Choice of Functional Form for Independent Variables in Accident Prediction Models

Joanna Kamińska
Wrocław University of Environmental and Life Sciences Department of Mathematics, Poland

The development of multivariate statistical models to identify factors that explain systematic variation in accident counts has been an active field of research in the past 20 years. During this period many different models and functional forms have been applied. This study, based on data for national roads in Norway, tests alternative functional forms of the relationship between independent variables and the number of injury accidents. The paper compares six different functional forms (sets of independent variables and specifications of the form of their relationship to accident occurrence) by means of Poisson-lognormal regression. The best model was identified in terms of five goodness of fit measures and a graphical method – the CURE plot (CURE = cumulative residuals). The coefficients estimated for the independent variables were found to vary according to functional form. It is therefore important to compare different functional forms as part of an exploratory analysis when developing accident prediction models.
Keywords: accident prediction models, Poisson lognormal model, cumulative residuals, functional form.

1. INTRODUCTION

The relationship between road accidents and factors influencing their number has been a focus of research for a long time (Satterthwaite, 1981; Hauer, 1986, 2004), but modern analyses, based on the generalised Poisson distribution have emerged in the last twenty years. Researchers have developed many different models and functional forms depending on data availability, local conditions and study purpose. Models based on the Poisson distribution or its generalisations are most widely used. Some common models include: Poisson lognormal – PLN (Miaou et al. 2005, Lord and Miranda-Moreno 2008, El-Basyouny and Sayed 2009B), multivariate Poisson lognormal - MPLN (Ma and Kockelman 2006, Tsionas 2001, Park and Lord 2007, El-Basyouny and Sayed,2009A, Ma et al.,2008), negative binomial – NB (Miaou 1996, Vogt 1999, Lord and Park 2008), zero-inflated Poisson – ZIP (Li et al.,1999), Conway-Maxwell-Poisson- COM-Poisson (Lord et al. 2009) and generalized estimating equations – GEE (Wang and Abdel-Aty 2007, Lord and Persaud 2000). All of these models contain functional forms which describe the relationship between independent variables and the dependent variable (number of accidents).

Hauer (2004) and Hauer and Bamfo (1997) stress that in most applications of standard statistical software, only monotonic functions are used to model the relationship between variables. They argue that the choice of functional form to describe the relationship between variables should be based on an exploratory analysis in which alternative functions are tested and compared. The main research problems addressed in this paper are:

1. What functional forms have been used in prior accident models?
2. Do different functional forms produce different estimates of the relationship between variables?
3. How can the best functional form be chosen?

Following a brief review of the literature, a Norwegian data set will be analysed by means of different functional forms and the results compared.

2. LITERATURE REVIEW

During more than 30 years of accident modelling, many functional forms have been applied. Let $P(n_i) = f(\mu_i, \varepsilon_i)$ denote the probability mass function with random error ε_i and parameter μ (more than one parameter is possible). Then a function can be written as $\mu_i = g(\beta, x_j)$ where X_j are independent variables characterising for example pavement, road geometry and traffic volume and β is a vector of parameters. Some of the functional forms used in prior studies are listed in Table 1. In every model AADT (annual average daily traffic), which represents traffic volume has been found to be the most important factor influencing the number of accidents.

All functional forms included in Table 1 can be represented as exponential functions of a linear combination of parameters and variables (in order to meet Poisson regression model assumption (Winkelmann, 2008)). The variables included in this linear combination can be entered either in natural units or can be non-linear transformations of the original variables.

The question is which functional form should be used in particular situation. Both type of model and functional form depend on characteristics of the data and accident severity (Lord at al., 2005).

Table 1. Overview of some functional forms used in previous studies

Functional forms	Model	Examples of studies
$\mu_i = \exp\left(\beta_0 + \sum_j X_{ij}\beta_j\right)$	MVPLN	Park and Lord 2007
	MVPLN	El-Basoyouny and Sayed 2009A
	PLN	El-Basoyouny and Sayed 2009B
	NB	Karlaftis and Tarko 1997
	NB	Anastasopoulos and Mannering 2009
$\mu_i = \beta_0\left(\frac{365 \cdot X_{1i} \cdot X_{2i} \cdot X_{3i}}{1000000}\right)\exp(\beta_1 X_{4i} + \beta_2 X_{5i})$	NB	Li, Lord, Zhang and Xie 2008
$\mu_i = \beta_0(X_{1i} + X_{2i})^{\beta_1}$	NB	Lord and Park 2008
$\mu_i = \beta_0 \cdot X_{1i}^{\beta_1} \cdot X_{2i}^{\beta_2}$	NB	Lord and Park 2008
$\mu_i = \beta_0(X_{1i} + X_{2i})^{\beta_1}\left(\frac{X_{2i}}{X_{1i}}\right)^{\beta_2}$	NB	Lord and Park 2008
$\mu_i = \exp(\beta_0 + \beta_3 X_{2i}) \cdot X_{1i}^{\beta_1} \cdot X_{2i}^{\beta_2}$	NB	Lord and Park 2008
$\mu_i = \beta_0 \cdot (X_{1i} X_{2i})^{\beta_1}$	NB COM-Poisson	Lord and Park 2008 Lord, Guikema and Geedipally 2008
$\mu_i = \beta_0 \cdot X_{1i} \cdot X_{2i}^{\beta_1}$	COM-Poisson	Lord, Guikema and Geedipally 2008
$\mu_i = \beta_0 \cdot X_{1i}^{\beta_1} \exp\left(\sum_j X_{ij}\beta_j\right)$	NB	Elvik 2008 Wang and Abdel-Aty 2007
$\mu_i = \beta_0 \cdot AADT_i^{\beta_1} \exp\left(\frac{AADT}{1000}\beta_2 + \sum_j X_{ij}\beta_j\right)$	NB	Elvik 2008
$\mu_i = \beta_0 \cdot X_{1i}^{\beta_1} \cdot X_{2i}^{\beta_2} \exp\left(\sum_j X_{ij}\beta_j\right)$	NB	Abdel-Aty and Radwan 2000
$\mu_{it} = \beta_0 \cdot Leng_{it}^{\beta_2} F_{it}^{\beta_3 \ln F_{it}} Leng_{it}^{\beta_4 \ln Leng_{it}}$.	NB GEE	Lord and Persaud 2000
$\exp\left(\begin{array}{l}\beta_5 IntDen_i + \ln\gamma TimeT_t + \beta_6(IntDen_i)^2 + \beta_7 \ln(F_{it})\ln(Leng_i) + \\ \beta_8 \ln(F_{it}) IntDen_i + \beta_9 \ln(Leng_{it}) IntDen_i\end{array}\right)$	NB	Couto and Ferreira 2011

