Accident prediction models for urban and rural carriageways

Martine Reurings & Theo Janssen

R-2006-14

Accident prediction models for urban and rural carriageways

Based on data from The Hague region Haaglanden

R-2006-14

Martine Reurings & Theo Janssen

Leidschendam, 2007
SWOV Institute for Road Safety Research, The Netherlands

Report documentation

Number:	R-2006-14
Title:	Accident prediction models for urban and rural carriageways
Subtitle:	Based on data from The Hague region Haaglanden
Author(s):	Martine Reurings & Theo Janssen
Project leader:	Theo Janssen
Project number SWOV:	39.150
Keywords:	Mathematical model, accident rate, risk, prediction, traffic concentration, accident prediction model, Netherlands

Contents of the project: In this report accident prediction models are discussed for carriageways of urban and rural distributor roads in the region Haaglanden. They express the number of injury accidents on a carriageway in its length and its average amount of daily traffic. Not only distinction is made between urban and rural carriageways, but also between carriageways with one and two driving directions.

Number of pages:	81
Price:	€ 15,-
Published by:	SWOV, Leidschendam, 2007

SWOV Institute for Road Safety Research
P.O. Box 1090
2260 BB Leidschendam
The Netherlands
Telephone +31 70 317 33 33
Telefax +31 70 320 12 61
E-mail info@swov.nl
Internet www.swov.nl

Summary

The SWOV-project *Infrastructure and Road Safety* aimed to find (mathematical) relations between characteristics of the Dutch road infrastructure and road safety. Such relations are often called accident prediction models (APMs). The SWOV-project developed APMs for distributor roads in The Hague region Haaglanden and for provincial roads in the provinces Gelderland and Noord-Holland. This report discusses the APMs for the distributor roads in Haaglanden. This part of the project was carried out in the European project RIPCORD-ISEREST.

In order to develop APMs a database is needed which contains several road characteristics, including the average amount of daily traffic (AADT) and road length. The number of road crashes in a certain period should also be known. For Haaglanden the database Wegkenmerken+ meets these conditions. Wegkenmerken+ is based on the Dutch National Roads Database (NWB). As a consequence single and dual carriageway roads are treated differently. A dual carriageway road is a road on which the driving directions are separated by a physical barrier, so a dual carriageway road consists of two carriageways and each of these carriageways has one driving direction. A single carriageway road consists of only one carriageway and this carriageway can have one or two driving directions. In the NWB, and hence in Wegkenmerken+, the road characteristics are listed per carriageway, so for a dual carriageway road the characteristics are given separately for each carriageway. The accident prediction models are therefore not models for roads, but for carriageways.

Examples of the road characteristics listed in Wegkenmerken+ are location (urban or rural), the speed limit, the type of road the carriageway is part of (single or double carriageway) and the number of driving directions. These characteristics are used to define the following carriaway types for which APMs are developed:

- carriageways of distributor roads inside urban areas;
- carriageways of distributor roads outside urban areas;
- carriageways of dual carriageway distributor roads inside urban areas, with a speed limit of 50 km/h, one lane in each driving direction;
- carriageways of single carriageway distributor roads inside urban areas, with a speed limit of 50 km/h, two lanes and two driving directions.

Two model forms are tested, namely

$$\mu_i = e^{\alpha} \cdot AADT_i^{\beta_1} \cdot L_i^{\beta_2}$$

and

$$\mu_i = e^{\alpha} \cdot AADT_i^{\beta_1} \cdot L_i^{\beta_2} \cdot e^{\beta_3 \cdot \frac{AADT_i}{1000}},$$

where μ_i is the expected number of road crashes on carriageway segment i in three years, $AADT_i$ is the AADT of that carriageway segment and L_i the length. The second form turned out to be the best for all carriageway types except for carriageways of single carriageway distributor roads inside urban areas, with a speed limit of 50 km/h, two lanes and two driving directions. The model parameters α, β_1, β_2 and β_3 are estimated with the GENMOD procedure of SAS 9.1. The procedure uses generalized linear modelling, a technique which is often used in the literature to develop APMs. The fit of the

models is extensively checked by conducting several statistical tests on the deviance, the parameter estimates and the standardized deviance residuals. The conclusion of these tests was that all models fit the data reasonably well.

The developed APMs were compared to each other. The following conclusions could be drawn:

- for AADT $\leq \pm$ 25000 carriageways inside urban areas generally have a lower crash rate (number of crashes per motor vehicle kilometre) than carriageways outside urban areas, see *Figure 4.27*;
- carriageways with a speed limit of 50 km/h or 80 km/h and one driving direction have a lower crash rate than carriageways with the same speed limit but with two driving directions;
- the average crash rate of urban carriageways with a speed limit of 70 km/h is lower than the crash rate of carriageways with a speed limit of 50 km/h
- the average crash rate of rural carriageways with a speed limit of 60 km/h is almost the same as the crash rate of rural carriageways with a speed limit of 80 km/h and two driving directions.

Some of these conclusions are counterintuitive. For example, the one which states that urban carriageways with a speed limit of 70 km/h have a lower crash rate than urban carriageways with a speed limit of 50 km/h. However, this conclusion does not state that reducing the speed limit increases the crash rate. For this type of conclusions before and after studies are necessary.

Because of the limited size of the database, it was not possible to develop APMs for more detailed carriageway types. Therefore we recommend to collect more data on more roads for further research. This data should not only include characteristics of road segments, but also characteristics of intersections. So far intersections were not considered separately, they were considered as part of carriageways. By developing models for intersections it is possible to investigate the influence of intersection characteristics on the safety of intersections.

Contents

1. Introduction

The research in this report is part of the project *Infrastructure and Road Safety* which was carried out at the SWOV Institute for Road Safety Research. The general goal of this project is to find relations between characteristics of the Dutch road infrastructure on the one hand and road safety on the other hand, using risk and exposure measures. This general goal is translated into the following two more specific goals:

- to get insight into the quantitative road safety aspects of infrastructural characteristics within certain road categories;
- to get insight into the quantitative road safety aspects of infrastructural characteristics between certain road categories.

The project is partly embedded in the European project RIPCORD-ISEREST. The aim of this project is to give scientific support to the European transport policy to reach the 2010th transport road safety target by establishing best practice tools and guidelines for road infrastructure safety measures. To do so, good insight is needed in the variables that explain the crash levels on roads and networks. Variables, for example, are the average amount daily traffic (AADT), the width of a road, the number of lanes, the presence or absence of bicycle lanes and the way the priority on an intersection is organized. The relation between the safety and these factors can be described by the mathematical models like the accident prediction model (APM) and the road safety impact assessment (RIA). These models are the subject of Workpackage 2 of RIPCORD-ISEREST, which started with making an overview of the state-of-the-art on accident prediction models and road safety impact assessments, see Reurings et al. (2005).

The next step in this workpackage consists of pilot studies which are carried out in the four participating countries: Austria, the Netherlands, Norway and Portugal. In these pilot studies accident prediction models are developed, based on the models and modelling techniques discussed in the state-of-the-art report. This report discusses the pilot study carried out by SWOV in the Netherlands. Accident prediction models have been derived for the Dutch city region Haaglanden, an area consisting of The Hague and surroundings. Haaglanden was chosen because the road characteristics database Wegkenmerken+ is most complete for this area. The models have been developed using the generalized linear modelling technique.

This report first makes some preliminary remarks about road characteristics which may have an influence on road safety and about the database containing the carriageways of Haaglanden in *Chapter 2*. Then in *Chapters 3, 4* and *5* several models are developed, compared and discussed for these carriageways. The report ends with conclusions and recommendations in *Chapter 6*. The report also explains in which way modelling results can be used by road authorities. The *Appendix* gives a summary of generalized linear modelling.

