Aspects of Road Safety: A Case of Education by Research – Analysis of Parameters Affecting Accidents

Magda Marek

Social Academy of Sciences in Kielce, 15 Peryferyjna Str., 25-562 Kielce, Poland

Abstract

Today, more attention is paid to road safety. Since an important issue in this safety is education and prevention, it is essential to determine the influence of potential factors on the occurrence of a traffic accident. This article focuses on analyzing the impact of various characteristics on safety on one of the most dangerous roads in Poland, the national road No. 7. A Bayesian model was created for this purpose. The aim of the article is also to present the possibility of using the SAS system for bayesian modelling. Discovering the knowledge from the data allows specific conclusions to be drawn that contribute to making people safer on the roads.

Keywords

Data analysis, Bayesian statistics, Road safety, Monte-Carlo method, Markov chain

1 Introduction

The development of information and communication technologies allows access to many services offered on the Internet [1, 2, 3]. In the era of information society development, systems and technologies that support the learning process and integrate academic networks play a key role. From the point of view of education, computer modelling (taking into account real data) has an important function in scientific research, which can be exemplified in management systems. The goal is to increase the widely understood safety in the public domain. One of the means of achieving this goal is the provision of communication in the situation of adverse events [4, 5, 6], which is possible through the selection of appropriate transmission techniques [7, 8, 9]. Nevertheless, the problems of education and safety can be considered on many different levels. Some activities, often of a local nature, are directed at the aspect of safety (there are various studies on factors determining safety).

One example is road traffic safety [10, 11, 12]. In this respect, education through the analysis of large data sets and the models developed may ultimately contribute to minimizing the causes of road accidents and thus increasing the level of safety of road users, which is the main objective of this article. In this case, the use of real road accident data and modern data analysis methods becomes essential. The safety aspect is influenced by transport psychology [13], traffic engineering [14], the system of periodic technical inspections of vehicles [15], and the formation of road safety programmes [16]. Driver skills, vehicle technical design, and road environment are also important. The condition of the road surface may be affected by weather conditions, therefore, at the stage of road infrastructure design, it is important to know the rainfall statistics in a given region or country [17]. In practice, rains pose the risk of slipping on slippery surfaces.

Education and Research in the Information Society, September 27–28, 2021, Plovdiv, Bulgaria
EMAIL: mmarek0987@gmail.com

2 Materials and Methods

Logistic regression is one of the regression analysis methods that is used when the model contains an explanatory variable with dichotomous values, thus taking only two values. For the purposes of Bayesian modelling, this study uses the SAS (Statistical Analysis System), which allows for logistic regression analysis. This makes it possible to assess the impact and magnitude of the interaction of particular (extracted) characteristics on the occurrence of an accident.

Generally, traffic incidents are characterized by traffic events such as collisions, minor accidents, severe accidents, and fatalities. However, from the point of view of this study, only accidents were selected for analysis. In this case, the objective variable takes the values: *WL* – light accident (treated as failure) or *WCS* – severe or fatal accident (treated as success from the point of research). *WL* is defined as an accident in which victims are at most slightly injured, and *WCS* as an accident in which at least one person is seriously injured or at least one person dies as a consequence of their injuries. This makes it possible to determine the likelihood of severe injury or death in a traffic accident, depending on the characteristics that describe the perpetrator of the accident. Incorrect and repetitive data were removed from the collection. The preparation of the data for analysis is presented in Figure 1 [21].

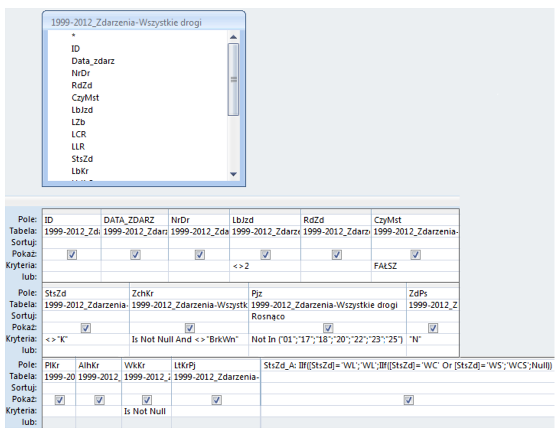

Figure 1: Query to prepare data for further analysis

For this study, a Bayesian regression model with data from the Polish national road No. 7 (DK7) in the Świętokrzyskie Voivodeship was used. It is worth noting that the Bayesian approach is based on simple probability rules [18]. The advantage of classical methods is that there is no need to repeat the experiments many times. In comparison, in the traditional approach, estimators obtained from a small sample size may be less reliable [19, 20, 21]. The choice of Bayes' model results from the author's research interests and the fact that Bayes' theorem can be successfully applied to supercomputers to model various changes [12, 19, 22]. Research experiments using regression most often concern analyses in the field of economics because analysts are more likely to undertake regression modelling in which, in addition to the objective variable and the explanatory variable, there are a number of other predictors. More information is given in [21], where the methodology applied is further described.