Source: personal study.

For example, a PLN model has been applied to crash count data recorded on urban roads in the city of Vancouver (El-Basyouny and Sayed, 2009b). A MPLN model has been applied to crash count data recorded on different types of roads: two-lane highway segments (Ma at al., 2008), three-leg un-signalised intersections in California highways (Park and Lord, 2007), signalised intersections in the city of Edmonton (El-Basyouny and Sayed, 2009a). A NB model has been applied to crash count data recorded on various types of roads such as two high speed roads: highways in Central Florida (Abdel-Aty and Radwan, 2000), rural interstate highways in Indiana (Anastasopoulos and Mannering, 2009), rural frontage roads in Texas (Li at al., 2008) and differentiated roads in 92 counties of Indiana (Karlaftis and Tarko 1997). The GEE model has been applied to crash count data recorded only on intersections: four-legged signalised intersections in Toronto (Lord and Park, 2008) and selected intersections in Florida (Wang and Abdel-Aty, 2007). The GLM (generalized linear model) has been used with Poisson error structure for establishing the relationship between accidents and geometry for many types of junctions (3,4-Arm roundabouts, major-minor, signalized and other) (Maher and Summersgill, 1995). The data considered in this paper represent national roads in all Norwegian counties (highways, rural roads and urban sections).

The PLN model was chosen because the empirical distribution of the count of accidents between road sections fitted best to the Poisson-lognormal distribution (see point 3.2.).

3. DATA AND METHODS

3.1 Functional forms tested

The PLN model was used in order to compare different functional forms in this paper. This model was chosen because it is easy to estimate and more flexible than the negative binomial with respect to over-dispersion (Lord and Mannering, 2010). However, the Poisson lognormal model is adversely affected by small sample size (Miaou et al., 2003). The sample set used in this study is large enough (25,739 road sections) to enable the effective use of PLN model. The probability β_j of road segment *i* having λ_i accidents is:

$$P(n_i) = \frac{\exp(-\lambda_i)\lambda_i}{n_i!} \tag{1}$$

where λ_i is the Poisson parameter representing expected number of accidents $E(n_i)$ for segment *i*. Poisson regression defines the parameter $\lambda_i = E(n_i)$ as a function of explanatory variables,

$$D(M_j) \tag{2}$$

where X_j are covariates representing traffic volume and characteristics of the roads, β_j are parameters, $g(X_{ji}, \beta_j)$ is a functional form, ε_i denote error terms distributed as $N(0,\sigma^2)$ (normal distribution with mean value of zero).

According to Noland and Karlaftis (2005) the selection of the appropriate functional form must be dedicated by the nature of the dependent variable. The functional forms considered below include variables that are intuitively connected with the number of accidents. On the basis of the literature review as well as available data, after analysis of scatter plots for number of crashes and independent variables it was decided that in addition to basic log-linear functional forms, functions with speed limit squared as in Ma et al. (2008), natural logarithm of number of lanes plus 1 and natural logarithm of number of junction plus 1, as in Elvik (2008) would be taken into consideration.

Six different functional forms have been compared (note that $\beta_o' = \exp(\beta_0)$):

$$g_1 = \beta_0' \cdot \exp\left(\beta_1 AADT + \sum_{i=1}^{9} \beta_{i+1} X_i\right) \tag{3}$$

$$g_2 = \beta_0' \cdot AADT^{\beta_1} \cdot \exp\left(\beta_2 AADT + \sum_{i=1}^{9} \beta_{i+2} X_i\right) \tag{4}$$

$$g_3 = \beta_0' \cdot AADT^{\beta_1} \cdot (X_2+1)^{\beta_2} \cdot \exp\left(\beta_3 AADT + \beta_4 X_1 + \sum_{i=3}^{9} \beta_{i+3} X_i\right) \tag{5}$$

$$g_4 = \beta_0' \cdot AADT^{\beta_1} \cdot \exp\left(\beta_2 AADT + \beta_3 {X_8}^2 + \sum_{i=1}^{9} \beta_{i+3} X_i\right) \tag{6}$$

$$g_5 = \beta_0' \cdot AADT^{\beta_1} \cdot (X_1+1)^{\beta_2} \cdot \exp\left(\beta_3 AADT + \sum_{i=1}^{9} \beta_{i+3} X_i\right) \tag{7}$$

$$g_6 = \beta_0' \cdot AADT^{\beta_1} \cdot \exp\left(\beta_2 AADT + \beta_3 {X_4}^2 + \beta_4 {X_8}^2 + \sum_{i=1}^{9} \beta_{i+4} X_i\right) \tag{8}$$

The subscripts from 1 to 9 denote different independent variables, identified in Table 5. Functional form g_1, as the only one, contains annual average daily traffic (AADT) in natural units. In all other models, the natural logarithm of AADT has been used. The parameters of each function have been estimated using the maximum log likelihood method with basic discrete Newton algorithm (Winkelmann, 2008). In this study, the

LIMDEP software package was used to estimate all the coefficients.