2. Preliminary remarks about the models for Haaglanden

This chapter discusses some preliminaries which are important for the development of accident prediction models for the carriageways in Haaglanden. First, we give examples of road characteristics which may have an influence on road safety. Then we will introduce and discuss the database which is used . Next, the different structures of the models to be developed are given, and finally we explain how the crash rate can be visualized.

2.1. Infrastructural characteristics which influence road safety

A large number of road characteristics have a possible influence on road safety. There are three different types of characteristics: function, design and use. The function of a road can be considered as the possibility that is offered to a moving vehicle on that road. In the Sustainable Safety programme a distinction is made between three types of road function. The two ’extreme’ types are through-roads, for traffic dispersion, and access roads, for access to the destination. The third type, the distributor roads, are intended to make a good link between the two extreme types, both literally and figuratively. Distributor and access roads exist inside and outside urban areas, which means that there is a total of five road categories.

In an ideal situation the road design should be determined by the function. Important differences between the designs of the road categories are:

- the number of main carriageways and service roads;
- the type of road surface;
- the presence and type of edge and lane marking;
- the parking possibilities;
- the presence and type of exit roads.

Certain road use characteristics also have a large influence on road safety. A few examples are:

- the amount of traffic, given the number of carriageways, lanes and specific facilities;
- the type of traffic, given the access limitations;
- the traffic speed, given the speed limit;
- the number of driving directions per carriageway;
- the speed enforcement and other behavioural rules by the police.

The database Wegkenmerken+ contains several design characteristics for distributor roads in Haaglanden. The database will be discussed in more detail in the following section.

2.2. The database

The database which is used for the research in this report contains information about carriageways in the city region Haaglanden, as we already mentioned in the introduction. One of the consequences of the road characteristics being listed per carriageway, for example, is that the average amount of daily traffic (AADT) ofF segments of dual carriageway roads is given separately for each carriageway and hence for each driving direction, whereas the AADT of single carriageway roads is the sum of both driving directions. All the carriageways are part of distributor roads, both inside and outside urban areas. The first type will be referred to as urban carriageways, whereas the latter will be called rural carriageways.

Besides the functional characteristic of the carriageways, that of urban or rural distributor road, the database also contains some design and use characteristics. The design characteristics which are reasonably well listed are the number of main carriageways of the road segment the carriageway belongs to and the presence of parallel facilities such as bicycle paths and service roads. In the database, the use of the carriageways is fairly well described by the average amount of daily traffic, the speed limit and the number of driving directions. Other road characteristics in the database are:

- the length of the carriageway in metres;
- the number of speed humps;
- the number of exits;
- the type of limited access;
- the bicycle and/or moped facilities;
- the road surface;
- the parking facilities;
- the type of edge marking.

Carriageways for which these characteristics are the same are taken together and form one new carriageway. This procedure results in 303 carriageways inside and 98 carriageways outside urban areas. For all these combined carriageways the database also contains the number of crashes which occurred on each carriageway in the 2000-2002 period. These crashes include those that happened on intersections.

Based on the available road characteristics in the database it is possible to define several road types (or actually carriageway types). Together with the working group Haaglanden we decided to distinguish the following types:

- carriageways of distributor roads inside urban areas, with a speed limit of 50 km/h and one driving direction;
- carriageways of distributor roads inside urban areas, with a speed limit of 50 km/h and two driving directions;
- carriageways of dual carriageway distributor roads inside urban areas, with a speed limit of 50 km/h, one lane and one driving direction;
- carriageways of single carriageway distributor roads inside urban areas, with a speed limit of 50 km/h, two lanes and two driving directions;
- carriageways of distributor roads inside urban areas, with a speed limit of 70 km/h;
- carriageways of distributor roads outside urban areas, with a speed limit of 60 km/h;
- carriageways of distributor roads outside urban areas, with a speed limit of 80 km/h and one driving direction;
- carriageways of distributor roads outside urban areas, with a speed limit of 80 km/h and two driving directions.

Table 2.1 shows the crash rate of these carriageway types, the number of injury crashes per year divided by the motor vehicle kilometres per year. It should be remarked that the vehicle kilometres are measured in 2003, while the number of crashes per year is the average over the years 2000-2002. It is possible to compute the vehicle kilometres for 2001 by assuming a constant traffic growth each year. This results in AADTs which are a constant factor smaller than the AADTs in 2003. Because this constant factor, i.e. the traffic growth, is not exactly known, we did not use this method for the research presented in this report. Instead we assumed that the AADT in 2003 is an appropriate estimate for the AADTs in the years 2000-2002.

Carriageway type	Total length	AADT	Vehicle km	Injury crashes	Crash rate
All carriageways	524	11934	2282	1051	0.46
All urban carriageways	413	11716	1765	944	0.54
Urban, 50 km/h, one direction	244	11955	1065	511	0.48
Urban, 50 km/h, two directions	146	9670	514	410	0.80
Urban, 50 km/h, one direction, dual	242	11966	1057	501	0.47
Urban, 50 km/h, two directions, single	145	9679	513	410	0.80
Urban, 70 km/h	13	29527	139	16	0.11
All rural carriageways	111	12746	517	107	0.21
Rural, 60 km/h	20	11275	84	24	0.28
Rural, 80 km/h, one direction	37	14502	198	26	0.13
Rural, 80 km/h, two directions	45	11296	184	48	0.26

Table 2.1. *The average crash rate over 2000-2002 for the different carriageway types.*

Based on *Table 2.1* several conclusions can be drawn. For example, carriageways with one driving direction are safer than carriageways with two driving directions. However, it is not clear how influential the AADT is on the crash rate, while it is intuitively clear that AADT does have an influence. Extensive models for each carriageway type are needed to determine the AADT influence. In *Chapters 3* and *4* models are developed for the complete selection of urban roads and the complete selection of rural roads. *Chapter 5* describes the problems with disaggregating the models to speed limit. Some general results are given.

2.3. The different forms of the models

Reurings et al. (2005) concluded that an accident prediction model for road segments should be of the following form:

$$\mu_i = \alpha \cdot AADT_i^{\beta} \cdot e^{\gamma_j \cdot x_{ij}},$$

where μ is the expected number of road crashes in a certain period, $AADT$ is the AADT in that same period, $\mathbf{x}_j$are other explanatory variables, α, β, γ_j are the parameters to be estimated and the subscript i denotes the value of a variable for the i-th road segment.

According to Reurings et al. (2005) the other explanatory variables should at least include the (logarithm of the) segment length, the number of exits, the carriageway width and the shoulder width. However, in this study we prefer to develop separate models for different road types instead of including the variables which characterize a particular road type in the models. The only two explanatory variables will be the carriageway length and the AADT. The main focus will be on the two main road types: urban and rural distributor roads. Also models should be developed for the other road types, but due to low numbers of carriageways for most of the road types this is not possible for all types. The types for which models are not developed will be compared with simple plots.

Two types of model are used, the first of which is directly based on the conclusions of Reurings et al. (2005). It is given by

$$\mu_i = e^{\alpha} \cdot AADT_i^{\beta_1} \cdot L_i^{\beta_2}, \tag{2.1}$$

where L_i is the value of the variable L for carriageway i, i.e., L_i is the length (in metres) of carriageway i. It is obvious that *(2.1)* can be rewritten as

$$\log(\mu_i) = \alpha + \beta_1 \cdot \log(AADT_i) + \beta_2 \cdot \log(L_i), \tag{2.2}$$

which actually is a generalized linear model. The values of the parameters α, β_1 and β_2 will be determined by using the GENMOD procedure in SAS in the three different ways described in the *Appendix*. It will be assumed that the number of crashes is Poisson or negative binomially distributed and should hence be integers. Therefore the parameters cannot be estimated based on the average number of crashes in 2000-2002 and hence the total number of crashes in 2000-2002 will be used as observations. As a consequence μ_i will not be the predicted number of crashes per year but per three years.

