performance in Low Frequency and High Severity (LFHS) data. The study underscores the significance of model combination methods in enhancing prediction accuracy for auto insurance claim costs.

David [9] provides a comprehensive overview of Generalized Linear Models (GLMs) techniques in the context of insurance premium calculation. The paper emphasizes the typical approach of combining the conditional expectation of claim frequency with the expected claim amount to obtain the insurance premium. Specifically, the author focuses on using GLMs to calculate the pure premium based on observable characteristics of policyholders. The article includes a numerical illustration based on a French auto insurance portfolio, demonstrating the practical application of GLMs in this domain. The statistical software SAS is utilized to perform the numerical analysis, highlighting the empirical relevance of the GLM techniques presented in the paper.

Meng et al. [10] developed a supervised driving risk scoring neural network model using automobile insurance claims data and telematics car driving data. Based on a one-dimensional convolutional neural network, this model generates risk scores for individual car driving trips and significantly improves the classical Poisson generalized linear model for predicting automobile insurance claims frequency. Telematics-based insurers can leverage this model to identify more heterogeneity in their portfolio and incentivize safer drivers with premium discounts.

Selvakumar et al. [11] aim to address the challenge of predicting insurance claim amounts for different vehicle Categories. They use 34 years of historical data and employ machine learning models to forecast claim amounts, including linear regression, exponential smoothing, ARIMA, artificial neural network, and hybrid ARIMA-ANN models. Their empirical analysis shows that the Artificial Neural Network model outperforms other time series models regarding predictive accuracy. The study highlights the potential of machine learning approaches to assist insurance companies in India provide better predictive models for improved claims settlement and management across different vehicle Categories.

The paper addresses the need for insurance companies to assess their customers' risk for Third-Party Liability (TPL) car insurance. The study analyzes a dataset of 13,388 insurance claims from an Iranian insurance company and uses logistic regression and random forest models with different resampling techniques to improve model performance. The results indicate that the random forest model with hybrid resampling methods is the most effective classifier. Additionally, the study identifies victim age, premium, car age, and insured age as the most important factors for predicting insurance claims. Overall, the study provides valuable insights for insurance companies looking to improve their risk assessment processes for TPL car insurance.

Jaworski and Czarnocha [12] identified the main determinants of the capital structure of energy industry companies in the European Union. This study was conducted based on a panel of 6,122 companies from 25 EU countries. Multiple regression analysis was used in this study. This study found strong evidence for a positive relationship between company debt and its size and a negative relationship between profitability and liquidity [10].

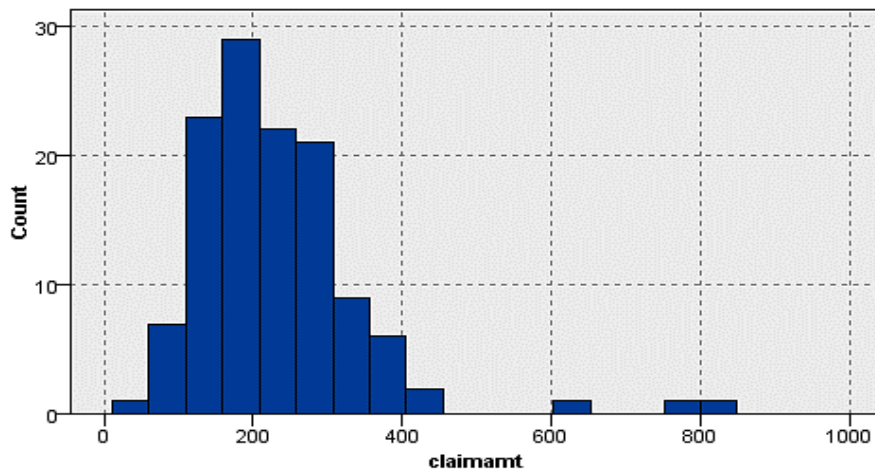
Kaushik et al. [13] developed an artificial intelligence network-based regression model to predict health insurance premiums with an impressive accuracy of 92.72%. Their research highlights the potential of AI and machine learning in revolutionizing the insurance industry and improving healthcare services for consumers. These technologies have also transformed the healthcare industry, enabling faster disease diagnosis and treatment anticipation and revolutionized the insurance sector, creating more efficient and accurate health insurance policies.

