Corpus Intellectual
ISSN PRINT 2811-3187 ONLINE 2811-3209 Volume 4 NO 1 Conf. Edition 2025

An Investigative Comparison of Bayesian and Classical Logistic Regression in Modeling Motor Insurance Preferences

Lead Author

Dorcas Modupe OKEWOLE

Affiliation:

Department of Mathematics and Statistics, Redeemer's University, Ede, Osun State

Abstract

This study illustrates the importance of investigative analysis (IA) in a statistical modeling activity. Wrong inferences are inevitable when a thorough investigation of the model variables is missing. The concept could be considered exploratory data analysis (EDA) but labeled as investigative analysis to portray the focus: investigating the particular specification of the model variables that leads to the correct inference. Logistic regression analysis was carried out to investigate factors influencing motor insurance subscription preferences, explicitly focusing on third-party and comprehensive insurance policies, using both Bayesian and classical approaches. The dataset of 59 subscribers includes variables such as type of policy, age, sex, profession, and area of residence. The dependent variable is the type of motor insurance policy subscribed to (third-party or comprehensive).Akaike Information criteria (AIC), Residual Deviance (RD), and area under the curve (AUC) showed that the final specification of the model (AUC = 79.46, RD = 67.46, AUC = 0.7417) was better than the first (AUC = 88.643, RD = 66.643, AUC = 0.7292).The analysis identified profession as a significant factor influencing the choice between third-party and comprehensive insurance, implying that individuals in certain professions are more likely to opt for comprehensive coverage. In contrast, others would prefer the third-party type and suggest the need for targeted strategies that consider specific professional groups. More importantly, the result revealed that, while profession plays a crucial role in motor insurance subscription decisions in the dataset, some other variable specifications had a contradicting result. The research underscores the

significance of investigative analysis in drawing out the information in a dataset under study. Careful exploration and analysis would ensure the elicitation of vital information.

Keywords: Investigative Analysis, Bayesian Estimation, Variable specification, logistic regression, classification.

Co-Authors: Shekemi Marvelous ODEYEMI, Ayobami Fadilat AKINTOLA and Olayinka Olusegun OLADIPUPO, Department of Mathematics and Statistics, Redeemer's University, Ede

1. Introduction

Statistical methodologies form a highly substantial component in various kinds of research. The significance of investigative analysis while applying statistical methods can be seen in Davidson's review of investigative projects (2023). Often, researchers approach statistical analysis as a one-way traffic of picking up a model, applying it to the data under consideration, determining the significance of the independent variables, and proceeding to conclude. However, detailed statistical investigations may provide more valid contrasting information while highlighting the need for further investigation, including more data. This study presents a case on analyzing a logistic regression model involving the choice of the motor insurance policy. The study highlights that essential information and inferences could still be drawn from available data, even when small. Logistic regression has continued to be a veritable tool for modeling binary classification, predicting the probability of an event occurring. In the simplest case, it describes the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. It is often necessary to validate the result of a statistical analysis using another method from which the similarity of solutions would assure their reliability. We consider the Bayesian and classical approaches for this purpose.

This study examines some factors that could influence the choice of motor insurance policy. Browne and Kim (1993**)** stated that occupation significantly influences insurance purchasing behavior, particularly in motor and health insurance, as job type correlates with income and perceived risk exposure. More recently, Ghafoor et al. (2016) also support this, showing how demographic and occupational factors shape preferences in insurance purchase decisions in emerging markets. There is a strong link between occupation, education level, and insurance literacy. Highly educated professionals are likelier to understand policy terms, compare different products,

and select insurance based on value rather than cost alone (Ahmed et al., 2019). As such, their occupation indirectly affects insurance preference through enhanced decision-making capability. Motor insurance preference is also shaped by how individuals use their vehicles, which is often tied to their job type. For instance, self-employed individuals or those in sales and marketing roles typically use their vehicles extensively, increasing the likelihood of claims and prompting a preference for policies with broader protection (Kumar & Bansal, 2014). In contrast, those with sedentary occupations may prefer less expensive, limited coverage.

Motor insurance in Nigeria is a vital financial protection mechanism for vehicle owners. Compulsory Third Party Liability (CTPL) insurance is mandatory for all vehicle owners in compliance with Nigerian law. This type of insurance covers injury, death, or property damage caused to third parties by the insured vehicle. In addition to Compulsory Third Party Liability (CTPL), many vehicle owners opt for Comprehensive Motor Insurance. This broader coverage extends beyond third-party liability to include protection against damages to the insured vehicle, theft, and other non-collision incidents.