The parameter β_2 will be very close to 1 for almost all models developed in *Chapter 3*. Ignoring this parameter and dividing both sides of *(2.1)* by L_i and 3 results in a model for the number of road crashes per metre per year. So if β_1 is positive, then the number of road crashes per metre is increasing for increasing $AADT$ and if β_1 is negative, then the number of road crashes is decreasing for increasing $AADT$. Neither of these possibilities is the case in practice. This is made clear by *Figures 2.1* and *2.2*. For the graph in *Figure 2.1* the urban carriageways were divided into the following AADT classes:

1. AADT $<$ 5000;
2. 5000 $\leq$ AADT $<$ 10000;
3. 10000 $\leq$ AADT $<$ 15000;
4. 15000 $\leq$ AADT $<$ 20000;
5. 20000 $\leq$ AADT $<$ 30000;
6. 30000 $\leq$ AADT $<$ 40000;
7. AADT $\geq$ 40000.

For each of the classes the average AADT is computed and the total number of crashes is divided by the total length of the carriageways in kilometres. *Figure 2.2* shows the number of road crashes per kilometre of ten subsequent urban carriageways, where the carriageways are ordered by increasing AADT.

These figures show that the number of crashes per kilometre neither just increases nor decreases, indicating that the models developed in *Chapter 3* are not of the appropriate structure. An explanation for the shape of the graph in *Figure 2.1* can be that the database used consists of carriageways of very different types. Therefore, *Figure 2.1* does not indicate that an increasing AADT causes a lower number of crashes per kilometre on the same carriageway, but that carriageways with high AADT have fewer road crashes per kilometre because they are designed to be safer.

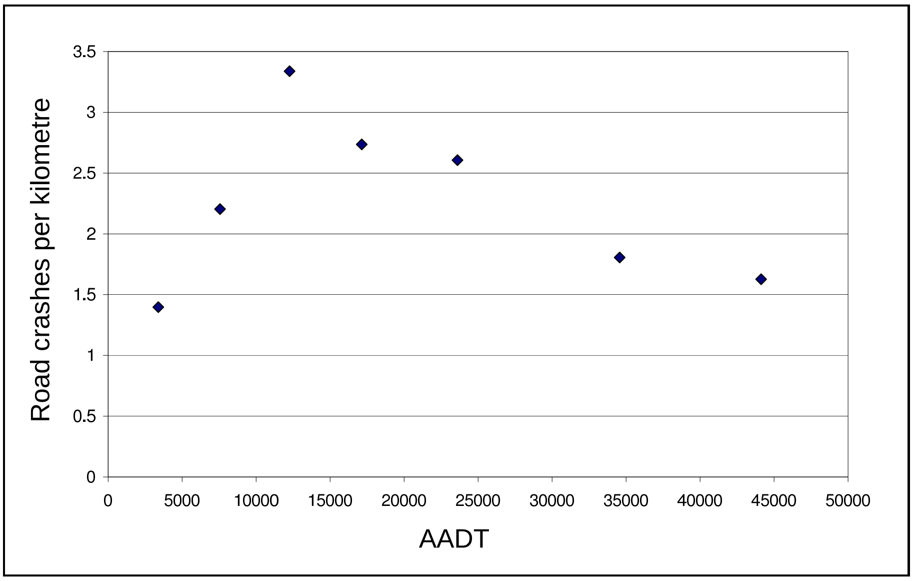

Figure 2.1. *The number of road crashes per kilometre per year against the average AADT for urban carriageways divided in seven AADT classes.*

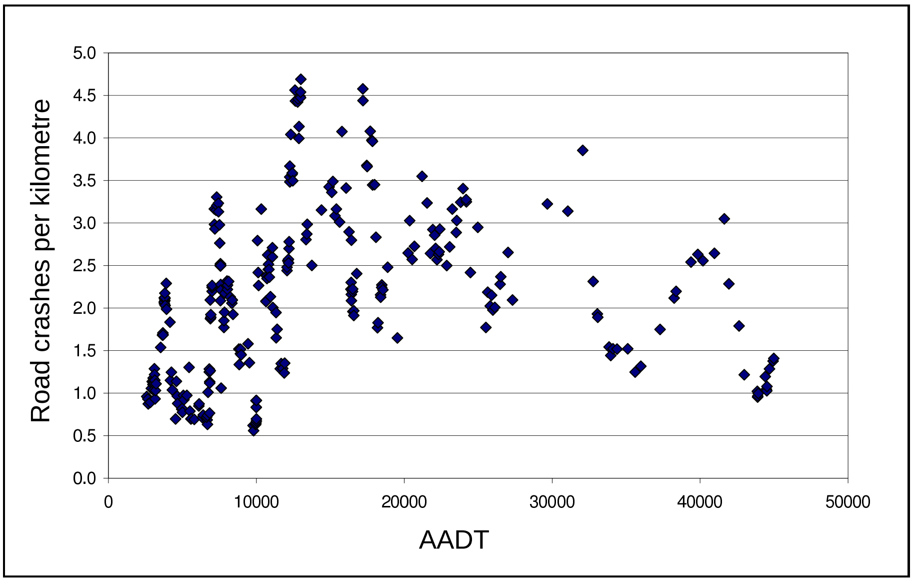

Figure 2.2. *The number of road crashes per kilometre per year for each ten subsequent urban carriageways against the average AADT.*

There are several ways to try to get models of a more appropriate structure. A first way is to define several classes for the AADT and to include the AADT as a class variable in the model rather than a continuous variable. This makes it possible to model a lower number of crashes for high AADTs. A disadvantage of this method is that only the intercept of the model has a different value for each class of the AADT; the parameter of log$(AADT)$ is the same for each class. This means that the number of road crashes per metre increases or decreases for all AADT classes. This is contradictory to the remarks above. A solution to this problem is to make the classes smaller or even to let each value of the AADT form its own class. This results in a large number of dummy variables and is hence not a practical solution.

Instead of adding the AADT as a class variable, it is also possible to develop a model for each level of the AADT separately. The parameter of log$(AADT)$ can then be different for each level. It can even be positive or negative for different levels. This comes close to the desired form of the model, but also this solution has a disadvantage: the model parameters have to be estimated based on a very small data set which makes the estimates unreliable.

Because of the mentioned disadvantages of the possible solutions, we decided to use a less insightful model structure:

$$\mu_i = \beta_0 \cdot AADT_i^{\beta_1} \cdot L_i^{\beta_2} \cdot e^{\beta_3 \cdot \frac{AADT_i}{1000}}. \tag{2.3}$$

The generalized linear model form of *(2.3)* is given by

$$\log(\mu_i) = \beta_0 + \beta_1 \cdot \log(AADT_i) + \beta_2 \cdot \log(L_i) + \beta_3 \cdot \frac{AADT_i}{1000}. \tag{2.4}$$

The AADT itself is added to model *(2.2)* next to its logarithm. The AADT could be considered as a property of the carriageway under consideration and hence as a sort of continuous dummy-variable. Because the AADT is very large compared to log(L), log$(AADT)$ and the number of road crashes, the estimated value of β_3 would be very small. Therefore the AADT is divided by 1000.