### 3 | Introduction of Problem Variables and Data Preprocessing

The problem in this paper has four input features and one target field. The input features are vehicle age, owner age, vehicle group, and number of claim occurrences. The target field is the claim loss amount. The data for this problem consists of 128 samples collected based on the behavior of Iran insurance customers.

The vehicle age feature is an ordinal discrete variable that includes values from 1 to 4. Vehicles with mileage between 0 to 3 years are classified in Category 1, vehicles with mileage between 4 to 7 years are classified in Category 2, vehicles with mileage between 8 to 11 years are classified in Category 3, and vehicles with mileage over 10 years are classified in Category 4. The vehicle group feature is a nominal variable that includes groups A, B, C, and D, assigned values 1, 2, 3, and 4, respectively. The owner age feature is also an ordinal discrete variable that includes values from 1 to 8. Owners aged between 17 to 20 years are classified in Category 1, owners aged between 20 to 24 years in Category 2, and so on, and owners aged over 60 years in Category 8. The number of claim occurrences feature is a discrete variable with integer values ranging from 0 to 434. As the average claim amount feature has a high correlation with the number of claim occurrences feature, the initial role of this feature is assigned none using the STATISTICS FILE

node in the MODELER software. The claim loss amount feature, considered the target field, is a continuous variable with a distribution shown in *Fig. 1*. It has a minimum value of 11.00, a maximum value of 850, a mean of 231.138, a standard deviation of 117.048, and a skewness value of 2.313. As seen in *Fig. 1*, the plot is skewed to the right, confirmed by the skewness value of 2.313.



**Fig. 1.** The shape of the target variable distribution.

The values of the first 10 records of the data (out of 128 records) are displayed in *Table 1*.

**Table 1.** The first 10 data records (out of 128 records).

Data	Owner's Age Category	Vehicle Category	Category Life Car	Amount of Damage Claim (10 Thousand Tomans)	Number of Claims
1	8	289	1	1	1
2	8	282	2	1	1
3	4	133	3	1	2
4	4	160	4	2	1
5	10	372	1	2	2
6	28	249	2	2	3
7	1	288	3	3	4
8	1	11	4	4	1
9	9	189	1	3	1
10	1	288	2	3	6

Considering that at the beginning, vehicle age, vehicle group, and owner age were categorized and converted into nominal or ordinal variables, outliers were also encountered in these variables because the categorization of variables also assigns outliers to Categories.

The claim loss amount variable is the only variable that needs to be dealt with outliers (data points greater than 3 times the variance from the mean) and extreme outliers (data points greater than 5 times the variance from the mean). We replace the outlier and extreme outlier values with the mean using the super node extreme and outlier (see *Fig. 7*).

Considering the absence of missing values and the inconsistency in the problem data, the data preprocessing process seems complete, and the data is ready for modeling. In the next section, we will proceed to the modeling section to discover a relationship for predicting the claimed loss amount using the introduced features in the problem.

## 4 | Data Analysis

In this section, we aim to predict the claim amount based on the characteristics of the vehicle, including vehicle age, the age of the vehicle owner, and the vehicle group by presenting a linear regression model.

One of the preparations required for regression is the transformation of nominal variables into dummy variables. This structure creates  $k-1$  variables, zero and one, for each  $k$  Category of the nominal variable. Since the "vehicle group" variable is a nominal variable with four states, we use the SET TO FLAG node in IBM SPSS MODELER software to transform this nominal variable into three binary variables (Flags) named vehiclegroup\_2, vehiclegroup\_4, and vehiclegroup\_3. We consider Category A of this variable as the reference Category.

Next, we proceed to the Partition node to determine the training and testing data. Given that the total number of data points is 128 and we have a limited amount of data, we select 50% of the data for training and 50% for testing. In the next step, we utilize the Regression node in the modeling section of MODELER to execute the linear regression model.

The results obtained from the execution of the model indicate that the vehicle age variable holds the highest importance among the other features. The predictive importance of the vehicle age variable in forecasting the claim amount was estimated to be 0.37 by the model. Fig. 3 illustrates the importance of variables in prediction.