The data used for Bayesian regression modelling come from the SEWiK system (Accident and Collision Recording System). They were made available earlier by the Regional Police Headquarters in Poland [21]. It is worth mentioning that national road No. 7 is among the most dangerous roads in Poland. To prepare the final sets for Bayesian regression analyses, the selected data were imported into the database system. Before the final data set was obtained for analysis, it was subjected to several stages of processing in the database system [21].

As previously written, SAS software for statistical analysis and data processing was used [23]. The 4GL (4th Generation Language) was applied for this purpose. A logistic regression model was created to estimate the probability of severe injury or death in a traffic accident, depending on the following characteristics of the perpetrator of the traffic accident: (1) presence of alcohol or narcotics at the time of the accident, (2) gender, (3) age group (driver experience), and (4) driver behaviour causing the traffic accident. On this basis, logistic regression methods can find significant factors for traffic accidents.

3 Model applied

Regression as a statistical tool allows the analytical representation of relationships that link certain characteristics [24]. The study is based on a random sample, which is a number of observation pairs. The regression model represents the formation of the value of a random variable under the influence of an independent variable [25]. In practice, the linear regression model is a certain variety of econometric models made to quantitatively describe the relationship of the variables under study [21, 26]. The complexity of global phenomena generally results in the observation of more than one exogenous variable in the model. The parameters of the linear regression model are estimated under the influence of procedures established in the field of correlation and regression theory [21]. For this research, the logit function, which is applied to a family of binomial distributions, is relevant. In practice, the logit model is a case of a generalized linear model. It is worth recalling that in logistic regression, unlike classical regression, the classical probability is not determined, but the chance is expressed by the ratio of the number of successes to the number of failures. This means that the interpretation of the individual parameters of Bayesian regression is done using the odds ratio. In logistic regression, the linking function (logit) takes the form:

$$\operatorname{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{k} \beta_i X_i \tag{1}$$

where:

- p – probability of success,
- $p\text{-}1$ – probability of failure,
- β_0 – intercept parameter,
- $\beta_1, \beta_2, \ldots, \beta_k$ – regression coefficients.

From a practical point of view, multiple regression can take the form of a linear or nonlinear model. The concepts involved in multiple regression analysis overlap with simple regression analysis. These include measures of model fit to empirical data, confidence intervals, graphical representation of regression, and hypothesis testing [27]. However, compared to linear modelling, multiple regression modelling is more accurate in forecasting [21, 28]. As noted in [21], differences arise when interpreting the coefficients of the equation that describe the model.

GENMOD procedure in SAS environment was used to build and analyze Bayesian model for national road No. 7 in Świętokrzyskie Voivodeship in Poland. The results of the procedure for this spatial aspect are included in Figure 2 [21].

Number of Observations Read	62
Number of Observations Used	62
Number of Events	148
Number of Trials	257

Figure 2: Result table with numerical values of the procedure observations GENMOD

Figure 2 presents the number of records in the data set (its number is equal to 62). The number of events represents the number of successes (accidents with the *WCS* status recorded in the entire data set). On the other hand, the number of trials represents the number of recorded events (accidents) in the entire data set.

Class-level information provides data on the qualitative explanatory variables of the model. As previously mentioned, these are the presence of alcohol (*AlhKr*), the gender of the driver (*PlKr*), the age group of the driver (*GrWkKr*), and the behaviour of the driver after aggregation (*ZchKr*).

$$Y = StZd\_A = f(AlhKr, PlKr, GrWkKr, ZchKr) \tag{2}$$

where:

- *AlhKr* – presence of alcohol/narcotics at the time of accident,
- *PlKr* – gender of the driver,
- *GrWkKr* – age group of the driver (number of years at a time of event; see Table 1),
- *ZchKr* – behaviour of the driver (see Table 2),
- *StZd_A* – status of the traffic incident (explanatory variable).

For analysis, some of the categories have been further aggregated. The following are highlights of the aggregation of the variables and the structure of the model. Parameter values will also be presented in this paper.

The variable *GrWkKr* is ordinal (its values can be ordered, however, they do not represent a scale of difference). Nevertheless, they may be used as information on traffic risks in a descriptive way (they do not appear in a mathematical expression) [21]. Coding made it possible to replace qualitative values with numerical values.

Table 1: Codification of the age of the driver (*GrWkKr*)

Code	Age group
01	<18
02	<25
03	<35
04	<50
05	<65
06	>65

In addition to the values previously described after aggregation, aggregated categories that describe driver behaviour were used for the analyses. These characteristics are qualitative variables.