Like model *(2.1)*, models of the type *(2.3)* are developed based on the three different ways described in the *Appendix*. The results will be discussed in *Chapter 4*. Model *(2.1)*, or equivalently *(2.2)*, will be referred to as the simple model, whereas model *(2.3)*, or equivalently *(2.4)*, will be referred to as the extended model.

2.4. Visualising the risk of carriageways

Plots of the number of crashes per kilometre against the AADT (like the plots in *Figures 2.1* and *2.2*) can be used to visualize the crash rate of carriageways, which is defined as the number of road crashes per million vehicle kilometres, both per year. In formula:

$$r = \frac{y}{\frac{L}{1000} \cdot AADT \cdot 365 \cdot 10^{-6}},$$

where y is the number of crashes per year. The angle α between the x-axis and the line connecting the origin and one of the plotted points is given by

$$\alpha = \arctan\left(\frac{y}{\frac{L}{1000} \cdot AADT}\right) = \arctan\left(\frac{365}{10^6} \cdot r\right).$$

This shows that the larger the angle, the higher the risk. This is illustrated in *Figures 2.3* and *2.4*. In these figures the number of road crashes per kilometre per year is plotted against the AADT for urban and rural carriageways respectively. Line A in *Figure 2.3* makes a larger angle with the x-axis than Line B in the same figure, which implies that the crash rate of the carriageway corresponding to Point I (which is $r =$ 4.6104) is higher than the crash rate of the carriageway represented by Point II (which is $r =$ 1.2139). The two lines in *Figure 2.4* show that the crash rates of the carriageways corresponding to Point I and II are almost equal. Indeed, the crash rate of the first carriageway is $r =$ 1.1497, whereas the crash rate of the second carriageway is $r =$ 1.1107.

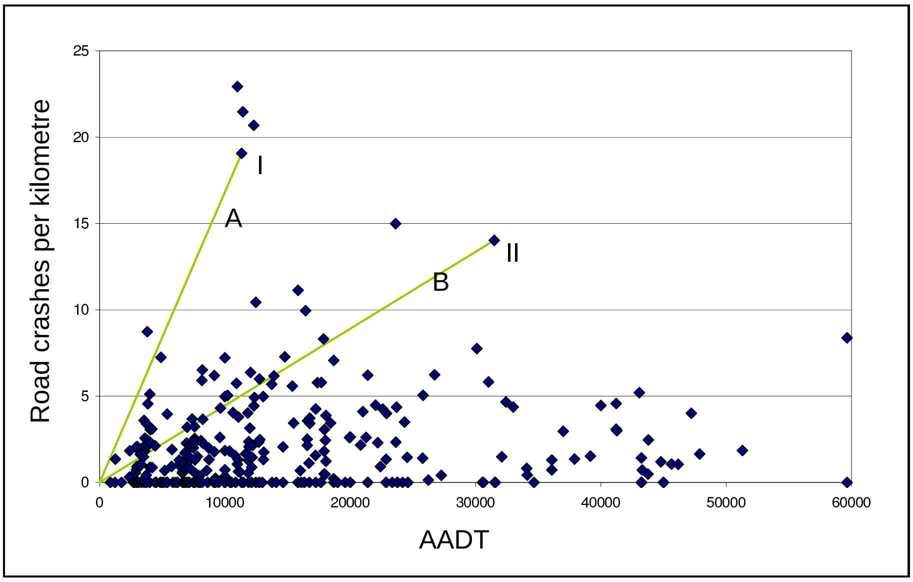

Figure 2.3. *The number of road crashes per kilometre against the AADT for urban carriageways.*

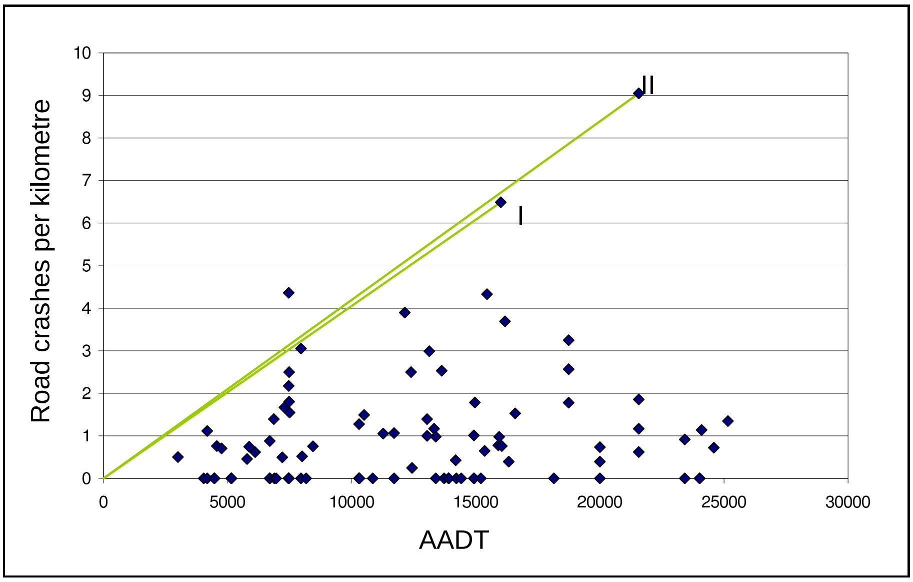

Figure 2.4. *The number of road crashes per kilometre against the AADT for rural carriageways.*

3. The simple model

In this chapter models of the form *(2.1)* are developed for a selection of carriageways in Haaglanden. The methods which are described in the *Appendix* will be used and compared. *Section 3.1* discusses the several models for urban carriageways, and the models for rural carriageways are the subject of *Section 3.2*.

3.1. Urban carriageways

3.1.1. *The Poisson distribution*

In this section a model is developed which describes the relation between the number of crashes on urban carriageways on the one hand, and the carriageway length and the AADT on the other, based on the assumption that road crashes are Poisson distributed. The statistics in *Table 3.1* describe the goodness-of-fit of the model.

Criterion	Degrees of freedom (DF)	Value	Value/DF
Deviance	300	1247.9565	4.1599
Pearson's χ^2	300	1459.4956	4.8650
Log likelihood		7044.5871	

Table 3.1. *Criteria for assessing the goodness-of-fit of the simple model for urban carriageways, based on the Poisson distribution.*

The deviance as well as Pearson's χ^2 is much larger than the number of degrees of freedom, which indicates the presence of overdispersion. This is not very surprising, because it already followed from the literature that the Poisson distribution is not the most appropriate distribution to use for the number of road crashes. A consequence of overdispersion is that carriageways with the same AADT and length can have a statistically significant different number of crashes, because the variance is rather large. Only the AADT and carriageway length are not enough to explain the number of crashes, therefore explanatory variables are missing.

In *Sections 3.1.2* and *3.1.3* two other types of models will be developed to solve the overdispersion problem. However, to allow comparison between the different models, the results of the analysis of the parameter estimates based on the Poisson distribution are given in the second column of *Table 3.2*.

