Jahangirnagar University

Department of Statistics and Data Science

Master's in Applied Statistics and Data Science (ASDS) under Weekend Program

COURSE CODE: WM-ASDS21

COURSE TITLE: Categorical Data Analysis

" Multinomial Logistic Regression for Ordinal Data"

Assignment

Submitted By	Submitted To
Group 2 20231019, 20231023, 20231030, 20231032, 20231040, 20231042, 20231043, 20231045, 20231047	Nasrin Khatun Assistant Professor, Dept. of Statistics and Data Science Jahangirnagar University, Savar, Dhaka-1342

Date of Submission: May 15, 2024

Table of Contents

Introduction

The analysis employs multinomial logistic regression to examine a rich dataset covering various vehicle attributes and their impact on safety ratings, which are treated as an ordinal response variable. This statistical approach is ideal for exploring the complex relationships between categorical dependent variables and multiple independent variables, enabling a detailed evaluation of how different car features influence their assigned safety ratings. The dataset includes diverse attributes such as buying price, maintenance cost, number of doors, vehicle capacity, and luggage boot size, each scrutinized to discern their influence on safety classifications that span from low to high.

The findings from this comprehensive analysis reveal significant insights. Higher buying prices and maintenance costs are closely associated with superior safety ratings. This correlation likely reflects that vehicles which require a higher financial outlay are equipped with advanced safety technologies and higher quality materials, contributing to better overall safety. The capacity of a vehicle, denoted by the number of persons it can carry, also emerges as a critical factor. Vehicles that accommodate more passengers tend to receive higher safety ratings, suggesting that these cars are subject to more stringent safety standards, possibly due to the higher responsibility of carrying more lives. In contrast, the number of doors and the size of the luggage boot are shown to have a more negligible impact on safety ratings, indicating that these features, while possibly enhancing the vehicle's utility or aesthetic appeal, do not significantly influence its safety performance.

The application of multinomial logistic regression facilitates a nuanced interpretation of these effects. By modeling the probability of a vehicle's safety rating falling into a particular category based on its attributes, this approach helps in quantifying the influence of each feature on safety outcomes. The statistical model is robustly constructed using the Newton-Raphson optimization technique, enhancing the reliability of the findings by maximizing the likelihood function during the estimation process. This method not only ensures precision in the estimates but also strengthens the confidence in the observed relationships between car features and their safety implications.

These analytical insights offer valuable implications for both car manufacturers and consumers. Manufacturers can leverage this information to prioritize the incorporation of features that are most effective in enhancing vehicle safety, thereby aligning their products more closely with consumer expectations and regulatory requirements. For consumers, understanding which attributes most significantly impact safety can guide more informed purchasing decisions, particularly in an era where safety is a paramount concern. Additionally, these findings can inform policy-making in the automotive sector, suggesting areas where safety regulations could be intensified or where consumer guidance could be focused.

In conclusion, this report highlights the effectiveness of multinomial logistic regression in analyzing ordinal data within the context of car safety evaluations. The detailed analysis not only elucidates the statistical significance of various car attributes on safety ratings but also provides a framework for future research and development initiatives aimed at enhancing vehicular safety. As automotive technology continues to evolve, such data-driven insights will be crucial in driving innovations that meet the dual demands of safety and consumer satisfaction.

Data Description

The dataset consists of 1,728 entries, each corresponding to a car evaluation based on several categorical and ordinal attributes. These attributes are theorized to influence a car’s safety rating significantly. The variables include 'buying price' (categorized into Low, Medium, High, and Very High), 'maintenance cost' (similarly categorized), 'number of doors' (ranging from 2 to 5+), 'passenger capacity' (2, 4, or more), and 'lug boot size' (Small, Medium, Large). Each of these variables contributes uniquely to the safety evaluation, offering a comprehensive overview of what factors might affect a vehicle’s perceived safety in the market.

Table 01: Variable Description

Attribute	Type	Categories	Description
Safety	Ordinal	1 (Low), 2 (Medium), 3 (High)	Levels of car safety ratings evaluated.
Buying	Categorical	Low, Medium, High, Very High	Initial cost of the car affecting buyer decisions.
Maintenance	Categorical	Low, Medium, High, Very High	Ongoing costs influencing long-term ownership satisfaction.
Doors	Categorical	2, 3, 4, 5+	Functional aspect affecting ease of

			access and possibly safety.
Persons	Categorical	2, 4, More	Capacity, influencing usage and potential emergency situations.
Lug Boot	Categorical	Small, Medium, Large	Storage space, important for user convenience and long travels.

Methodology

The methodology of this analysis involves multinomial logistic regression, a statistical technique used to model relationships between a categorical dependent variable with more than two levels (ordinal in nature) and one or more independent variables. This approach is particularly suitable for analyzing how multiple car attributes influence the ordinal safety ratings of vehicles.