Table 2: Codification of behaviour of the driver (*ZchKr*) before and after data aggregation*

Code	Description	Name
01	Inappropriate speed to road conditions	*NdsPr*
02	Failure to give priority	*NdzPrwPrz*[1]
03	Incorrect passing	*NprWprz*[4]
04	Incorrect overtaking	*NprOmj*[4]
05	Incorrect evasive action	*NprWmj*[2]
06	Incorrect driving through a crosswalk	*NprPPsz*
07	Incorrect turning	*NprSkr*[3]
08	Incorrect stopping/parking	*NprZtPsPj*[2]
09	Incorrect reversing	*NprCfn*[2]
10	Driving on the wrong side of the road	*JzdNwSDr*
11	Driving through a red light	*WjdCzSw*[1]
12	Failure to observe other signals	*NprzInSg*
13	Failure to keep a safe distance between vehicles	*NzOd*[6]
14	Rapid braking	*GwHmw*[6]
15	Driving without the required lighting	*JzWmOs*
16	Tiredness, falling asleep	*ZmZsn*[5]
17	Reduced psychomotor ability	*OgSpPsch*[5]

* In addition, incorrect driving in place for other users, including bicycles (*NprPrzPRw*[2]), incorrect lane changing (*NprZmPsRch*), and incorrect U-turning (*NprZwr*[3]) were taken into account. Thus, the resulting data set is a complete set of possible events on the road.
Selected parameters after data aggregation:
[1] *NdzPrwPrz* + *WjdCzSw* = *NdzPrw* (failure to give priority)
[2] *NprCfn* + *NprWmj* + *NprZtPsPj* + *NprPrzPRw* = *NprMnIn* (incorrect maneuvers on the road: reversing, evasive action, stopping/parking, or driving in place for other users)
[3] *NprSkr* + *NprZwr* = *NprSkZw* (incorrect turning or U-turning)
[4] *NprWprz* + *NprOmj* = *NprWO* (incorrect passing or overtaking)
[5] *ZmZsn* + *OgSpPsch* = *NspKr* (tiredness, falling asleep, reduced psychomotor ability)
[6] *NzOd* + *GwHmw* = *NzOdl* (failure to keep a safe distance, rapid braking as the cause of an accident)

Figure 3 presents the response profile, which characterizes the target variable by assigning a value of 1 as success (Event) and 2 as failure (Nonevent) [21]. This figure shows that 148 successes (*WCS*) and 109 failures (*WL*) were observed among the generated data set for analysis.

Response Profile		
Ordered Value	Binary Outcome	Total Frequency
1	Event	148
2	Nonevent	109

Class Level Information		
Class	Levels	Values
ZchKr_A	8	JzdNwSDr NdsPr NdzPrw NprMnIn NprSkZw NprWO NspKr NzOdl
GrWkKr_a	6	01 02 03 04 05 06
AlhKr	2	N T
PlKr	2	K M

Figure 3: Result table for response profile and class-level information of the GENMOD procedure

The list presented in Figure 4 includes a maximum likelihood parameter estimate [21]. It contains a table of estimates of logistic regression parameters for the classical variant (assuming that the parameters are constant). The standard error column reports the value of the standard error (standard deviation) of the estimate. In reality, the true exact value of this error is unknown, so it is assumed to be some estimate of the deviation, indicating the difference between the measured value and the true value. Figure 4 allows us to compare the point estimators with the mean values. *DF* (Degree of Freedom) indicates the number of degrees of freedom, which is the number of independent parameters to estimate a particular regression parameter.

It is important to know the accuracy of the model so that a range of individual values can be predicted. It becomes possible to determine, e.g., the likelihood ratio 95% confidence limits. Based on this, the interval in which a parameter is found with a certain probability may be specified. It is worth emphasizing that it is essential that the expected values fall within the assumed confidence intervals. By default, in the GENMOD procedure, the significance level α is equal to 0.05. Therefore, the probability of the confidence interval is 95%. For example, for the driver behaviour parameter *JzdNwSDr* this value is equal to 2.8639, *NdsPr* – 0.9271, *NdzPrw* – 1.0757, *NprMnIn* – 1.4515, *NprSkZw* – 0.5601, *NprWO* – 2.0216, *NspKr* – 0.3724, and for *NzOdl* – 0. The level of spread of values depends on the count of given events, therefore, some values differ from others. Based on this, it is possible to evaluate the impact of particular model parameters on accident risk based on the set of processed data.

Analyzing the data obtained, it can be observed that the probability of success for the youngest age group of drivers is much higher than for the other groups. Generally, the direct causes of accidents are most often: driving on the wrong side of the road, incorrect manoeuvres on the road: incorrect passing or overtaking, reversing, evasive action, stopping/parking, driving in place for other users, and failure to give priority. It is important to assess the quality of a posterior distribution. Due to the limited volume of this article, only the most important information about the diagnostic tests performed is included (Section IV). More information can be found in [21].