Parameter	Estimate	Standard error	Wald's 95% confidence interval	Wald's χ^2	p-value
Intercept	-8.4758	0.3411	(-9.1444, -7.8072)	617.32	< 0.0001
log(L)	1.1210	0.0156	(1.0905, 1.1516)	5180.19	< 0.0001
log$(AADT)$	0.2703	0.0296	(0.2123, 0.3283)	83.40	< 0.0001

Table 3.2. *Analysis of the parameter estimates for the simple model for urban carriageways, based on the Poisson distribution.*

The modelled relationship between the expected number of crashes on urban carriageways in three years, the AADT and the carriageway length is hence given by

$$\hat{\mu}_i = e^{-8.4758} \cdot AADT_i^{0.2703} \cdot L_i^{1.1210} = 0.00021 \cdot AADT_i^{0.2703} \cdot L_i^{1.1210}.$$

The exponent of L is almost equal to 1, which shows that the number of road crashes is approximately proportional to the carriageway length, for constant AADT. In other words, the number of crashes per meter on a carriageway is independent of the length of the carriageway length. This could be counterintuitive, because there are reasons to expect that the risk for short carriageways is different than for long carriageways. For example, on short carriageways there is comparatively more accelerating and breaking and on long carriageways the average driven speed is probably higher. However, the exponent of L almost being equal to one does not indicate that this difference in crash rate exists.

In the third column of *Table 3.2* the standard errors of the estimates are stated, which are equal to the square roots of the estimated variances. The parameter estimates lie in the interval given in the fourth column with a probability of 0.95. The bounds of this interval for the j-th explanatory variable are computed as follows:

$$\text{Parameter estimate} \pm \xi_{0.975}\hat{\sigma}_j.$$

Here ξ_α is the α-quantile of the standard normal distribution and $\hat{\sigma}_j$ is the standard error of the j-th explanatory variable. The values in the fifth column are the values of Wald's χ^2-statistic, which is defined as

$$\chi^2 = \left(\frac{\hat{\beta}_j}{\hat{\sigma}_j}\right)^2.$$

This statistic follows a χ_1^2 distribution. The last column of *Table 3.2* gives the p-values corresponding to Wald's χ^2, i.e. the smallest possible value of the confidence level α at which the null hypothesis that the value of the parameter is equal to zero would be rejected for the derived value of χ^2. With other words, the probability that the null hypothesis is falsely rejected is smaller than the p-value. All parameters are hence statistically significant for all confidence levels higher than 0.0001.

It is interesting to study the influence of the individual variables on the model. This can be done in SAS with a Type 1 or a Type 3 analysis, which generate statistical tests for the significance of these influences. A Type 1 analysis involves fitting a sequence of models, starting with the most simple model containing only the intercept. In each step a variable is added to the model. For every two successive models the difference of the log likelihoods times two is computed, which is equal to the difference of the scaled deviances if φ is held fixed for all models and hence to the difference of the deviances in case of the Poisson distribution. In the *Appendix* we remark that this difference is χ_1^2 distributed, under the null hypothesis that the parameter of the added variable is equal to 0. So if the p-value for this parameter value is smaller than α, then the null hypothesis can be rejected and the added variable is statistically significant for confidence level α. The results of a Type 1 analysis depend on the order in which the variables are added to the model. In *Table 3.3* the results of the Type 1 analysis are given. From the Type 1 analysis it follows that both explanatory variables are statistically significant for all confidence levels higher than 0.0001.

Source	Scaled deviance (SD)	Difference between SD's	p-value
Intercept	8337.5883		
log(L)	1331.8325	7005.76	$<$ 0.0001
log$(AADT)$	1247.9565	83.88	$<$ 0.0001

Table 3.3. *Statistics for the Type 1 analysis of the simple model for urban carriageways, based on the Poisson distribution.*

A Type 3 analysis computes the likelihood ratio statistic for each variable $\mathbf{x}_j$, that is two times the difference between the log likelihood for the model containing all variables and the log likelihood for the model with all variables except $\mathbf{x}_j$. The likelihood ratio statistic follows a χ_1^2 distribution under the hypothesis that the parameter of $\mathbf{x}_j$ is equal to zero. The results of the Type 3 analysis are given in *Table 3.4*. The Type 3 analysis leads to the same conclusion as the Type 1 analysis.

Source	Difference between scaled deviances	p-value
log(L)	7048.64	$<$ 0.0001
log$(AADT)$	83.88	$<$ 0.0001

Table 3.4. *Statistics for the Type 3 analysis of the simple model for urban carriageways, based on the Poisson distribution.*

All the statistics described above were used to check the validity of the model. For this purpose also three types of plot are very useful. In the first type the standardized deviance residuals are plotted against the explanatory variables in the linear predictor, see *Figures 3.1* and *3.2*. The null pattern of this type of plot is a distribution of residuals with mean zero and constant range. Both plots show a zero mean, but according to *Figure 3.2* there is no constant range. This indicates heteroscedasticity.

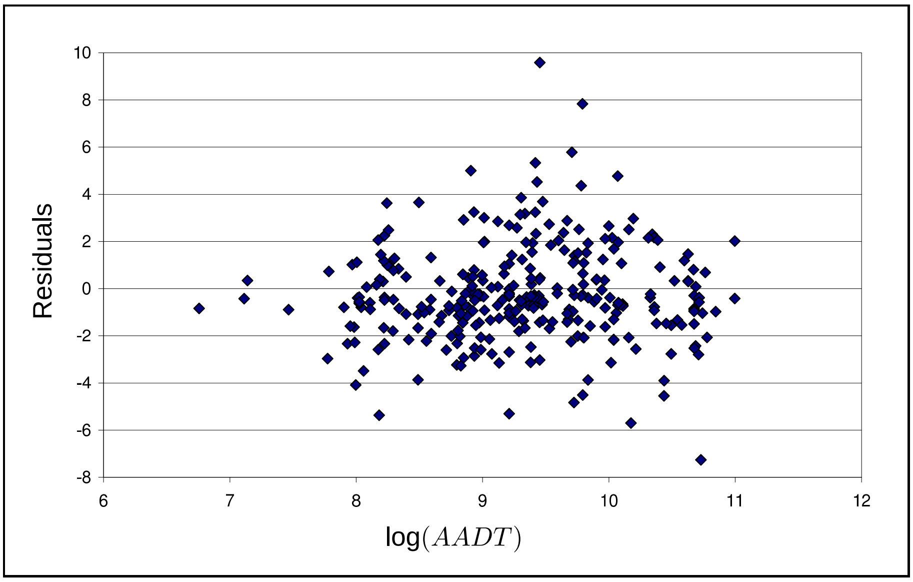

Figure 3.1. *The standardized deviance residuals of the simple model for urban carriageways, based on the Poisson distribution against log$(AADT)$.*

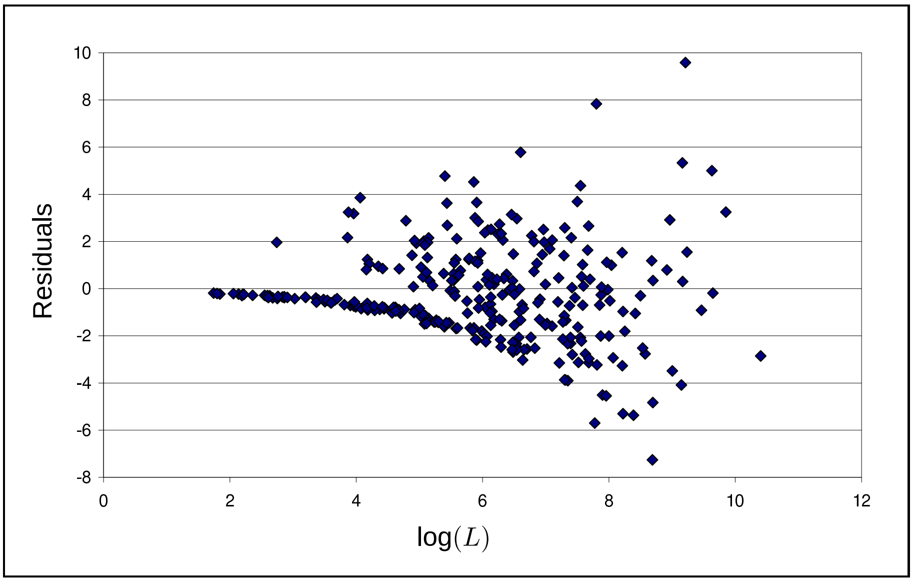

Figure 3.2. *The standardized deviance residuals of the simple model for urban carriageways, based on the Poisson distribution against log(L).*

The second plot type is a plot of the standardized deviance residuals against the linear predictor, see *Figure 3.3*. The null pattern of this plot is the same as for the previous type, so the residuals should be scattered around the x-axis with constant range. In addition, the contours of fixed y (the observed values) should be 'parallel' curves. In *Figure 3.3* the curves are more or less visible, but the constant range condition is violated.