Applications of logistic regression are vast; for example, Gifford and Bayrak (2023) constructed predictive analytics models to forecast the North American National Football League (NFL) game outcomes in a season using decision trees and logistics regression. The binary win-loss outcome measure was the dependent variable, with several independent variables. The authors constructed decision tree and binary logistic regression models to describe the relationships between the predictors and football game outcomes in the NFL. Moomen et al. (2019) used the logistic model to study the factors influencing truck crashes and downgrades in Wyoming, US. Some other applications are Robles-Velasco et al. (2020) on a study of pipe failures in water supply networks in a Spanish city and Sze et al. (2014) on applying logistic regression to a study on road safety issues.

This study aims to contribute to statistical modeling issues by emphasizing the need for data exploration and consequent careful specification of the independent variables, using a case of motor insurance preferences.

2. Data, Study Variables and Measurements

The data is the Motor insurance policy subscription from Prestige Assurance Plc, one of Nigeria's leading general insurance Companies

with offices nationwide. It contains data on 59 subscribers made available at one of the offices in Lagos.
The variables and Measurements are as follows:

Dependent Variable
Type of Policy - Categorical (Third Party, Comprehensive)
Independent Variables
Age – Numerical (Years)
Sex – Categorical (Male, Female)
Area of Residence - Qualitative
Profession – Qualitative

3. Model
The fundamental concept is to transform the output of a linear regression model into a probability using the logistic function (also known as the sigmoid function). This transformation ensures that the predicted values lie between 0 and 1, representing probabilities. The model for the study is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \quad (1)$$

Y is the dependent variable, representing the linear combination of the independent variables and parameters, $X_i's$ are the independent variables, and $\beta_i's$ are the coefficients.
X_1 = Age
X_2 = Sex
X_3 = Profession
X_4 = area of resident
Logistic regression starts by estimating the linear combination of the input features (independent variables) weighted by coefficients as given in equation (1). The linear combination is then imputed into the logistic function given as:

$$\text{Log odds} = logit = ln\left(\frac{P_X}{1-P_X}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \quad (2)$$

Where,
$\widehat{odds} = \exp(logit)$

The logistic function is therefore

$$p_x = \frac{\exp(logit)}{1+\exp(logit)} = \frac{\widehat{odds}}{1+\widehat{odds}} \quad (3)$$

The Bayesian Model

The Bayesian estimation approach involves specifying a probability function to represent the prior belief on the parameter of interest. The prior distribution is combined with the likelihood function based on the

data, leading to the posterior distribution on which parameter estimates are drawn.

Prior Distribution

We use the normal distribution with the mean stated as zero and a large value (22) set for the variance to represent insufficient information. $\beta_j \sim N(\mu_j, \sigma_j^2),\ j = 0,1,2,3,4$

Likelihood function

Since the dependent variable(Y_i) is nominal of the binary case, we specify the Bernoulli distribution for each Y_i

$$f(Y_i) = p^{y_i}(1-P)^{(1-y_i)}$$
$$= \left(\frac{\exp(\widehat{logit})}{1+\exp(\widehat{logit})}\right)^{y_i}\left(1-\frac{\exp(\widehat{logit})}{1+\exp(\widehat{logit})}\right)^{(1-y_i)} \quad (4)$$

$$L(B/Y,X) = \prod_{i=1}^{n}\left[\left(\frac{\exp(\widehat{logit})}{1+\exp(\widehat{logit})}\right)^{y_i}\left(\frac{\exp(\widehat{logit})}{1+\exp(\widehat{logit})}\right)^{(1-y_i)}\right] \quad (5)$$

Where $B = \beta_0, \beta_1, \beta_2, \beta_3, \beta_4$

We assume that the individual choices of the motor insurance policy type are independent, which implies that the likelihood function over a data set of *n* subjects is as follows:

Posterior Distribution

This is the product of the likelihood function and the prior distribution

$$g(B/Y) = \prod_{i=1}^{n}\left[\left(\frac{\exp(\widehat{logit})}{1+\exp(\widehat{logit})}\right)^{y_i}\left(\frac{\exp(\widehat{logit})}{1+\exp(\widehat{logit})}\right)^{(1-y_i)}\right] X \prod_{j=0}^{4}\frac{1}{\sigma_j\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{\beta_j-\mu_j}{\sigma_j}\right)^2\right\} \quad (6)$$

Through numerical integration implemented by brms package in R, the estimation involves drawing samples from the posterior distribution, and a measure of central tendency is applied to represent the posterior estimate, the Bayesian estimate.

Data

The data source is the Prestige Assurance PLC, Lagos, on 59 motor insurance subscribers. It contains the kind of motor insurance the people subscribed to and other information such as age, gender,

profession, and area of residence. In the original data, the measurements of the variables are as follows:

Age – quantitative (age last birthday in years)
Sex – qualitative (Male, Female)
Area of residence – qualitative (respondents' area of residence, such as Isolo Lagos, Gbagada, Somolu, etc.)
Profession – qualitative (subscriber's profession, such as Manufacturing, Engineering, Pharmacy, Trading, Nursing, Civil service, Retired, etc.)