3.2 Measures of goodness-of-fit

To compare the crash prediction models, a single measure of goodness of fit is not enough (Lord and Park, 2008). In the present study five measures of goodness-of-fit were employed:

1. Akaike's Information Criterion (AIC)

$$AIC = \frac{-2\ln L(M_j) + 2k}{N} \tag{9}$$

where $lnL(M_j)$ is a log-likelihood value of model j, k is the number of parameters and N is the number of observations (here $N=25739$).

2. Bayesian Information Criteria (BIC)

$$BIC = D(M_j) + k \cdot \ln N = -2\ln L(M_j) + k \cdot \ln N \tag{10}$$

where $D(M_j)$ is a deviance of model M_j.

3. Mean absolute deviation (MAD)

$$MAD = \frac{1}{N}\sum_{i=1}^{N} | \hat{y}_i - y_i | \tag{11}$$

where $\hat{y}_i$ is the estimated number of crashes in segment I and y_i is the observed number of crashes in segment i.

4. Mean squared prediction error (MSPE)

$$MSPE = \frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{12}$$

5. MAPE measure of fit

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{13}$$

Because of the occurrence of cases where observed number of crashes equaled zero, the smoothed out version (14) was applied:

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N} \left| \frac{\hat{y}_i - y_i}{\bar{y}} \right| \tag{14}$$

where $\bar{y}$ is the mean of the observed number of crashes per road section.

The model performance is better if the test values are smaller, except for the R-squared measure, where a high value indicates a good fit.

Another measure of goodness of fit, indicating whether a certain functional forms first the data closely was proposed by Hauer and Bamfo (1997) and is known as the CURE method (CURE = cumulative residuals). This graphical method has been used extensively in safety analysis (e.g. Lord and Park, 2008; Lord and Persaud, 2000; Wang and Abdel-Aty, 2007, Couto and Ferreira, 2011) and in many other scientific areas (e.g. Lin at all, 2002) to compare different functional forms. The cumulative residuals are plotted for each explanatory variable. The residuals (e_i) represent the difference between the observed (y_i) and estimated $(\hat{y}_i)$ number of crashes. The closer the residuals oscillate around zero line, the better the model fits to the data. For variables with a limited range of values (less than 30) cumulative residuals box and whiskers plot (CRBW) have been applied. This method permits an evaluation of how well the model fits the data set especially for the one chosen explanatory variable. The standard CURE is very useful for continuous variables like AADT.

In order to choose between the most often used models for accident prediction analysis (PLN or NB – negative binomial) the matching between each models to entry data (numbers of accidents) was checked. The original Poisson's model was also considered for the sake of comparison. The resulting parameters are shown in Table 2.

Table 2. Parameter estimates with standard errors (in parentheses)

Distribution	Parameter	
Poisson	lambda	1,9982 (0,005)
Negative – binomial	lambda	1,9982 (0,015)
	alpha	2,6475 (0,049)
Poisson lognormal	lambda	0,5039 (0,013)
	sigma	1,3332 (0,005)

Source: personal study.

The goodness of fit for each distribution was measured by three measures: chi-squared statistics value, log-likelihood value and AIC. The results were listed in table 3.

Table 3. Goodness of fit measures for different models

Goodness of fit measure	Poisson	NB	PLN
Log-likelihood	-60473	-38576	-37908
AIC	4,6989	2,9976	2,9457
χ^2	233303	43793	45128

Source: personal study.

In order to enable the visual evaluation of the matching of each of the distribution types, the estimated number of crashes for each type of the accidents within each of the three types of distribution was listed below:

Table 4. Distribution of road sections by total number of crashes

Number of accidents	Accidents frequency			
	data	Poisson	NB	PLN
0	14109	3490	12851	14045
1	5689	6973	4082	5666
2	2397	6967	2365	2490
3	1217	4640	1576	1253
4	714	2318	1120	707
5	386	926	824	443
6	287	309	621	296
7	199	88	476	200
8	129	22	369	132
9	92	5	289	87
10	71	1	228	62
11	60	0	181	50
12	62	0	144	45
13	46	0	116	42
14	29	0	93	39
15	26	0	75	34
16	29	0	60	28
17	20	0	49	22
18	16	0	40	16
19	15	0	32	12
20	13	0	26	8
>20	39	0	119	62
χ^2		19754,9	15924,3	477,7

Source: personal study.

3.3 Data

The data refer to 25,739 segments on national roads in Norway. For these sections, data on accidents and a number of variables associated with the number of accidents were obtained for the period 1993-2000. The dependent variable in the analysis was the total number of injury accidents. The data contain ten explanatory variables describing geometric characteristics, traffic flow and additional information. There were 33,691 accidents in total, 1437 (4.26 %) of which were fatal. There were 14,109 (55 %) sections with a zero count of accidents. The highest number of accidents recorded for a road section was 96. Nearly all road sections (21,044) had a length of 1 kilometre. Table 2 contains summary statistics for all variables.