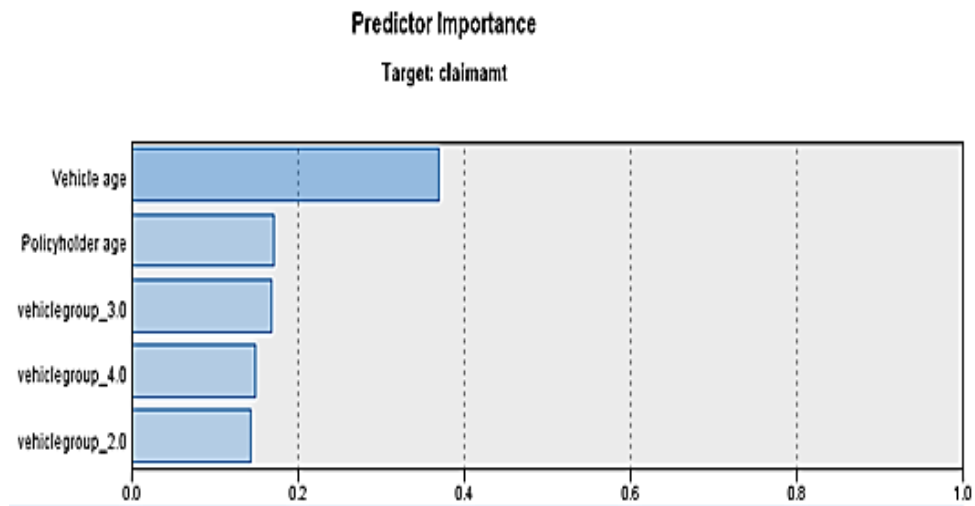


Fig 2. Importance of predictive variables.

Table 2 displays the Analysis of Variance (ANOVA) for the regression model. Based on the results obtained from the ANOVA table, it can be concluded that the created regression model is suitable. This conclusion is supported by the  $p$ -value being less than 0.05, indicating that the Mean Square of Regression (MSR) is significantly greater than the Mean Square Error (MSE). Additionally, considering the  $F$ -statistic in the ANOVA table, it is evident that MSR is approximately 16.451 times greater than MSE. This signifies that the model effectively predicts the target values.

Table 2. Analysis of Variance.

	Sum of Squares	df	Mean Square	F	Sig.
Regression	233006.376	5	46601.2	16.45	.000
Residual	150134.607	53	2832.72		
Total	383140.983	58			

Based on Table 3, the adjusted R square value for the regression model is 0.571, indicating that the variables, including vehicle group Categories 2, 3, and 4, vehicle age, and owner's age, collectively explain approximately 57% of the variance in the target variable, which is the claimed amount. The multiple correlation coefficient (R) indicates a moderately strong positive linear relationship between predictors and the claimed amount. R square ( $R^2$ ) shows that about 60.8% of the variance in the claimed amount is explained by predictors. The standard error of the estimate (53.22338) represents the accuracy of predictions. Considering that the value of the Durbin-Watson statistic is 2.148 and falls within the range of 1.5 to 2.5, it can be argued that there is no significant autocorrelation among the variables [9]. The



critical point is that the Durbin-Watson statistic is acceptable because the error term follows a normal distribution. Therefore, testing the assumption of normality for the errors in the subsequent analysis is essential.

**Table 3. Analysis of Variance.**

R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
.780	.608	.571	53.22338	2.148

The *Table 4* displays regression coefficients and related statistics for the model. It also presents the Variance Inflation Factor (VIF) for each coefficient, which helps assess multicollinearity among the variables.

**Table 4. Regression coefficients.**

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	Collinearity Statistics VIF
	B	Std. Error				
(Constant)	324.2	24.1		13.4	.00	
Owner's age	-2.34	3.07	-.067	-.76	.45	1.06
Vehicle age	-50.09	6.39	-.690	-7.8	.00	1.04
Category 2 of vehicle group	12.90	19.1	.071	.67	.50	1.49
Category 3 of vehicle group	24.85	20.1	.128	1.23	.22	1.45
Category 4 of vehicle group	67.40	19.4	.356	3.46	.00	1.43

The table presents regression coefficients and associated statistics for a regression model. It includes unstandardized coefficients (B) that show the effect of each predictor variable on the dependent variable. For instance, "Owner's age" has a coefficient of -2.34, indicating a decrease in the claimed amount with increasing owner's age. Standardized coefficients (Beta) are also provided, with "vehicle age" having a significant negative impact with a coefficient of -0.690. The t-statistic measures the significance of each coefficient, where higher absolute values indicate greater significance. Some variables, like "vehicle age" and "Category 4 of vehicle group," have p-values of 0.00, indicating high significance. Collinearity Statistics (VIF) assess multicollinearity, and in this case, all VIF values are below 10, suggesting no problematic multicollinearity. Overall, the table offers insights into the significance and impact of predictor variables on the claimed amount in the regression model.