Analysis Of Maximum Likelihood Parameter Estimates				
Parameter		DF	Estimate	Standard Error
Intercept		1	1.2585	0.9371
ZchKr_A	JzdNwSDr	1	2.8639	1.1526
ZchKr_A	NdsPr	1	0.9271	0.5056
ZchKr_A	NdzPrw	1	1.0757	0.5268
ZchKr_A	NprMnln	1	1.4515	0.9936
ZchKr_A	NprSkZw	1	0.5601	0.6173
ZchKr_A	NprWO	1	2.0216	0.6896
ZchKr_A	NspKr	1	0.3724	0.7486
ZchKr_A	NzOdl	0	0.0000	0.0000
GrWkKr_a	01	1	21.7273	80020.04
GrWkKr_a	02	1	-0.7296	0.7084
GrWkKr_a	03	1	-0.7390	0.6879
GrWkKr_a	04	1	-1.1320	0.6810
GrWkKr_a	05	1	-0.6250	0.6960
GrWkKr_a	06	0	0.0000	0.0000
AlhKr	N	1	-1.1929	0.6062
AlhKr	T	0	0.0000	0.0000
PlKr	K	1	-0.3069	0.4423
PlKr	M	0	0.0000	0.0000

Figure 4: Analysis of maximum likelihood parameter estimates of the GENMOD procedure

4 Diagnostic graphs

The results of the GENMOD procedure also include diagnostic graphs. The number of graphs obtained corresponds to the number of parameters of the logistic regression equation. These are presented in Figures 5-11 [21].

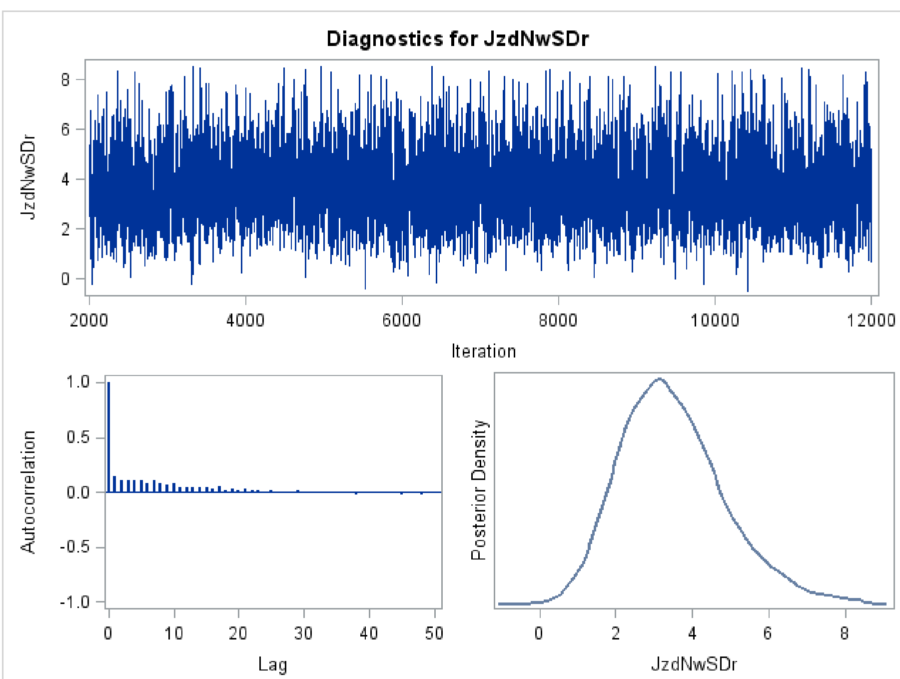

Figure 5: Results for the *JzdNwSDr* (driving on the wrong side of the road) parameter

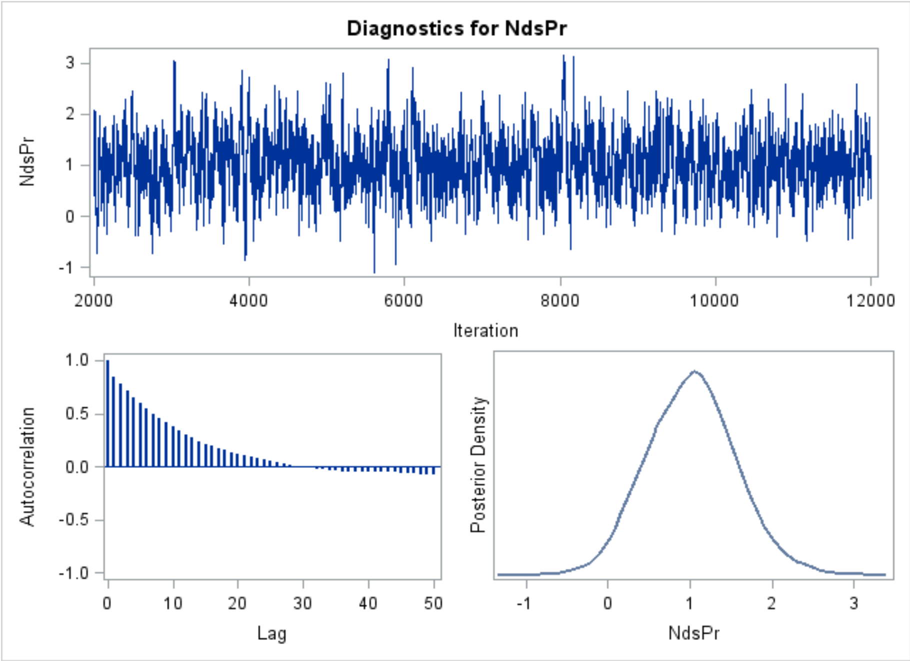

Figure 6: Results for the NdsPr (inappropriate speed to road conditions) parameter