4. Results and Discussion

Exploratory data analysis provides essential guidance on the inferential analysis required. It enables an overview of the data prior to the analysis. An exploration of the descriptive statistics on the variables (Table 1) drew attention to the possibility of the profession being significant, depending on the grouping. There were more subscribers to the comprehensive policy than third-party policy in groups 3, 5, and 7, while subscribers in groups 4 and 6 were more of the third party, emphasizing the possibility of the significance of the profession variable. The first attempt at the modeling (Table 2) showed all the independent variables as insignificant (at the 5% level).The earlier indication (Table 1) of the possibility of the profession being significant suggests the need to regroup the variable.

The results of both estimation methods (Bayesian and classical) are similar in terms of which independent variable is significant in the model and which is not. In practice, the Bayesian 95% confidence interval serves as a significance test: a CI containing zero indicates non-significance, while one that does not include zero implies significance. In this paper, the L-95% CI and U-95% CI represent the CI's lower and upper limits, respectively.

Table 1: Cross-tabulation of the Qualitative Variables with Type of Policy in A First grouping

Variable	Groups	Type of Policy		Total
		Third-Party	Comprehensive	
Sex	Male	14	22	36
	Female	10	13	23
Area of residence	1	3	6	9
	2	16	23	39
	3	5	6	11
Profession	1	1	0	1

	2	1	0	1
	3	1	2	3
	4	2	1	3
	5	4	16	20
	6	2	0	2
	7	13	16	29

Table 2: Logistic Regression with the first grouping of the Profession variable.

	Classical		Bayesian		
	Estimate	P-value	Estimate	L-95% CI	U-95% CI
(Intercept)	1.18017	0.3839	1.54	-1.30	4.56
Age	-0.02078	0.4897	-0.03	-0.09	0.04
Sex	0.28605	0.6596	0.34	-1.02	1.72
Area_of_Residence_1	-0.41090	0.6215	-0.51	-2.33	1.23
Area_of_Residence_2	-0.40387	0.7146	-0.48	-2.90	1.96
Proffession_1	-17.47850	0.9965	-17.14	-49.44	0.34
Proffession_2	-17.35382	0.9965	-16.70	-47.62	0.33
Proffession_3	0.47134	0.7288	0.80	-2.27	4.41
Proffession_4	-0.69774	0.6122	-1.01	-4.57	2.03
Proffession_5	1.30701	0.0759	1.54	0.04	3.19
Proffession_6	-17.07881	0.9951	-16.99	-48.66	0.00

Table 3: Cross-tabulations ofAll Independent Variables with the Final Grouping of The Profession Variable

Variable	Categories	Type of Policy		Total
		Third-	Comprehensive	
Sex	Male	14	22 (61.1%)	36
	Female	10	13 (56.5%)	23
Area of Residence	Mainland	16	23 (59.0%)	22
	Island	3 (33.3%)	6 (66.7%)	26
	Other	5 (45.5%)	6 (54.5%)	11

Profession	Group 1	8 (23.5%)	26 (76.5%)	34
	Group 2	16	9 (36.0%)	25

Further regroupings were carried out and analyzed until a particular grouping criterion (Table 3) finally presented the variable as significant. The regroupings follow from the fact that the distribution across categories of the variable revealed the pattern of third-party versus comprehensive insurance preferences. In the final regrouping, professions that are more similar in their third-party versus comprehensive insurance preference were in the same group. Of course, interpretations of categorical variables should be specific to the groups involved. Table 3 contains the composition of the final form of the variables in the model relative to the type of policy. For instance, for males and females, there are more subscribers to comprehensive insurance, which suggests that the type of policy is not associated with sex. The area of residence revealed a similar pattern to that of sex. The case is, however, different in the Profession data. To further buttress the point of this investigative analysis, it is worthy of note that it is not in every dataset that regrouping the categorical variable will produce different results. There are instances where the result remains the same irrespective of the grouping criteria adopted. One such instance is the case of the area of residence in this study. Despite all the regroupings carried out, the results consistently indicated that the area of residence does not affect the subscriber's choice of motor insurance policy.

In the final grouping (Table 3) of the profession variable, there were two groups, Group 1 and Group 2, each representing distinct characteristics among policyholders. Upon this final grouping, the profession emerged as the sole significant variable (Table 4) in predicting motor insurance subscription types in the dataset. Group 1, consisting of professions and sectors such as business, finance, law, Health, management, insurance, and Haulage, exhibited a higher likelihood of comprehensive insurance subscriptions than Group 2, which comprised professions and sectors such as trading, freelance, manufacturing; marketing; hospitality; logistics; consultancy; farming; civil service; restaurant; retired; who mostly preferred third-party insurance.