We utilized the "analysis" node in the software to examine the model's predictive performance, and the results are observed in *Table 5*. Given that the correlation value in the test data is 0.84, the results indicate a high degree of correlation between the predicted values and the actual target values. This suggests the regression model's appropriateness, as the strong correlation reflects the model's capability to make accurate predictions that closely align with the actual outcomes.

**Table 5. The results of running the "analysis" node in the modeler software.**

Partition'	Training	Testing
Minimum error	-123.455	-116.934
Maximum error	134.255	137.609
Mean error	0.0	7.237
Mean absolute error	36.703	34.187
Standard deviation	50.878	46.89
Linear correlation	0.78	0.844
Occurrences	59	61

Fig. 3 displays a scatter plot showing the relationship between predicted values and actual target values. The plot demonstrates a strong linear correlation between the predicted and actual values, confirming their high association level. This visual representation aligns with the previous observation of a high correlation coefficient, further supporting the effectiveness and appropriateness of the regression model.

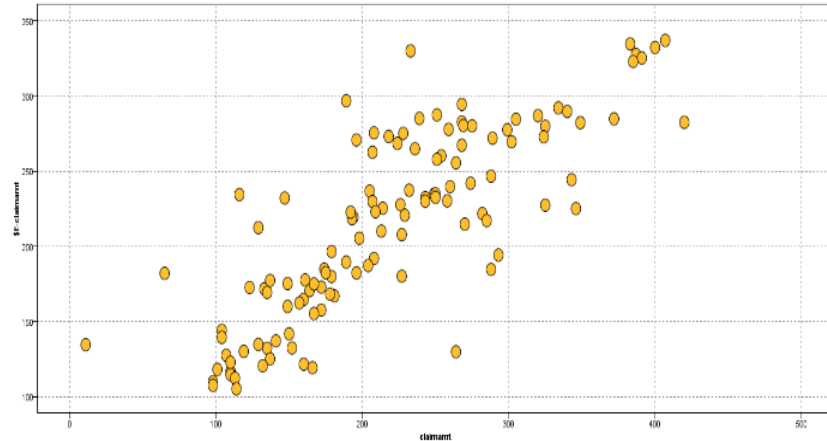


Fig. 3. Scatter plot.

This scatter plot illustrates the relationship between predicted values and actual target values. Based on the results obtained, there is an average error of approximately 26%, with a median error of about 11%. (Refer to Table 6 for details). This information highlights prediction errors in the model, with the median error being a useful measure of central tendency. Further insights can be gained from Table 6 for a detailed breakdown of these error statistics.

Table 6. Descriptive statistics for the variable "percentage error".

Mean	26.082
Min	0.250
Max	1122.316
Standard deviation	103.139
Standard error of mean	9.415
Median	10.866

Table 6 presents detailed descriptive statistics for the "percentage error" variable, shedding light on the distribution of prediction errors in the model. The mean error is approximately 26.082%, indicating an overall positive or negative deviation from the actual values. The minimum error recorded is a mere 0.250%, reflecting the smallest underestimation or overestimation in the model's predictions. In contrast, the maximum error is a substantial 1122.316%, representing the most significant deviation from actual values. The standard deviation of 103.139% illustrates the extent of variation or dispersion in these prediction errors.

In contrast, the standard error of the mean, at 9.415%, gauges the precision of the mean error estimate. The median error of 10.866% serves as a measure of central tendency, indicating the middle point in the distribution of errors. These statistics provide a comprehensive and insightful perspective on the model's prediction errors, encompassing their range, central tendency, and overall variability.