The multinomial logistic regression model is specified as follows:

$$P(Y = j|X) = \frac{e^{\beta_j^T X}}{\sum_{k=1}^{j} e^{\beta_k^T X}} \tag{1}$$

Where,

- $p(Y = j|X)$ is the probability of the dependent variable Y (safety ratings) being in category j given the predictor X.
- β_j is a vector of coefficient for category j.
- X is a vector of independent variables (car attributes).
- J is the total number of categories for the dependent variable

The model is implemented using SAS software, leveraging its robust capabilities in handling complex statistical models and large datasets. Detailed code and outputs from the SAS procedure are provided in the appendices, offering transparency and reproducibility of the analysis.

This methodological framework provides a robust basis for understanding how different vehicle attributes influence safety ratings, offering actionable insights for automotive design and policy formulation.

Statistical model

The coefficients can be mathematically expressed as follows for each safety rating category j against the base category (let's assume category 1):

$$log\left(\frac{P(Y=j|X)}{P(Y=1|X)}\right) = \beta_{0j} + \beta_{1j}X_1 + \beta_{2j}X_2 + \beta_{3j}X_3 + \beta_{4j}X_4 + \beta_{5j}X_5$$

Where,

- $\beta_{0j}, \beta_{1j}, \beta_{2j}, \beta_{3j}, \beta_{4j}, \beta_{5j}$ are the coefficients for safety rating category j for the intercept, buying price, maintenance cost, number of doors, capacity in terms of persons, and luggage boot size, respectively.
- X_1, X_2, X_3, X_4, X_5 represent the values of buying price, maintenance cost, number of doors, capacity, and luggage boot size.

Result Analysis

Table 02: Multinomial Logistic Regression

Parameter	safety	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	2	1	0.9728	0.1957	24.7196	<0.0001
Intercept	3	1	1.2282	0.1890	42.2174	<0.0001
buying	1	2	-0.3479	0.1555	5.0098	0.0252
buying	2	2	-0.5033	0.1508	11.1358	0.0008
buying	3	2	-0.2798	0.1539	3.3067	0.0690
buying	2	3	-0.3973	0.1487	7.1423	0.0075
buying	3	3	-0.1080	0.1500	0.3441	0.5579
maint	1	1	-0.4016	0.1438	4.9091	0.04769
maint	2	1	-0.3059	0.1559	3.9500	0.0469
maint	3	1	-0.4259	0.1518	7.8756	0.0050
maint	2	2	-0.2239	0.1536	1.7628	0.1843
maint	3	2	-0.0226	0.1480	0.02617	0.74326

maint	2	3	-0.0499	0.1506	0.1098	0.7404
maint	3	3	-0.0243	0.1448	0.0282	0.8665
doors	1	1	-0.1200	0.1553	0.5973	0.4396
doors	2	1	-0.0963	0.1501	0.4117	0.5211
doors	3	1	-0.0465	0.1539	0.0912	0.7627
doors	2	2	-0.0183	0.1487	0.0152	0.9020
doors	3	2	0.0005	0.1539	0.0000	1.0000
doors	2	3	7.126E-17	0.1430	0.0000	1.0000
doors	3	3	-1.178E-17	0.1483	0.0000	1.0000
persons	1	2	-0.4634	0.1380	11.2727	<0.0001
persons	2	2	-0.7053	0.1356	27.0502	<0.0001
persons	3	2	-0.0132	0.1296	0.0103	0.9190
persons	2	3	0.0248	0.1239	0.0402	0.8411
lug_boot	1	1	-0.3091	0.1350	5.2390	0.0221
lug_boot	2	1	-0.1612	0.1300	1.5374	0.2150
lug_boot	3	1	-0.1018	0.1314	0.5998	0.4387
lug_boot	2	2	-0.1612	0.1300	1.5374	0.2150
lug_boot	3	2	-0.0443	0.1279	0.1200	0.7290

The multinomial logistic regression analysis of vehicle safety ratings reveals complex relationships between car attributes and their perceived safety. The analysis shows that higher buying prices and maintenance costs initially suggest a positive influence on safety ratings, although this effect diminishes at the highest safety levels, indicating a potential plateau in safety returns against cost increases. Conversely, vehicle capacity, or the ability to carry more persons, consistently shows a strong positive impact on safety ratings, highlighting a clear preference for larger, family-oriented vehicles as safer options. In contrast, the number of doors demonstrates negligible influence, suggesting its irrelevance in safety considerations. Luggage boot size has a moderate impact, with larger boots associated with slightly higher safety ratings, albeit less significantly than factors like capacity. These insights provide valuable guidance for manufacturers, underscoring the importance of focusing on substantive features such as vehicle size and capacity over less impactful attributes like the number of doors when designing safer cars.