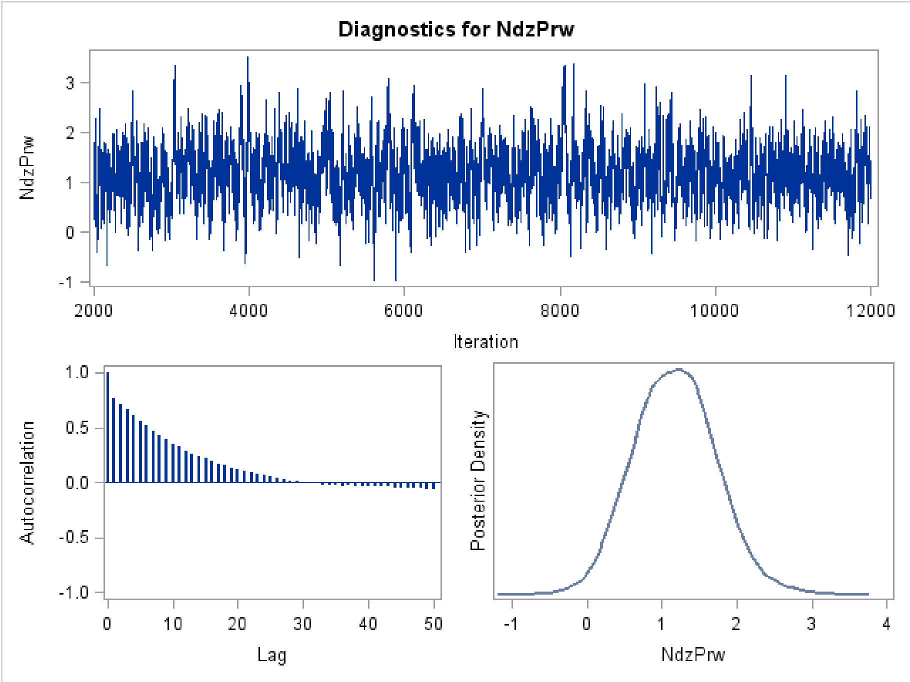

Figure 7: Results for the NdzPrw (failure to give priority) parameter

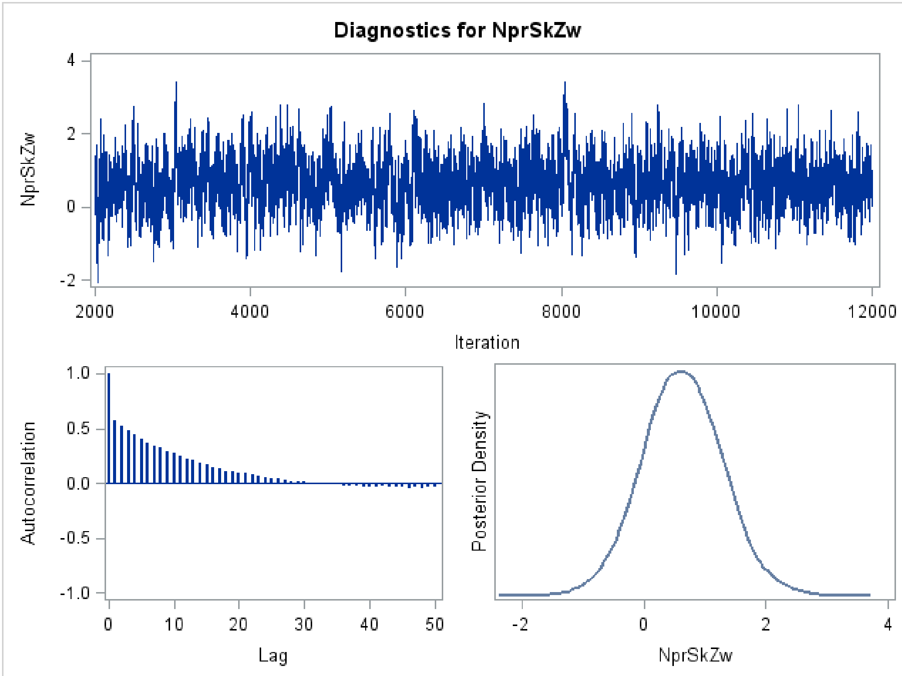

Figure 8: Results for NprSkZw (incorrect turning or U-turning) parameter

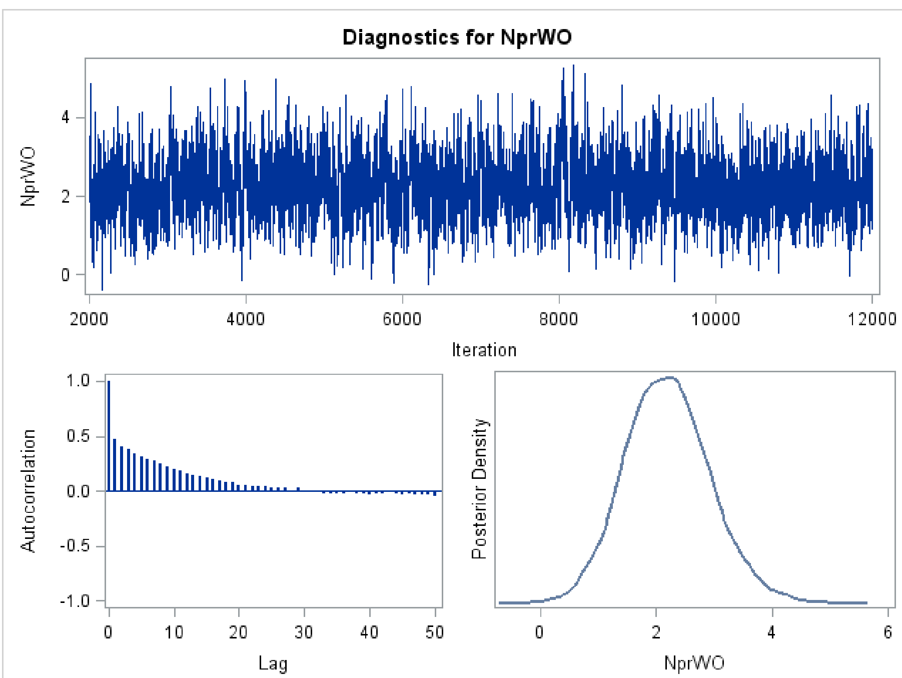