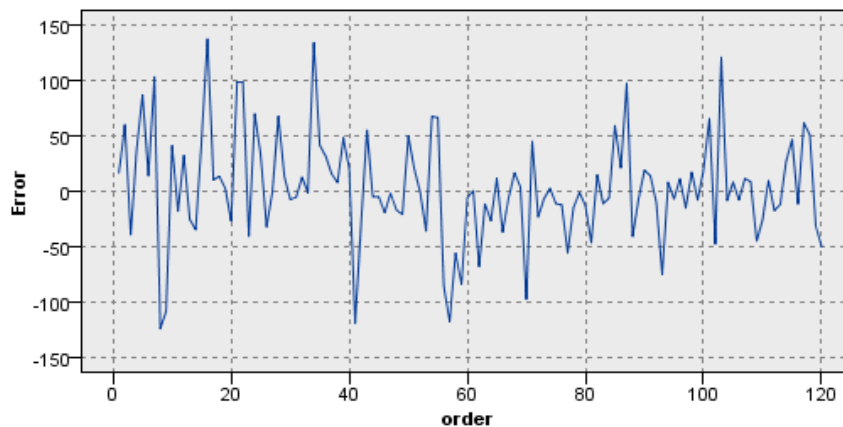
Table 7. Kolmogorov-smirnov test results.

N		120
Normal parameters <sup>a,b</sup>	Mean	3.67865
	Std. Deviation	48.819931
Most extreme differences	Absolute	.107
	Positive	.107
	Negative	-.074
Test statistic		.107
Asymp. Sig. (2-tailed)		.052 <sup>c</sup>

Table 7 provides the results of the kolmogorov-smirnov test, which is a critical evaluation for assessing the normality assumption of error values in the context of linear regression. This test is fundamental as it ensures the validity of the regression model, as one of its core assumptions is that the error values follow a normal distribution [9]. The sample size, denoted as "N," comprises 120 data points. The "normal parameters" section reveals that the error values mean approximately 3.67865, with a standard deviation of around 48.819931. These statistics help describe the central tendency and spread of the error values, which are essential for evaluating normality.

The "most extreme differences" section displays the most extreme absolute, positive, and negative differences observed in the dataset, indicating the variability in the error values. The "test statistic" of 0.107 provides the kolmogorov-smirnov test's result, indicating the maximum difference between the empirical distribution function of the error values and the theoretical normal distribution. A smaller test statistic suggests a closer fit to a normal distribution. The "asympt. Sig. (2-tailed)" value of approximately 0.052 represents the asymptotic significance level for a two-tailed test. In this context, it indicates the likelihood of observing these differences if the error values were sampled from a normal distribution.

A significance level below a predetermined threshold (e.g., 0.05) would suggest a significant departure from normality. In this case, the Kolmogorov-Smirnov test yields an asymptotic significance of approximately 0.052. While this value is slightly above the typical 0.05 significance level, it suggests that the normality assumption is not strongly violated. This is a positive outcome as it supports the validity of the linear regression model by confirming that the error values follow a distribution close to normal, in alignment with the underlying assumption. Having established the acceptance of the normality assumption for the data, we can now rely on the durbin-watson statistic. It indicates that because this statistic falls within the range of [1.5, 2.5], it rejects the hypothesis of autocorrelation in the residuals [9]. Consequently, we conclude that the residuals are not autocorrelated. Fig. 4 visually depicts the variation in residuals concerning the order or index of data points. Observing the graph, it becomes apparent that the residuals display no meaningful dependence on their previous values and do not follow a specific pattern. This further supports the conclusion that the residuals are not autocorrelated.



**Fig. 4. Visual representation of the relationship between the residuals and their position in the dataset.**

Fig. 5 presents a histogram of the error variable. The graph shows that the data distribution is approximately symmetric and resembles a normal distribution. This observation implies that the error values are relatively well-behaved, displaying characteristics akin to a normal distribution. The proximity of the data to a normal distribution is a positive indicator, as it aligns with one of the key assumptions in linear regression, namely the normality of errors. This conformity to a normal distribution further supports the reliability and validity of the regression model.