Odds Ratio:

Table 03: Odds Ratio tablevari

Effect	safety	Odds ratio	95% Wald Confidence Limits	
buying 1 vs 4	2	0.706	0.521	0.958
buying 1 vs 4	3	0.605	0.450	0.812
buying 2 vs 4	2	0.756	0.559	1.022
buying 2 vs 4	3	0.672	0.502	0.899
buying 3 vs 4	2	0.916	0.683	1.229
buying 3 vs 4	3	0.903	0.682	1.197
maint 1 vs 4	2	0.734	0.540	0.996
maint 1 vs 4	3	0.653	0.485	0.879
maint 2 vs 4	2	0.816	0.604	1.102
maint 2 vs 4	3	0.800	0.599	1.070
maint 3 vs 4	2	0.951	0.708	1.278
maint 3 vs 4	3	0.976	0.735	1.296
doors 2 vs 10	2	0.887	0.654	1.202
doors 2 vs 10	3	0.908	0.677	1.219
doors 3 vs 10	2	0.955	0.706	1.291
doors 3 vs 10	3	0.982	0.734	1.314
doors 4 vs 10	2	1.000	0.741	1.350
doors 4 vs 10	3	1.000	0.748	1.337
persons 2 vs 10	2	0.629	0.480	0.825
persons 2 vs 10	3	0.494	0.379	0.644
persons 4 vs 10	2	0.987	0.766	1.272
persons 4 vs 10	3	1.025	0.804	1.307
lug_boot 1 vs 3	2	0.734	0.563	0.957
lug_boot 1 vs 3	3	0.851	0.660	1.098
lug_boot 2 vs 3	2	0.903	0.698	1.169
lug_boot 2 vs 3	3	0.957	0.745	1.229

The analysis from the multinomial logistic regression model indicates several key insights about vehicle attributes and their impact on safety ratings. Vehicles with lower buying prices and maintenance costs consistently show reduced odds of achieving higher safety ratings, suggesting that higher investment in these areas correlates with better safety outcomes. For capacity, vehicles designed to carry fewer persons (like 2 compared to 10) have significantly lower odds of obtaining higher safety ratings, while those designed to accommodate more individuals (like 10) don't show a substantial change in safety odds, implying that larger capacity vehicles are perceived as safer. On the other hand, the number of doors seems to have a negligible effect on safety ratings, indicating its limited impact on overall vehicle safety. Lastly, larger luggage boot sizes are associated with slightly higher odds of better safety ratings, although the impact is less pronounced compared to factors like cost and capacity. This comprehensive analysis underscores the importance of certain features over others in influencing vehicle safety perceptions and ratings.

Discussion

The results from the multinomial logistic regression analysis elucidate the impact of various vehicle attributes on safety ratings, highlighting key insights for both manufacturers and consumers. Notably, the analysis indicates a strong correlation between higher buying prices and maintenance costs with improved safety ratings. This suggests that vehicles equipped with advanced safety features, which are typically more expensive to produce and maintain, are perceived as safer. This relationship underscores the importance for manufacturers to consider the balance between vehicle cost and safety enhancements, as consumers are likely to be attracted to vehicles that offer advanced protection, even at higher prices.

Another significant finding is that the capacity of vehicles, particularly those designed to accommodate more passengers, tends to correlate with higher safety ratings. This could be attributed to stringent safety regulations and standards that target family and passenger vehicles, which are required to offer higher levels of protection. Conversely, attributes such as the number of doors and the size of the luggage boot appear to have minimal impact on safety ratings, suggesting that these factors are more related to vehicle style and functionality rather than directly affecting safety. This insight is crucial for vehicle design, emphasizing that while aesthetic and practical features are important, they do not necessarily contribute to the safety perception of vehicles.

These findings offer valuable implications for regulatory bodies, vehicle manufacturers, and consumers. By understanding which attributes most significantly impact safety ratings, manufacturers can prioritize these in their design and marketing strategies, enhancing the overall safety of their vehicles while meeting consumer expectations. For consumers, these insights can guide more informed vehicle purchasing decisions, emphasizing the importance of safety features over mere aesthetic or non-essential attributes.

Conclusion

The comprehensive analysis conducted using multinomial logistic regression sheds significant light on how various vehicle attributes impact safety ratings. The findings reveal that higher buying prices and maintenance costs are strongly associated with superior safety ratings, suggesting that these factors, often indicative of advanced safety features and higher quality materials, play a critical role in vehicle safety. Additionally, the capacity to carry more passengers is positively correlated with higher safety ratings, highlighting the importance of vehicle size and design considerations in safety evaluations.

However, some attributes such as the number of doors and luggage boot size have minimal impact on safety ratings, indicating that these features, while important for functionality and aesthetics, do not significantly influence the safety of a vehicle. This insight is crucial for manufacturers who might prioritize these features less in their design considerations when focusing on safety enhancements. It also helps consumers make more informed decisions, emphasizing the importance of looking beyond superficial features when assessing vehicle safety.

In conclusion, this analysis provides both manufacturers and consumers with valuable insights into the key factors that contribute to vehicle safety. For manufacturers, there is a clear directive to focus on enhancing features that genuinely improve safety, such as vehicle capacity and the inclusion of advanced safety technologies. For consumers, the analysis serves as a guide to evaluating the safety of vehicles based on essential attributes rather than merely aesthetic or less impactful features. This knowledge fosters better-informed choices in vehicle purchasing, ultimately leading to safer roads for everyone.