4. RESULTS

Table 6 gives the estimates of regression coefficients β and their standard errors based on the Poisson-lognormal (PLN) model. In addition table 6 contains values of the goodness-of-fit measures. It is easily seen that the parameters differ significantly between g_1 and the other functions. The difference is attributable to the inclusion of the logarithm of AADT as an explanatory variable. When the natural logarithm of AADT is used model fit improves significantly (from MAPE=93% to MAPE=74% when comparing models g_1 and g_2.

Table 5. Data used in present study

Variable name	Symbol	Mean	Standard deviation	Minimum	Maximum
Dependent variable					
Total number of accidents		1.3090	3.4446	0	96
Independent variables					
Average annual daily traffic (AADT)	AADT	2347.3	4891.220	8	86307
The natural logarithm of AADT		6.981	1.164	2.08	11.37
Number of lanes	X_1	2.016	0.232	1	8
Number of junctions per kilometre	X_2	0.200	0.425	0	28
The natural logarithm of (number of junction + 1)		0.387	1.052	0	3.37
Dummy for trunk road (1 = yes, 0 = no)	X_3	0.269		0	1
County (identified by number from 1 to 20)	X_4	12.016	5.878	1	20
Motorway type B (1 = yes, 0 = no)	X_5	0.010		0	1
Motorway type A (1 = yes, 0 = no)	X_6	0.001		0	1
Rural road with speed limit 90 km/h (1 = yes, 0 = no)	X_7	0.039		0	1
Speed limit (km/h)	X_8	75.454	10.107	30	100
Speed limit squared (km^2/h^2)		5795.443	0.086	900	10000
Segment length multiplied by years of data (km per year) (SLTM)	X_9	7.599	0.991	2.08	8
The natural logarithm of (number of lanes + 1)		0.697	0.086	0	2.08

Source: personal study.

From tables 2-4 it can be seen that the PLN (Poisson lognormal) and NB (negative binomial) distributions fit much better than standard Poisson distribution. Both of them generate long tails which are very important especially for variables with higher variation. The hypothesis on the matching of the distributions with theoretical values was tested. The values of statistics χ^2 clearly indicate that PLN empirical model is the best fit. Therefore the PLN model was used in further considerations.

Table 6. Estimation results for Poisson-lognormal model – coefficients and standard errors

	Functional form g_1		Functional form g_2		Functional form g_3		Functional form g_4		Functional form g_5		Functional form g_6	
Variable name	**Coeffi-cient**	**Standard error**	**Coeffi-cient**	**Standard error**	**Coeffi-cient**	**Standard error**	**Coeffi-cient**	**Standard error**	**Coeffi-cient**	**Standard error**	**Coeff-icient**	**Standard error**
Constant	2.8587	0.0314	-6.3089	0.0814	-6.3485	0.0843	-4.7315	0.1787	-6.2742	0.0832	-4.8903	0.1749
County	-0.0462	0.0008	0.0008	0.0009	0.0015	0.0010	0.0010	0.0009	0.0006	0.0009	-0.0256	0.0052
County squared	-	-	-	-	-	-	-	-	-	-	0.1490	0.0284
Speed limit	-0.0502	0.0003	-0.0268	0.0004	-0.0272	0.0004	-0.0761	0.0050	-0.0268	0.0004	-0.0762	0.0050
Speed limit squared	-	-	-	-	-	-	0.0004	0.0000	-	-	0.0004	0.0000
Number of junctions	0.0888	0.0023	0.0397	0.0022	-	-	0.0381	0.0022	0.0399	0.0022	0.0460	0.0054
Dummy for trunk road	0.3402	0.0104	-0.1227	0.0115	-0.1196	0.0119	-0.1255	0.0117	-0.1205	0.0115	-0.1224	0.0116
Motorway type B	0.7987	0.0448	-0.1088	0.0445	-0.0992	0.0459	-0.2631	0.0477	-0.1110	0.0444	-0.2648	0.0475
Motorway type A	0.2326	0.0705	-0.5885	0.1107	-0.5855	0.1159	-0.7206	0.1134	-0.6003	0.1109	-0.7462	0.1113
Rural road with speed limit 90 km/h	0.2191	0.0328	-0.2470	0.0325	-0.2435	0.0335	-0.3909	0.0362	-0.2617	0.0324	-0.3833	0.0360
SLTM (length · years)	0.2476	0.0033	0.1819	0.0042	0.1845	0.0044	0.1831	0.0043	0.1815	0.0043	0.1843	0.0043
AADT/1000	0.0374	0.0003	-0.0015	0.0007	-0.0019	0.0007	-0.0011	0.0007	-0.0005	0.0007	-0.0009	0.0006
Natural logarithm of AADT	-	-	0.9228	0.0072	0.9252	0.0075	0.9239	0.0073	0.9175	0.0074	0.9219	0.0073
Number of lanes	-0.4172	0.0041	0.0100	0.0085	0.0143	0.0094	-0.0045	0.0089	-0.1109	0.0279	-0.0360	0.0170
Natural logarithm of (num. of lanes + 1)	-	-	-	-	-	-	-	-	0.3580	0.0710	-	-
Natural logarithm of (num. of junctions + 1)	-	-	-	-	0.1096	0.0071	-	-	-	-	-	-
Error term (over-dispersion parameter)	0.3187	0.0032	0.2395	0.0039	0.2612	0.0042	0.2475	0.0041	0.2380	0.0039	0.2395	0.0039
AIC	2.7775		2.3975		2.3916		2.3936		2.3980		2.3965	
BIC	71511		61735		61584		61639		61751		61718	
MAD	1.2227		0.9686		0.9715		0.9684		0.9683		0.9668	
MSPE	9.9507		4.8592		4.9490		4.8362		4.8393		4.7937	
MAPE	93.35		73.76		73.97		73.71		73.77		73.57	
Restricted log-likelihood	-39424		-32145		-32187		-32110		-32140		-32096	
Log-likelihood at convergence	-35729		-30839		-30763		-30789		-30845		-30828	

Source: personal study.