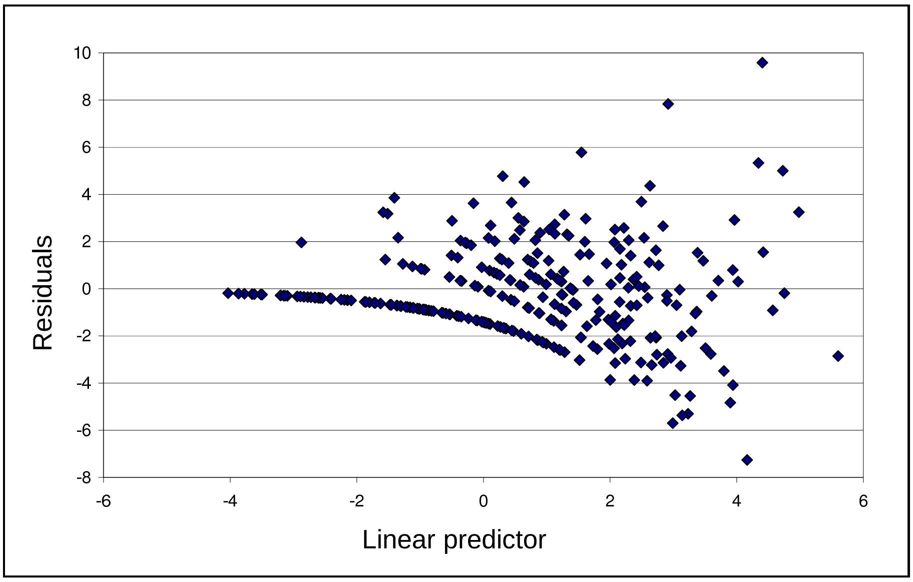

Figure 3.3. *The standardized deviance residuals of the simple model for urban carriageways, based on the Poisson distribution against the linear predictor.*

The third plot type is the so-called QQ-plot. This plot displays the following points:

$$\left\{\left(\Phi^{-1}\left(i/304\right), DR_{(i)}\right) : i = 1, \ldots, 303\right\},$$

where Φ is the distribution function of the standard normal distribution and $DR_{(i)}$ is the i-th order statistic of the standardized deviance residuals, i.e., the i-th standardized deviance residual when they are ordered in increasing order. These points should show a scatter around a straight line with slope 1.

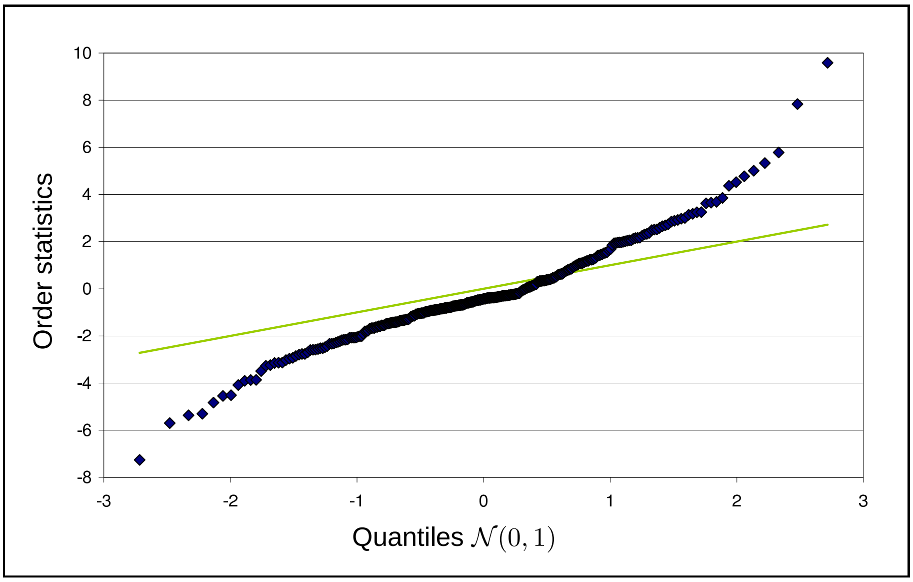

Figure 3.4. *The QQ-plot for the standardized deviance residuals of the simple model for urban carriageways, based on the Poisson distribution.*

At first sight, the points in the QQ-plot of the fitted model (*Figure 3.4*) form a straight line reasonably well, although not with slope 1. However, drawing a straight line through them shows that the points are more like a curve than like a straight line. Therefore the conclusion can be drawn that the residuals are certainly not standard normally distributed, which means that the conclusions based on the statistics in *Tables 3.2 – 3.4* are questionable.

3.1.2. *The negative binomial distribution*

In this section the model which is obtained under the assumption that road crashes are negative binomially distributed will be discussed. The statistics in *Table 3.5* describe the goodness-of-fit of the model. If the deviance is compared to its χ^2_{300} distribution a p-value of 0.24 is found. This implies that the null hypothesis that the fitted model is the right model can not be rejected on basis of all confidence levels greater than 0.24. A similar but less convincing result follows from Pearson's χ^2; its p-value is 0.07.

Criterion	Degrees of freedom (DF)	Value	Value/DF
Deviance	300	316.8576	1.0562
Pearson's χ^2	300	337.4371	1.1248
Log likelihood		7349.8350	

Table 3.5. *Criteria for assessing the goodness-of-fit of the simple model for urban carriageways, based on the negative binomial distribution.*

The estimates for the model parameters and several statistics are given in *Table 3.6*. Also $1/\nu$,the scale parameter of the Gamma distribution, is estimated.

Parameter	Estimate	Standard error	Wald's 95% confidence interval	Wald's χ^2	p-value
Intercept	-8.0049	0.8619	(-9.6942, -6.3155)	86.25	$<$ 0.0001
log(L)	0.9988	0.0415	(0.9174, 1.0801)	578.43	$<$ 0.0001
log$(AADT)$	0.3181	0.0815	(0.1584, 0.4778)	15.24	$<$ 0.0001
$1/\nu$	0.5821	0.0828	(0.4199, 0.7443)		

Table 3.6. *Analysis of the parameter estimates for the simple model for urban carriageways, based on the negative binomial distribution.*

The relation between the expected number of crashes on urban carriageways in three years, the AADT and carriageway length is given by

$$\hat{\mu}_i = e^{-8.0049} \cdot AADT_i^{0.3181} \cdot L_i^{0.9988} = 0.00033 \cdot AADT_i^{0.3181} \cdot L_i^{0.9988}.$$

Once more, the exponent of L is almost equal to 1. This value is even an element of the 95%-confidence interval corresponding to log(L). Although the estimates are not very different from those in *Table 3.2*, the standard errors are a factor 2.5 to 2.8 larger. However, the variables still are statistically significant for all confidence levels higher than 0.0001.