Figure 9: Results for the NprWO (incorrect passing or overtaking) parameter

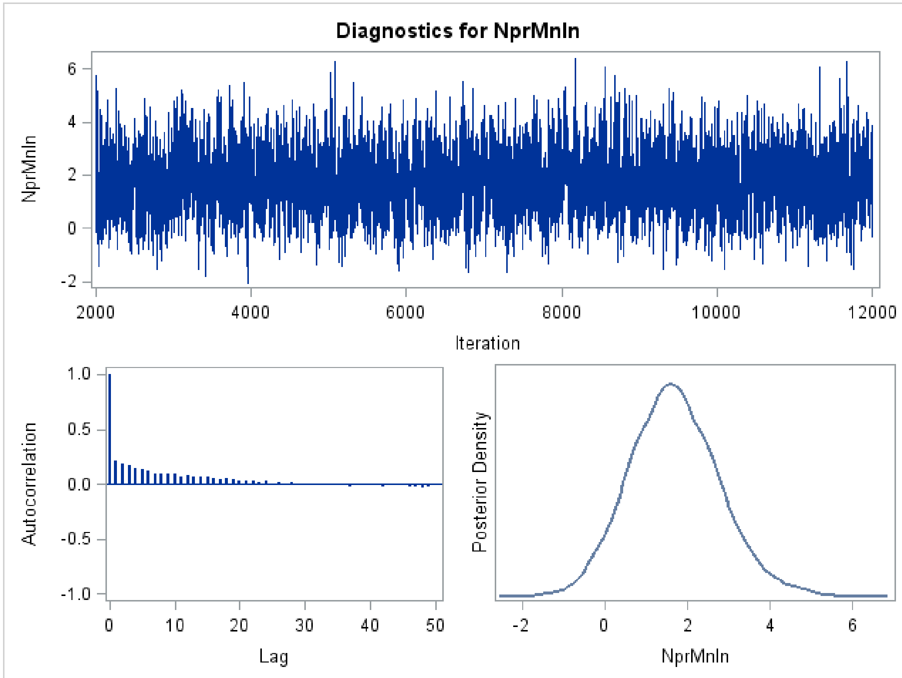

Figure 10: Results for the NprMnIn (incorrect manoeuvres on the road: reversing, evasive action, stopping/parking, driving in place for other users) parameter