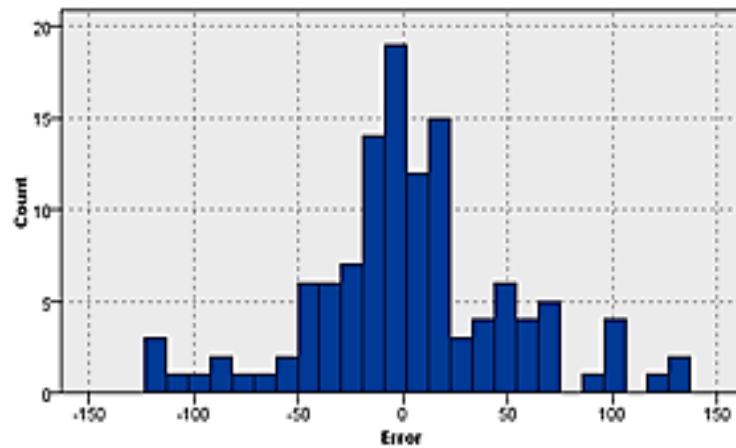


Fig. 5. A histogram of the error variable.

The next assumption in regression analysis is the constancy of error variances. To test this assumption, the values of errors are plotted against the predicted values (values forecasted by the model). *Fig. 6* depicts a scatter plot showing the distribution of residuals concerning the predicted values. *Fig. 6* illustrates the relationship between the residuals and their predicted values. By examining this scatter plot, one can assess whether the variances of errors remain relatively constant across the range of predicted values or if there is a systematic pattern of change. A consistent and random distribution of points would suggest that the assumption of constant error variances is met. However, if there is a discernible pattern, such as a funnel shape or any systematic change in the spread of residuals, it may indicate heteroscedasticity, violating the assumption of constant error variances. Analyzing the scatter plot helps ensure the reliability of the regression model.

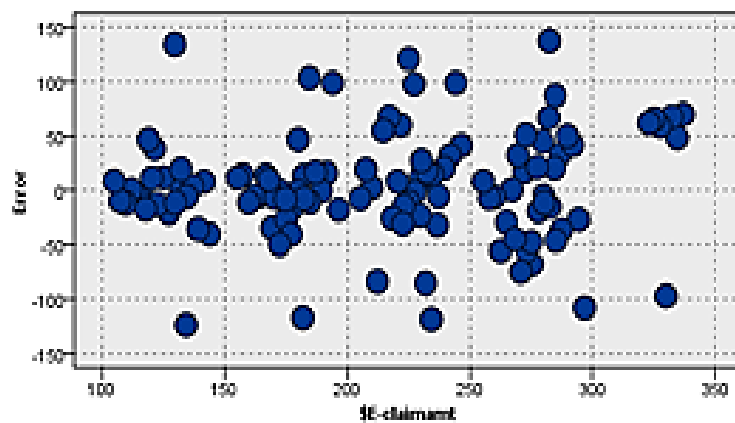


Fig. 6. The distribution of residuals.

As evident from *Fig. 6*, the change in residuals concerning the predicted values exhibits a relatively constant trend. This consistency suggests that the assumption of constant error variances can be reasonably accepted for the predicted values. The constancy of error variances across the range of predicted values is a positive outcome, as it aligns with a fundamental assumption in linear regression. This assumption assures that the variability of errors remains relatively consistent and does not systematically increase or decrease as the predicted values change. The acceptance of this assumption further supports the reliability of the regression model. Now that all the regression model assumptions are satisfied, we examine and interpret the regression coefficients presented in *Table 4*.

Upon reviewing *Table 4*, it becomes apparent that the regression coefficients for the "vehicle age" and "Category 4 of vehicle group" variables are statistically significant at the  $\alpha = 0.05$  level. In other words, the t-statistic, derived from dividing the coefficient value (B) by the standard error, shows a significant difference from zero at the  $\alpha = 0.05$  level, indicating that the regression coefficients for these variables have a statistically meaningful impact. Interpreting the regression coefficient for "vehicle age," which is -

50.092, can be expressed as follows: With each one-year increase in the age of the vehicle, the amount of claimed damages decreases by 50.092 units, equivalent to 5,009,200 Iranian rials.

Additionally, considering that the regression coefficient for "Category 4 of vehicle group" is 67.405 and given that Category a is the reference Category for the vehicle group (Category 1), the interpretation of this coefficient is as follows: vehicles belonging to Category 4 of the vehicle group exhibit a 67.405-unit increase in the claimed damages compared to Category a (Category 1).