Goodness-of-fit measures based on log-likelihood values (AIC and BIC) show that g_3 is the best functional form. However, goodness-of-fit measures based on residuals (MAD, MSPE, MAPE) show that g_6 gives the best fit. It is important to point out that all functional forms from g_2 to g_6 have very similar goodness-of-fit values. The only function that differs markedly from the others is g_1. This indicates that in accident modelling the natural logarithm of AADT should be used to represent traffic volume. Other changes in the set of independent variables do not produce large differences in modelling results. The difference between g_3 and g_2 concerns how the number of junctions was included in the model (in natural units in g_2 and by means of the natural logarithm in g_3. This change does not change model fit importantly (AIC, BIC and are a little smaller for g_3 but MAD MSPE and MAPE are bigger). Adding speed limit squared variable (g_4) to the set of variables changes goodness-of-fit measures by less than 0.5 %. A similar conclusion applies to the comparison between g_2 and g_5 with respect to the natural logarithm of the number of lanes.

To decide which functional form (g_3 or g_6) fits best, CURE plots with box-whiskers extension for the variables with a limited range of values (figures 1 - 4) and a traditional CURE plot for AADT (figure 5) were used.

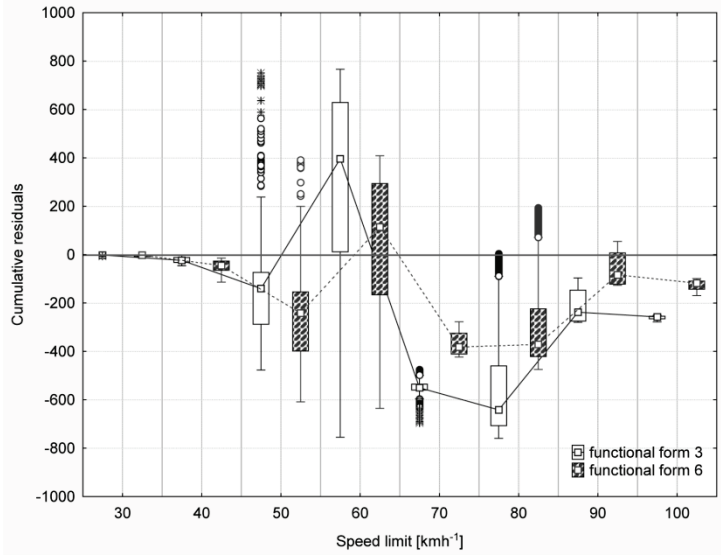

Fig. 1. Cure plot comparing functional forms 3 and 6 with respect to speed limit
Source: personal study

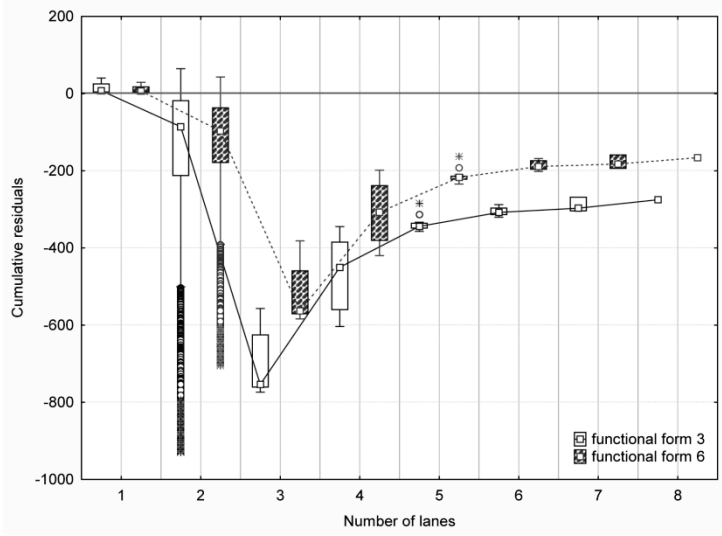

Fig. 2. Cure plot comparing functional forms 3 and 6 with respect to number of lanes
Source: personal study.

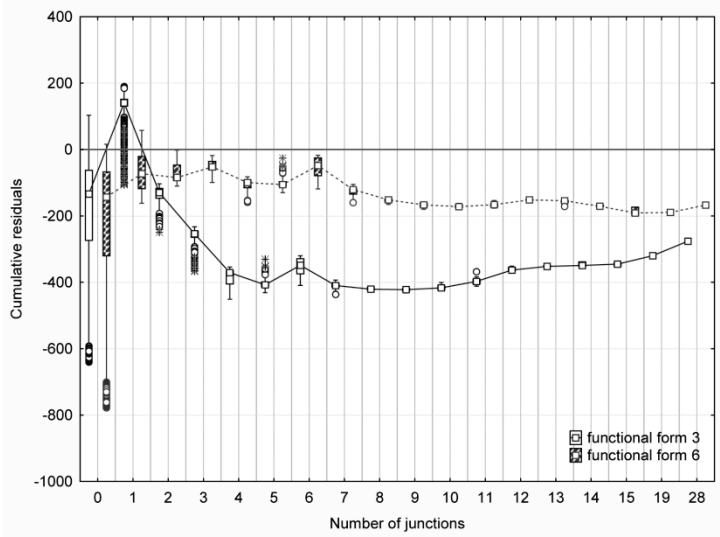

Fig. 3. Cure plot comparing functional forms 3 and 6 with respect to number of junctions per kilometre
Source: personal study.