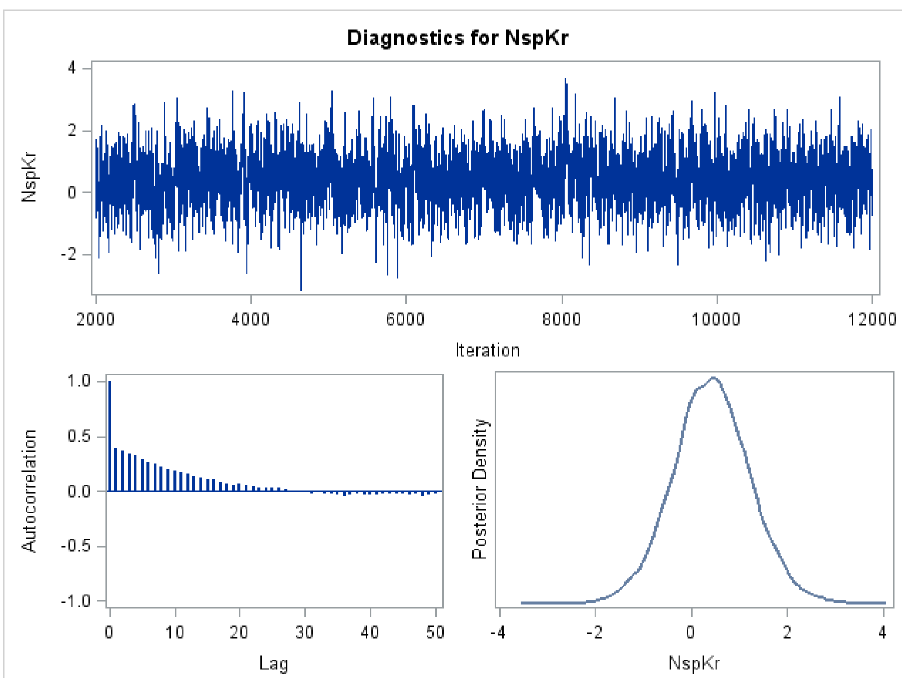

Figure 11: Results for the NspKr (tiredness, falling asleep, reduced psychomotor ability) parameter

Diagnostic figures (Figures 5-11) for each parameter consist of several parts. The first one contains a graph showing the mixing quality of the Markov chain, which thus denotes the estimation quality of the regression parameter distributions. For example, the best mixing is presented by the graph in Fig. 5 for the regression parameter *JzdNwSDr* (driving on the wrong side of the road). The chain is convergent and stabilized in this case [21]. Another part of the diagnostic graph is the bar graph of autocorrelation of individual parameters. In practice, when the autocorrelation plot decreases relatively slowly, it is an undesirable phenomenon. A much faster decrease in the autocorrelation of the parameters was observed, for example, for the *JzdNwSDr* parameter. Moreover, it is also possible to read the expected values of the individual parameters. Information in this regard is included in Table 3.

Table 3: Excepted values of regression parameters

Parameter	Value
Intercept	1,3262
JzdNwSDr	3,4961
NdsPr	1,0193
NdzPrw	1,1741
NprMnIn	1,6920
NprSkZw	0,6372
NprWO	2,2006
NspKr	0,4209

The study considered logistic regression, assuming that the parameters are random variables, that is, they can take multiple values. However, the occurrence of the values given in Table 3 is the most likely. The use of the Bayesian approach results in good identifiability of parameters in short samples (estimators obtained from a small sample may be less reliable in the classical approach) [21].

5 Conclusions

The paper as an example of education by research presents a Bayesian model built to assess the impact of various features on road safety. The model was developed by applying a logistic regression function realized by SAS software to find significant factors in traffic accidents. The cause of most accidents is improper driving behaviour (driving on the wrong side of the road and incorrect passing or overtaking, as well as failure to give priority). In addition, the failure to adapt the speed to road conditions is a common cause of road accidents. These features have the greatest influence on the regression model in the aspect of the data set, which comes from the analyzed national road No. 7 in Świętokrzyskie Voivodeship in Poland (real data had been used on one of the most dangerous roads in Poland). Besides, taking into account all the results obtained, the perpetrators of accidents on road No. 7 are most often young people. The probability of success for the youngest age group of drivers is much higher than for the other groups. This raises the question of whether a driving license should be available at such a young age when severe or fatal road accidents involving young drivers occur?

Through research, it is possible to educate not only drivers but society as a whole by presenting the factors that primarily affect road safety. This is important because it can contribute to the discussion of educational needs and requirements for future and current drivers. The development of information and communication technology is helpful in this regard. From an educational point of view, the use of modern techniques to analyze large and reliable data sets is a key issue, as presented in this article.

References

[1] J. Wilk, Home Networking. Sieci domowe. Specyfika sieci oraz rozwiązania techniczne, Wydawnictwo Stowarzyszenia Współpracy Polska-Wschód, Oddział Świętokrzyski, Kielce, 2012.

[2] J. Wilk-Jakubowski, Usługi oferowane we współczesnych sieciach komputerowych, AUTOBUSY – Technika, Eksploatacja, Systemy Transportowe 18 (2017) 1796–1798.

[3] J. Wilk-Jakubowski, Analysis of broadband informatics services provided via the Internet, and the number of internet users on a global scale, GSI J. 1 (2020), Online 1–12.

[4] M. Sasanuma, H. Uchiyama, T. Nagoya, M. Furukawa, T. Motohisa, Research and development of very small aperture terminals (VSAT) that can be installed by easy operation during disasters – Issues and the solutions for implementing simple and easy installation of VSAT earth station, IEICE Tech. Rep., Vol. 112 (2013), pp. 1–3.

[5] S. Kameda, T. Okuguchi, S. Eguchi, N. Suematsu, Development of satellite-terrestrial multi-mode VSAT using software defined radio technology, in: Asia-Pacific Microwave Conference, IEEE, Sendai, 2014.

[6] N. Suematsu, H. Oguma, S. Eguchi, S. Kameda, M. Sasanuma, K. Kuroda, Multi-mode SDR VSAT against big disasters, in: European Microwave Conference, EuMC, Nuremberg, 2013. doi: 10.23919/EuMC.2013.6686788.