The results of the Type 1 and 3 analyses are stated in the *Tables 3.7* and *3.8*. Both analyses indicate that the carriageway length and AADT are statistically significant for all confidence levels higher than 0.0001.

Source	Twice the log likelihood	Difference of scaled deviances	p-value
Intercept	14320.4013		
log(L)	14684.6662	364.26	$<$ 0.0001
log$(AADT)$	14699.6700	15.00	0.0001

Table 3.7. *Statistics for the Type 1 analysis of the simple model for urban carriageways, based on the negative binomial distribution.*

Source	Difference of scaled deviances	p-value
log(L)	377.14	$<$ 0.0001
log$(AADT)$	15.00	0.0001

Table 3.8. *Statistics for the Type 3 analysis of the simple model for urban carriageways, based on the negative binomial distribution.*

Again several plots involving the standardized deviance residuals were drawn, see *Figures 3.5 – 3.8*. These scatter plots look better than the scatter plots corresponding to the Poisson distribution. Indeed, the variance in *Figures 3.6* and *3.7* is smaller than in *Figures 3.2* and *3.3*. Furthermore, the scatter plot in *Figure 3.8* closely resembles a straight line with slope 1, although the ends tend to deviate from that line.

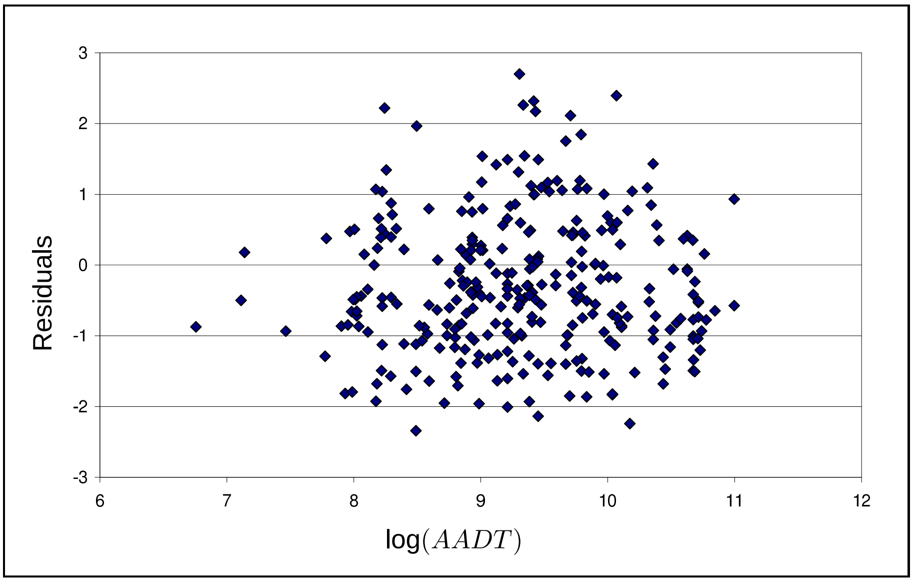

Figure 3.5. *The standardized deviance residuals of the simple model for urban carriageways, based on the negative binomial distribution against log$(AADT)$.*

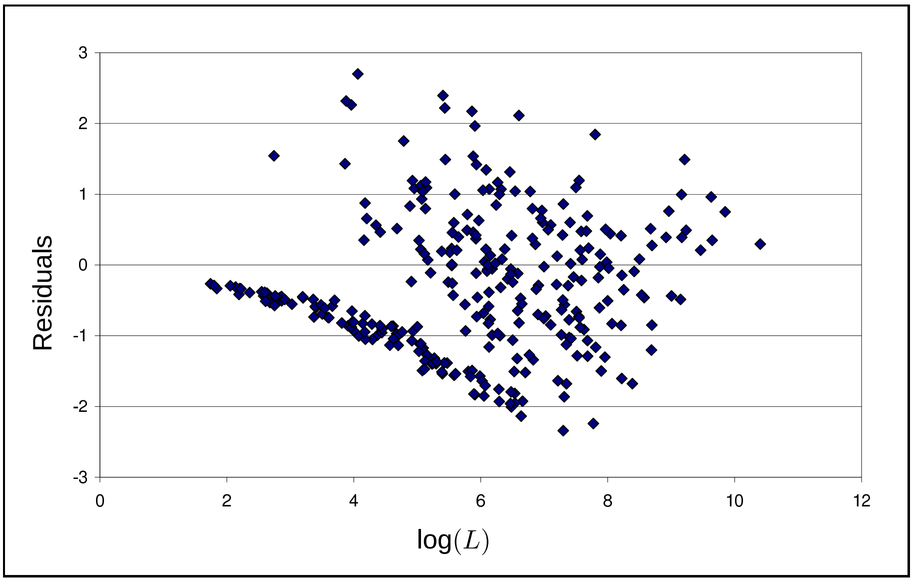

Figure 3.6. *The standardized deviance residuals of the simple model for urban carriageways, based on the negative binomial distribution against log(L).*

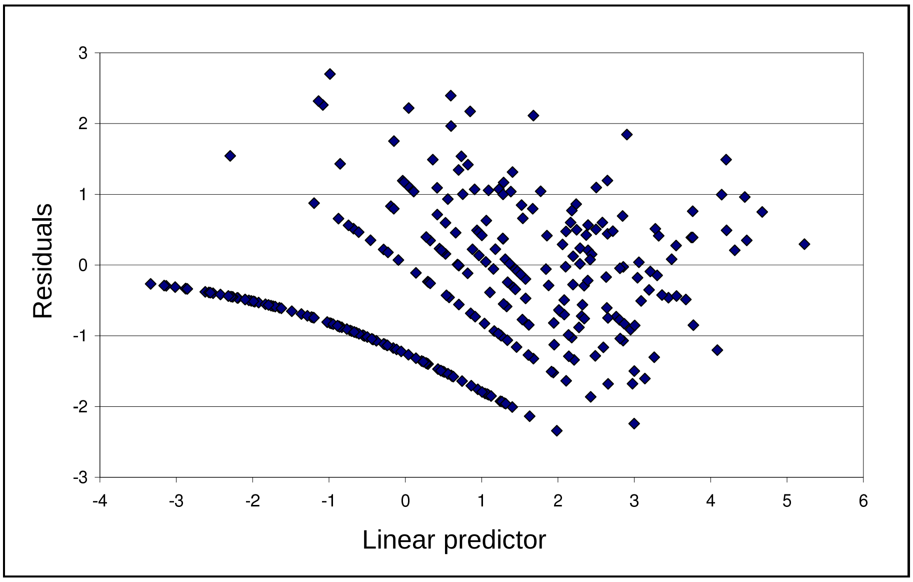

Figure 3.7. *The standardized deviance residuals of the simple model for urban carriageways, based on the negative binomial distribution against the linear predictor.*