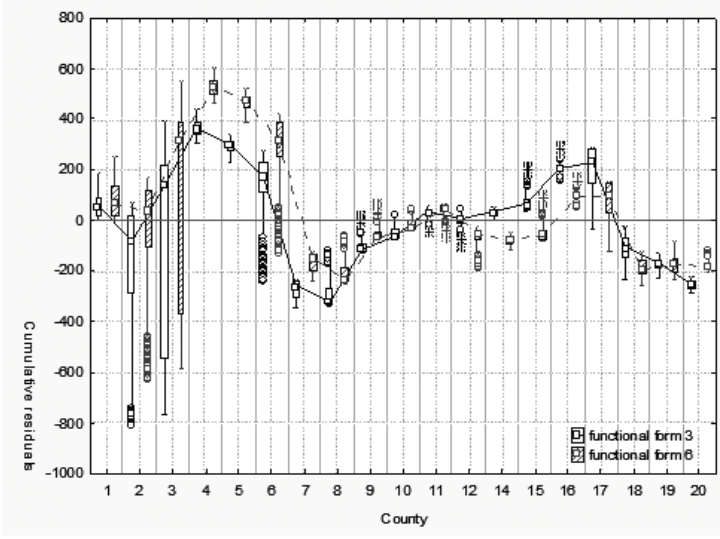

Fig. 4. Cure plot comparing functional forms 3 and 6 with respect to county
Source: personal study.

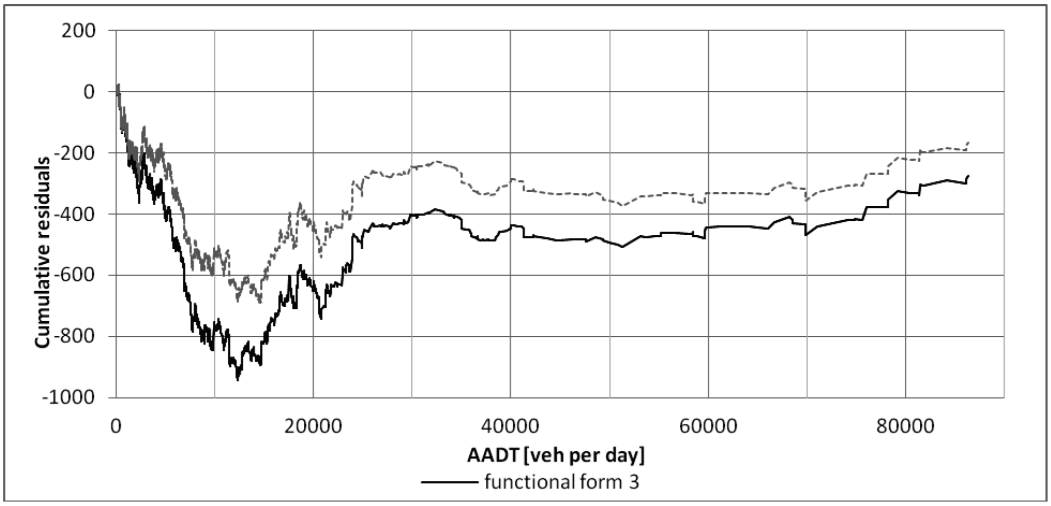

Fig. 5. Cure plot comparing functional forms 3 and 6 with respect to AADT
Source: personal study.

The plots in figures 1 to 5 can be used to decide which functional form g_3 or g_6 is better. Irrespective of the values of number of lanes, number of junctions and AADT, g_6 fits the data better than g_3 does. Figure 1 shows that the curve for g_3 oscillates closer to the zero line than for g_6, but only for speed limit lower than 60 km per hour. were observed for the segments within county Oslo area, which includes the city that is biggest agglomeration in Norway. The fit for the areas of the biggest density of population is the worst. The number of residuals in each range is very similar for functional forms $g_2 - g_6$, therefore it is difficult to pick out the best variant on this basis.

Table 7. Residua for zero accidents sectors

Residua	Functional form g_1	Functional form g_2	Functional form g_3	Functional form g_4	Functional form g_5	Functional form g_6
>100	2	0	0	0	0	0
<100, 50)	15	5	5	5	5	5
<50,30)	16	11	12	12	12	12
<30,10)	143	87	86	89	87	91
<10,8)	91	59	58	54	57	54
<8,6)	149	134	136	145	133	135
<6,4)	380	388	393	370	392	380
<4,2)	1445	1269	1257	1256	1267	1251
<2,1)	2616	2310	2313	2319	2306	2327
<1,0>	9263	9846	9839	9859	9850	9854

Source: personal study.

Box and whiskers plot indicate that cumulative residuals for g_6 are less differentiated (smaller boxes and shorter whiskers) than for functional form g_3 for each speed limit between 50 and 90 km per hour. Figure 3 shows that the curve for g_3 is crossing the zero line around value 1. That means that there is a significant change of residual values. Although both curves oscillate in similar distance from the zero line for number of junctions between 0 and 2, function g_6 generates smaller changes of values and therefore fits the data more closely. As far as counties are concerned, the two functions do not differ greatly, see Figure 4. With respect to traffic volume (AADT), functional form g_6 generally fits the data better than functional form g_3. Based on the CURE plots, it is therefore concluded that functional form 6 is the best model for predicting the number of accidents on national roads in Norway.