Category D of the "vehicle group" variable has a higher claimed damages amount of 67.405 units than Category A. This suggests that vehicles falling under Category D of the vehicle group tend to have significantly higher claimed damages than those in Category A.

These interpretations provide insights into how changes in the independent variables influence the predicted values and contribute to a better understanding of the relationships within the regression model.

The flowchart in *Fig.7* displays the sequence of calculations performed within the MODELER software. This visual representation illustrates the step-by-step process and dependencies between different computations and data manipulations within the software, providing a clear overview of the analytical workflow. Analyzing this flowchart helps users understand the data processing and modeling steps, ensuring transparency and reproducibility in the analysis.

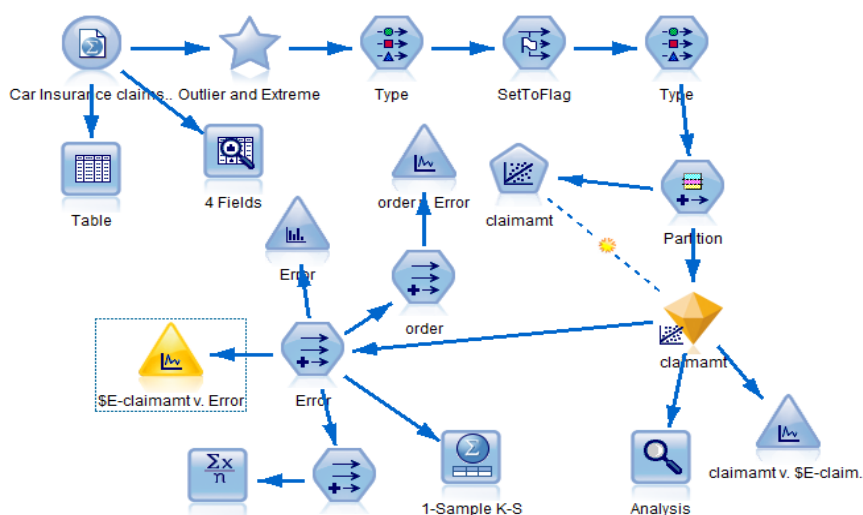


Fig. 7. Sequence of calculations performed within the MODELER.

## 6 | Conclusion

In this study, we used data from 128 customers of Iran insurance company to predict the amount of damage claims. The dataset included features such as the policyholder's age, vehicle type, vehicle age, the number of claim occurrences, and the number of claims (in thousands of Tomans) collected in 1400. Except for the claim amount, all features were discretized as ordinal variables to improve the analysis and handle outlying data. Subsequently, we employed multiple linear regression to estimate the target field and investigate the influence of each of these variables on predicting the desired damage amount. Using IBM SPSS modeler for data analysis, we evaluated the assumptions associated with regression. These assumptions included normality of error distribution, lack of autocorrelation in errors, and constant error variance, all of which were satisfied. The adjusted R square value for the regression model was 0.571, indicating that the variables, including vehicle group Categories 2, 3, and 4, vehicle age, and policyholder's age, collectively explain 57% of the variance in the target variable, which is the claimed damage amount. The analysis also showed that the MSR was significantly larger than the mean square of error (MSE). The F-statistic in the ANOVA table confirmed this by indicating that MSR is 16.451

times larger than MSE, demonstrating that the model adequately predicts the target values. These findings suggest the effectiveness of the multiple linear regression model in predicting damage claims.

Overall, this study provides insights into the factors influencing insurance claim amounts and demonstrates the model's ability to predict these amounts based on customer and vehicle characteristics. The regression model met the key assumptions, confirming its suitability for the given dataset.

## 6.1 | Future Research Directions

While this study provides valuable insights into predicting damage claims within the insurance domain, several lines for future research can be explored. Extending the dataset to include more diverse and detailed variables, such as driving behavior data or geographical information, can enhance predictive accuracy. Additionally, employing advanced machine learning techniques, like deep learning or ensemble methods, may offer superior predictive capabilities. Furthermore, investigating the temporal aspects of claims data and the evolving impact of socio-economic factors could lead to more dynamic predictive models, considering changing circumstances.