Table 4: Logistic Regression with the Final Regroupingof the Profession Variable

	Classical		Bayesian		
	Estimate	P-value	Estimate	L-95% CI	U-95% CI
(Intercept)	2.79	0.0386	3.25	0.57	6.24
Age	-0.04	0.1595	-0.04	-0.10	0.01
Sex	0.11	0.8623	0.11	-1.28	1.46
Area_of_Residence_1	-0.03	0.9705	-0.09	-1.94	1.67
Area_of_Residence_2	-0.23	0.8289	-0.33	-2.61	1.85
Proffession_Group_2	-1.77	0.0036	-1.98	-3.39	-0.70

We obtained the Akaike Information Criterion (AIC), Residual deviance (RD), and Area under the curve (AUC) for the classical model as a comparison of the first grouping and final grouping models. In comparing model specifications, the model with the smaller AIC, higher RD, and AUC has the best fit. As expected, the AIC was lower for the final model, while RD and AUC were higher than the first (Table 5). This result further supports that the final grouping fit the data better than the first grouping.

Table 5: Comparison of the Model under the First Grouping and the Final Grouping

	Model with the First Grouping	Model with the Final Grouping
AIC	88.643	79.46
RD	66.643	67.460
AUC	0.7292	0.7417

The significance of the profession variable suggests that certain occupations are more likely to opt for comprehensive insurance coverage over third-party insurance. The result revealed that individuals in professions requiring high levels of risk management, such as finance and law, were more inclined towards comprehensive insurance—conversely, those in less financially secure professions, such as freelancers and retired, preferred third-party coverage.

Conclusion

This study underscores the importance of investigative analysis, as it identified an essential predictor of the dependent variable, which would have remained hidden. The Results align with previous studies (Brown & Kim, 1993; Ghafoor et al., 2016; Ahmed et al., 2019), highlighting the influence of socioeconomic factors, including occupation, on insurance preferences. Grouping professions provides a nuanced understanding of how specific occupational categories contribute to insurance subscription patterns. The findings emphasize the need for insurance companies to tailor their marketing strategies and product offerings based on occupation-specific insights. Understanding these nuances can help insurers better meet the diverse needs of their customer base and improve their competitive edge in the insurance market. The variable specification issue, which was the focus of this paper, is one of the many issues that researchers should consider from the exploratory stage to the modeling and analysis stage. Approaching the application of statistical methodologies in research could be implemented in an investigative manner so that results represent actual realities. This study's limitation is the small sample size and number of independent variables. A more elaborate study might provide further information.

References

Ahmed, H., Tufail, M., & Farooq, S. (2019). Determinants of insurance purchasing behavior in Pakistan. Journal of Risk and Insurance, 86(3), 755–780.

Browne, M. J., & Kim, K. (1993). An International Analysis of Life Insurance Demand. Journal of Risk and Insurance, 60(4), 616–634.

Davidson, A. (2023). A Review of the Use of Investigative Projects in Statistics and Data Science Courses. Journal of Statistics and Data Science Education, 32(2), 199–201. https://doi.org/10.1080/26939169.2023.2240385.

Ghafoor, A., Khan, M. N., & Shaikh, F. M. (2016). The impact of demographic factors on the demand for general insurance in Pakistan. Journal of Finance and Economics Research, 1(1), 31–42. https://doi.org/10.20547/jfer1601103

Kumar, N., & Bansal, S. (2014). Motor insurance preferences in urban India: An empirical study. Journal of Insurance and Risk Management, 6(2), 65–80.

Gifford, M., &Bayrak, T. (2023). A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression. Decision Analytics

Journal,8,100296,ISSN 2772-6622, https://doi.org/10.1016/j.dajour.2023.100296.

Sze, N.N., Wong, S.C. &Lee C.Y. (2014). The likelihood of achieving quantified road safety targets: A binary logistic regression model for possible factors. Accident Analysis & Prevention, 73, 242-251, ISSN 0001-4575, https://doi.org/10.1016/j.aap.2014.09.012.

Moomen M.,Rezapour M., &Ksaibati K. (2019). An investigation of influential factors of downgrade truck crashes: A logistic regression approach. Journal of Traffic and Transportation Engineering (English Edition), 6(2), 185–195, ISSN 2095-7564, https://doi.org/10.1016/j.jtte.2018.03.005.

Robles-Velasco, A.,Cortés P., Muñuzuri J., &Onieva L. (2020). Prediction of pipe failures in water supply networks using logistic regression and support vector classification. Reliability Engineering & System Safety,196,106754. ISSN 0951-8320, https://doi.org/10.1016/j.ress.2019.106754.