In order to compare the exactness of accidents numbers predictions, the analysis of the residuals for the biggest group of segments – with the zero number of accidents – constituting 55% of all sections. The functional form g_1 definitely showed the worst fit within the range of zero values.

Using each of the remaining functional forms ($g_2 - g_6$) allows us to reach the prediction errors that don't exceed one (accident) in 70% of cases (see table 7.). The biggest difference between the real and predicted value is accordingly: for g_1 – 149 and 78 for the others. The biggest differences

For the detailed analysis of the quantitative influence of particular factors (variables) on the accidents number the elasticities were appointed (see table 8.). For the model used (logarithmic) elasticities are equivalent to the value of the coefficient estimates. Elasticities E_k for dummy variables X_k were appointed according to the following dependence (Halvorsen and Palmquist 1980).

$$E_k = \exp(\beta_k) - 1 \qquad (15)$$

As the model g_1 was definitely less adequate, it was omitted in further considerations.

Table 8. Elasticity estimates

Variable name	Functional form g_2	Functional form g_3	Functional form g_4	Functional form g_5	Functional form g_6
County	0.0008	0.0015	0.0010	0.0006	-0.0256
County squared	-	-	-	-	0.1490
Speed limit	-0.0268	-0.0272	-0.0761	-0.0268	-0.0762
Speed limit squared	-	-	0.0004	-	0.0004
Number of junctions	0.0397	-	0.0381	0.0399	0.0460
Dummy for trunk road	-0.1154	-0.1127	-0.1179	-0.1136	-0.1152
Motorway type B	-0.1031	-0.0944	-0.2313	-0.1050	-0.2326
Motorway type A	-0.4449	-0.4432	-0.5135	-0.4513	-0.5258
Rural road with speed limit 90 km/h	-0.2189	-0.2162	-0.3235	-0.2302	-0.3184
SLTM (length · years)	0.1819	0.1845	0.1831	0.1815	0.1843
AADT/1000	-0.0015	-0.0019	-0.0011	-0.0005	-0.0009
Natural logarithm of AADT	0.9228	0.9252	0.9239	0.9175	0.9219
Number of lanes	0.0100	0.0143	-0.0045	-0.1109	-0.0360
Natural logarithm of (num. of lanes + 1)	-	-	-	0.3580	-
Natural logarithm of (num. of junctions + 1)	-	0.1096	-	-	-

Source: personal study.

On the basis of the values of elasticity coefficients shown in table 8 it should be assumed that the models presented match each other. Apart from the number of lanes the signs and values of elasticities are similar. In case of country number for g_6 the coefficient has opposite sign that in all the other cases. It is the result of introducing into the model the g_6 variable that is county number square for which the coefficient is positive. As both variables describe the same feature, their elasticities should be joined. Because of the power relations between variables the coefficients cannot be added. Nevertheless it can be stated that their compound will result in positive value for the values of the variable being less than 33, that is for all the counties in consideration (there are 20 of them).

Similarly in case of speed limit and speed limit squared: in case of g_4 and $g_6$the values are joined and the actual influence of this variable on the number of accidents is similar as in the other cases. Generally on the basis of the conducted analysis it can be stated that as the traffic flow, number of junctions, county number, segment length and time of measure increases the number of accidents within the segment also rises.

Traffic flow has the biggest impact. The increase of traffic flow by 10% causes estimated raise in the number of accidents by circa 9%. Negative values of the elasticities for variable AADT/1000 should be treated as additive correction for multiplicative variable. The influence of the number of junctions on the number of accidents differs even by 30% between the particular models.

Model g_6 ascribes the greatest importance to the number of junctions. According to the models the bigger the speed limit is the smaller the number of accidents in the given segment. This correlation is also supported by negative coefficients for dummy of road type variable. This phenomenon is also consistent with intuition, as within the areas of the biggest traffic flow in the built-up areas where the speed limits are lower (40 or 50 km/h) the number of accidents is the biggest.

Milton and Mannering (1998) also got negative values of speed limit coefficient (-0,0072 and –0,0064). The bigger the number of junctions within a segment the bigger – according to the models – the number of accidents. Increasing the number of junctions by 10% causes estimated rise in the number of accidents by values from 0,38% (for g_4) to 0,46% (for g_6). Similar results were obtained by El-Besyouny and Sayed (2009) who got coefficient equalling 0,96 for PLN model.

The authors did not take into consideration the influence of speed limit which is relevantly correlated with the number of junctions. As a result the number of junctions in El-Besyouny and Sayed corresponds with two variables in the present models which allows as to draw the conclusion on the matching of the results.

5. CONCLUSIONS

The following points summarise the main findings of the research reported in this paper:

1. Poisson log-normal model is the most adequate (in comparison with the negative binomial and Poisson models) for modelling the number of accidents on the basis of the Norwegian data analysed in this paper.
2. There are substantial differences between modelling results for the first functional form and all the others, which shows that in a list of independent variables both AADT and its natural logarithm (lnAADT) should be present.
3. The choice of functional form is very important in accident prediction models. One should always check a few functional forms and choose the most appropriate functional form based on goodness-of-fit measures, CURE plots and other methods.
4. Making a decision solely on the basis of goodness-of-fit measures based on log-likelihood values (AIC, BIC) should be avoided as it can lead to wrong conclusions.
5. Functional form number 6 was found to be the best to model to predict the number of injury accidents on national roads in Norway.