## 6.2 | Limitations of This Research

This research has certain limitations that should be acknowledged. First, the study used a relatively modest dataset, which might limit the generalizability of the findings. A larger and more diverse dataset would allow for more robust model development. Additionally, the study assumed linear relationships between predictor variables and claim amounts, potentially overlooking nonlinear effects that may exist in the real world. Furthermore, the research focused on quantitative factors, ignoring potential qualitative influences on claims, such as customer behavior, or external factors like natural disasters. Lastly, as with any predictive model, there is always an inherent level of uncertainty in real-world predictions that should be considered in practical applications.

## References

- [1] Rasi Nojehdehi, R., Bagherzadeh Valami, H., & Najafi, S. E. (2023). Classifications of linking activities based on their inefficiencies in network DEA. *International journal of research in industrial engineering*, 12(2), 165–176. [https://www.riejournal.com/article\\_178844.html](https://www.riejournal.com/article_178844.html)
- [2] Rasinojehdehi, R., & Valami, H. B. (2023). A comprehensive neutrosophic model for evaluating the efficiency of airlines based on SBM model of network DEA. *Decision making: applications in management and engineering*, 6(2), 880–906. <https://dmame-journal.org/index.php/dmame/article/view/729>
- [3] Azizi, S., & Mohammadi, M. (2023). Strategy selection for multi-objective redundancy allocation problem in a k-out-of-n system considering the mean time to failure. *Opsearch*, 60(2), 1021–1044. <https://doi.org/10.1007/s12597-023-00635-2>
- [4] Najafi, E., Aryanezhad, M., & others. (2011). A BSC-DEA approach to measure the relative efficiency of service industry: a case study of banking sector. *International journal of industrial engineering computations*, 2(2), 273–282. [http://growingscience.com/ijiec/Vol2/IJIEC\\_2010\\_20.pdf](http://growingscience.com/ijiec/Vol2/IJIEC_2010_20.pdf)
- [5] Nojehdehi, R. R., Abianeh, P. M. M., & Valami, H. B. (2012). A geometrical approach for fuzzy production possibility set in data envelopment analysis (DEA) with fuzzy input-output levels. *African journal of business management*, 6(7), 2738. <https://www.researchgate.net/profile/Hadi-Bagherzadeh>
- [6] Bagherzadeh Valami, H., & Raeinojehdehi, R. (2016). Ranking units in data envelopment analysis with fuzzy data. *Journal of intelligent & fuzzy systems*, 30, 2505–2516. DOI:10.3233/IFS-151756
- [7] Afshar-Nadjafi, B., Pourbakhsh, H., Mirhabibi, M., Khodaei, H., Ghodami, B., Sadighi, F., & Azizi, S. (2019). Economic production quantity model with backorders and items with imperfect/perfect quality options. *Journal of applied research and technology*, 17(4), 250–257. [https://www.scielo.org.mx/scielo.php?script=sci\\_arttext&pid=S1665-64232019000400250](https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-64232019000400250)
- [8] Ye, C., Zhang, L., Han, M., Yu, Y., Zhao, B., & Yang, Y. (2022). Combining predictions of auto insurance claims. *Econometrics*, 10(2). <https://www.mdpi.com/2225-1146/10/2/19>

- [9] David, M. (2015). Auto insurance premium calculation using generalized linear models. *Procedia economics and finance*, 20, 147–156. <https://www.sciencedirect.com/science/article/pii/S2212567115000593>
- [10] Meng, S., Wang, H., Shi, Y., & Gao, G. (2022). Improving automobile insurance claims frequency prediction with telematics car driving data. *ASTIN bulletin*, 52(2), 363–391. DOI:10.1017/asb.2021.35
- [11] Selvakumar, V., Satpathi, D. K., Kumar, P. P., & Haragopal, V. V. (2021). Predictive modeling of insurance claims using machine learning approach for different types of motor vehicles. *Accounting and finance*, 9(1), 1–14.
- [12] Jaworski, J., & Czerwonka, L. (2021). Determinants of enterprises' capital structure in energy industry: evidence from European Union. *Energies*, 14(7), 1–21. <https://www.mdpi.com/1996-1073/14/7/1871>
- [13] Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Machine learning-based regression framework to predict health insurance premiums. *International journal of environmental research and public health*, 19(13). <https://www.mdpi.com/1660-4601/19/13/7898>