[7] J. Wilk-Jakubowski, Overview of broadband information systems architecture for crisis management, Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska 10 (2020) 20–23. doi: 10.35784/iapgos.1608.

[8] J. Ł. Wilk-Jakubowski, Zapewnienie łączności globalnej podczas zarządzania kryzysowego w świetle Europejskiego Projektu Badawczego ICT Cost Action IC0802, in: A. Malarczyk, G. Wilk-Jakubowski, R. S. Brzoza (Eds.), Polska – Europa – Świat Wobec Przemian Oraz Zjawisk Kryzysowych. Szanse I Zagrożenia, Wydawnictwo Stowarzyszenia Współpracy Polska-Wschód, Oddział Świętokrzyski, Kielce, 2017.

[9] J. Wilk-Jakubowski, A review on information systems engineering using VSAT networks and their development directions, Yugoslav Journal of Operations Research, Online-First 00 (2021) 15–15. doi: 10.2298/YJOR200215015W.

[10] A. Wojtas, M. Szkoda, Analiza wybranych czynników wpływających na bezpieczeństwo w ruchu drogowym, Autobusy 6 (2018) 1149–1154. doi: 10.24136/atest.2018.244.

[11] D. Stavrinos, J. L. Jones, A. A. Garner, R. Griffin, C. A. Franklin, D. Ball, S. C. Welburn, K. K. Ball, V. P. Sisiopiku, P. R. Fine, Impact of distracted driving on safety and traffic flow, Accident Analysis and Prevention 61 (2013) 63–70. doi: 10.1016/j.aap.2013.02.003.

[12] M. Marek, Bayesian regression model estimation: A road safety aspect. To appear, submitted to the SCA 21 (Sixth Smart City Applications International Conference), Safranbolu, 27–29 October, 2021.

[13] D. Bąk-Gajda, J. Bąk, Psychologia transportu i bezpieczeństwo ruchu drogowego, Difin, Warszawa, 2010.

[14] S. Gaca, W. Suchorzewski, M. Tracz, Inżynieria ruchu drogowego. Teoria i praktyka, Wydawnictwa Komunikacji i Łączności, Warszawa, 2008.

[15] W. Jarosiński, System okresowych badań technicznych pojazdów a bezpieczeństwo ruchu drogowego i liczba zdarzeń drogowych z udziałem ofiar śmiertelnych, Eksploatacja i Niezawodność 16 (2014) 105–111.

[16] K. Jamroz, A. Romanowska, Wpływ Wizji Zero na kształtowanie programów bezpieczeństwa ruchu drogowego w Polsce, Pomorskie Forum Drogowe, Sopot, 4–6 June, 2018.

[17] J. Wilk-Jakubowski, Measuring Rain Rates Exceeding the Polish Average by 0.01%, Pol. J. Environ. Stud. 27 (2018) 383–390. doi: 10.15244/pjoes/73907.

[18] F. Koop, Wprowadzenie do ekonometrii, Wolters Kluwer Polska Sp. z.o.o., Warszawa, 2011.

[19] H. T. Jelenkowska, Zastosowanie metody bayesowskiej do analizy statystycznej wielocechowych doświadczeń rolniczych, in: M. Fotyma (Ed.), Fragmenta agronomica, Polskie Towarzystwo Nauk Agrotechnicznych, Puławy, 2001.

[20] M. Nowakowska, Modelowanie związków między cechami drogi a zagrożeniami w ruchu drogowym, Prace Naukowe Politechniki Warszawskiej, TRANSPORT 88 (2013) 3–266.

[21] M. Marek, Aspekt przestrzenny w zadaniu estymacji bayesowskiego modelu regresji, Master's thesis, Kielce University of Technology, Kielce, Poland, 2014.

[22] H. Thornburg, Introduction to bayesian statistics, 2001. URL: http://ccrma.stanford.edu/jos/bayes/bayes.html.

[23] SAS Help and Documentation, 2014. URL: https://support.sas.com/en/documentation.html.

[24] E. Frątczak, A. Korczyński, Statystyka od podstaw z systemem SAS, Oficyna Wydawnicza Szkoła Główna Handlowa w Warszawie, Warszawa, 2013.

[25] J. Jóźwiak, J. Podgórski, Statystyka od podstaw, Polskie Wydawnictwo Ekonomiczne, Warszawa, 2009.

[26] M. Marek, Wykorzystanie ekonometrycznego modelu klasycznej funkcji regresji liniowej do przeprowadzania analiz ilościowych w naukach ekonomicznych, in: Z. E. Zieliński (Ed.), Rola informatyki w naukach ekonomicznych i społecznych. Innowacje i implikacje interdyscyplinarne, Wydawnictwo Wyższej Szkoły Handlowej w Kielcach, Kielce, 2013.

[27] J. Koronacki, J. Mielniczuk, Statystyka dla studentów kierunków technicznych i przyrodniczych, Wydawnictwa Naukowo-Techniczne, Warszawa, 2006.

[28] D. Larose, Metody i modele eksploracji danych, Wydawnictwo Naukowe PWN, Warszawa, 2008.