REFERENCES

[1] Abdel-Aty, M. A., Radwan, A. E., 2000. Modeling traffic accident occurrence and involvement. Accident Analysis and Prevention, 32, 633-642.

[2] Anastasopoulos, P. C., Mannering, F. L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention, 41, 153-159.

[3] Couto, A., Ferreira S., 2011. A note on modeling road accident frequency: A flexible elasticity model. Accident Analysis and Prevention, 43, 2104-2111.

[4] El-Basyouny, K., Sayed, T., 2009A. Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis and Prevention, 41, 820-828.

[5] El-Basyouny, K., Sayed T., 2009B. Accident prediction models with random corridor parameters. Accident Analysis and Prevention, 41, 1118-1123.

[6] Elvik, R., 2008. The predictive validity of empirical Bayes estimates of road safety. Accident Analysis and Prevention, 40, 1964-1969

[7] Halvorsen, R., Palmquist, R., 1980. The interpretation of dummy variables in semilogarithmic equation. American Revew 70 (30), 474-475.

[8] Hauer, E., 1986. On the estimation of the expected number of accidents. Accident Analysis and Prevention, 18, 1-12.

[9] Hauer, E. 2004. Statistical road safety modeling. Transportation Research Record, 1897, 81-87.

[10] Hauer, E., Bamfo, J., 1997. Two tools for finding what function links the dependent variable to the explanatory variables. In: Proceedings of the ICTCT 1997 Conference, Lund, Sweden.

[11] Karlaftis, M. G., Tarko, A. P., 1997. Heterogeneity considerations in accident modeling. Accident Analysis and Prevention, 30, 425-433.

[12] Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes Support Vector Machine models. Accident Analysis and Prevention 40, 1611-1618.

[13] Li, C. C., Lu, J. C., Park, J., Kim, K., Brinkley, P. A., Peterson, J. P., 1999. Multivariate zero-inflated Poisson models and their applications. Technometrics, 41 (1), 29-38.

[14] Lin, D. Y., Wei, L. J., Ying, Z., 2002. Model-Checking Techniques Based on Cumulative Residuals. Biometrics, 58, 1-12.

[15] Lord, D., Guikema, S. D., Geedipally, S. R., 2009. Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis and Prevention, 41, 1123-1134.

[16] Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A, 44, 291-305.

[17] Lord, D., Miranda-Moreno, L. F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: a Bayesian perspective. Safety Science, 46, 751-770.

[18] Lord, D., Park, P. Y-J., 2008. Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. Accident Analysis and Prevention, 40, 1441-1457.

[19] Lord, D., Persaud, B. N., 2000. Accident prediction models with and without trend: application of the Generalized Estimating Equations (GEE) procedure. Transportation Research Record, 1717, 102-108.

[20] Lord, D., Washington S. P., Ivan, J. N., 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention, 37, 35-46.

[21] Ma, J., Kochelman, K. M., 2006. Bayesian multivariate Poisson regression for models of

injury count by severity. Transportation Research Record 1950, 24-34.
[22] Ma, J., Kochelman, K. M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Analysis and Prevention, 40, 964-975.
[23] Maher, M., Mountain, L., 2009. The sensitivity of estimates of regression to the mean. Accident Analysis and Prevention, 41, 861-868.
[24] Maher, M.J., Summersgill, I., 1996. A comprehensive methodology for the fitting of predictive accident models. Accident Analysis and Prevention, vol.28, no.3, 281-296.
[25] Miaou. S. P., 1996. Measuring the Goodness-of-fit of Accident Prediction Models. Publication FHWA-RD-96-040. FHWA, U.S.DOT.
[26] Miaou, S.-P., Bligh, R. P., Lord, D., 2005. Developing median barrier installation guidelines: A benefit/cost analysis using Texas data. Transportation Research Record, 1904, 3-19.
[27] Miaou, S.-P., Song, J. J., Mallick, B. K., 2003. Roadway traffic crash mapping: a space-time modeling approach. Journal of Transportation and Statistics, 6, 33-57.
[28] Milton, J., Mannering, F., 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. Transportation 25, 395-413.
[29] Noland, R.B., Karlaftis, M.G., 2005. Sensitivity of crash models to alternative specifications. Transportation Research Part E, 41, 439-458.
[30] Park, B-J., Lord D., 2009. Application of finite mixture models for vehicle crash data analysis. Accident Analysis and Prevention, 41, 683-691.
[31] Park, E. S., Lord, D., 2007. Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. Transportation Research Record, 2019, 1-6.
[32] Satterthwaite, S.P., 1981. A survey of Reasearch into Relationship Between Traffic Accidents and Traffic Volumes. Supplementary Report 692. Transport Research Laboratory, Crowthorne, Berkshire.
[33] Tsionas, E. G., 2001. Bayesian Multivariate Poisson Regression. Communication in Statistics – Theory and Methods, Vol.30, no. 2, 243-255.
[34] Vogt, A., 1999. Crash Models for Rural Intersection: Four-lane by Two-lane Stop-controlled and Two-line by Two-line Signalized. Publication FHWA-RD-99-128. FHWA, U.S.DOT.
[35] Wang, X., Abdel-Aty, M., 2007. Right-Angle Crash Occurrence at Signalized Intersections. Transportation Research Record, 2019, 156-168.
[36] Winkelmann, R., 2008. Econometric Analysis of Count Data. Fifth edition. New York, Springer

Joanna Kamińska
Wrocław University of Environmental and Life Sciences Department of Mathematics, Poland
joanna.kaminska@up.wroc.pl