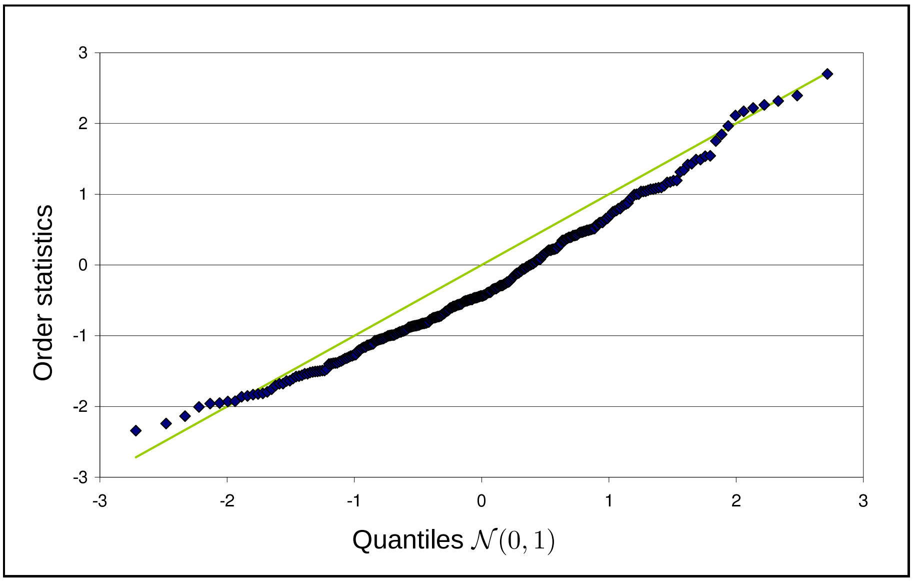

Figure 3.8. *The QQ-plot for the standardized deviance residuals of the simple model for urban carriageways, based on the negative binomial distribution.*

3.1.3. *The quasi-likelihood method*

The subject of this section is the model for urban carriageways obtained by applying the quasi-likelihood method. As explained in the *Appendix* this method does not require any assumptions about the underlying distribution, but only needs an assumption about the variance. In this case this assumption is that the variance of Y_i is given by $\mathbb{V}\text{ar}(Y_i) = \sigma^2\mu_i$, where σ^2 is possibly unknown. The parameter σ^2 can be estimated with the deviance or Pearson's χ^2 divided by the number of degrees of freedom. Here the first possibility is chosen, resulting in a scaled deviance equal to 1, see *Table 3.9*. The deviance and Pearson's χ^2 are the same as for the model based on the Poisson distribution. However, they are not equal to their scaled versions anymore, because φ is now not equal to 1.

Criterion	Degrees of freedom (DF)	Value	Value/DF
Deviance	300	1247.9565	4.1599
Scaled deviance	300	300.0000	1.0000
Pearson's χ^2	300	1459.4956	4.8650
Scaled Pearson's χ^2	300	350.8525	1.1695
Log likelihood		1693.4694	

Table 3.9. *Criteria for assessing the goodness-of-fit of the simple model for urban carriageways developed using the quasi-likelihood method.*

The estimated values for the intercept and the parameters of the two explanatory variables are equal to the estimated values under the assumption that the number of road crashes is Poisson distributed, but the values of the corresponding statistics are different. These values are given in the *Table 3.10*. It is easy to check that the standard errors are indeed a factor σ larger than in the case of the Poisson distribution. Consequently, Wald's 95%-confidence intervals are slightly wider as its bounds are given by

$$\text{Parameter estimate} \pm \xi_{0.975}\hat{\sigma}_j.$$

Finally, Wald's χ^2 is a factor σ^2 smaller, which follows immediately from its definition and the fact that the standard errors are a factor σ larger.

Parameter	Estimate	Standard error	Wald's 95% confidence interval	Wald's χ^2	p-value
Intercept	-8.4758	0.6958	(-9.8395, -7.1121)	148.40	< 0.0001
log(L)	1.1210	0.0318	(1.0588, 1.1833)	1245.28	< 0.0001
log$(AADT)$	0.2703	0.0604	(0.1520, 0.3886)	20.05	< 0.0001
$\sigma = \sqrt{\varphi}$	2.0396	0.0000	(2.0396, 2.0396)		

Table 3.10. *Analysis of the parameter estimates for the simple model for urban carriageways developed using the quasi-likelihood method.*

The Type 1 and Type 3 analyses also show that the parameters corresponding to the logarithms of the AADT and carriageway length are statistically significant for all confidence levels higher than 0.0001.

Source	Deviance	Difference of SDs	p-value
Intercept	8337.5883		
log(L)	1331.8325	1684.13	< 0.0001
log$(AADT)$	1247.9565	20.16	< 0.0001

Table 3.11. *Statistics for the Type 1 analysis of the simple model for urban carriageways developed using the quasi-likelihood method.*

Source	Difference of scaled deviances	p-value
log(L)	1694.44	$<$ 0.0001
log$(AADT)$	20.16	$<$ 0.0001

Table 3.12. *Statistics for the Type 3 analysis of the simple model for urban carriageways developed using the quasi-likelihood method.*

Also in this case it is possible to give the plots of the standardized deviance residuals against the explanatory variables and against the linear predictor. However, from *(A.3)* it follows that the standardized deviance residuals are a factor σ smaller, because $\hat{\phi}$ is now equal to σ^2 instead of equal to 1. Hence the plots of the residuals against the explanatory variables and the linear predictor are similar to *Figures 3.1 – 3.3*. Due to the decrease of the standardized deviance residuals their QQ-plot is different than *Figure 3.4*: the points form an approximately straight line with slope 1, see *Figure 3.9*.

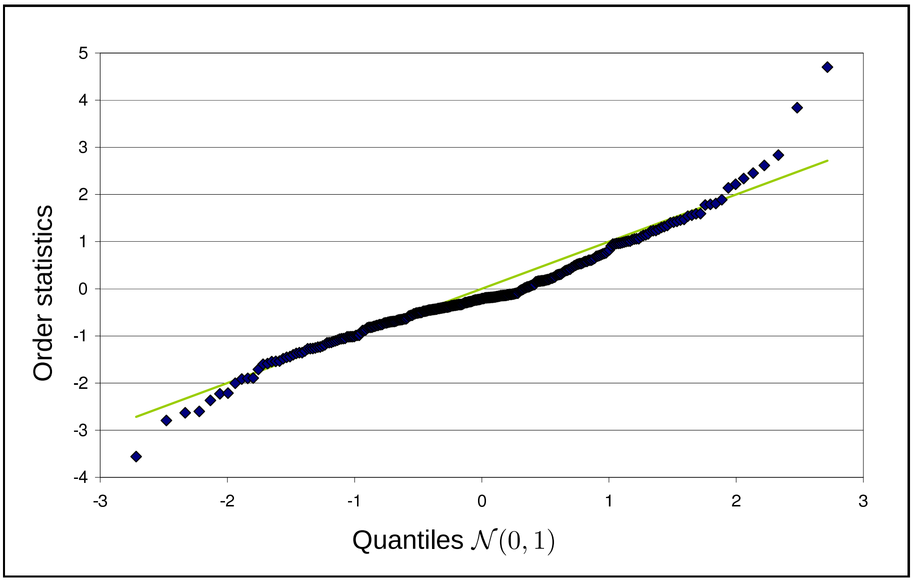

Figure 3.9. *The QQ-plot for the standardized deviance residuals of the simple model for urban carriageways developed using the quasi-likelihood method.*

3.1.4. *Discussion*

In *Sections 3.1.1 – 3.1.3* two different models are derived which describe the relation between the number of road crashes on urban carriageways in three years, the AADT and carriageway length. These models are

$$\hat{\mu}_i = 0.00021 \cdot AADT_i^{0.2703} \cdot L_i^{1.1210}, \qquad (3.1)$$

$$\hat{\mu}_i = 0.00033 \cdot AADT_i^{0.3181} \cdot L_i^{0.9988}. \qquad (3.2)$$

Model *(3.1)* was derived in two different ways: 1) by assuming that the number of road crashes follows a Poisson distribution, and 2) by applying the quasi-likelihood method. The quasi-likelihood method is preferred, because then the model is not affected anymore by overdispersion. However, the scatter plots of the standardized deviance residuals indicate that these residuals are not normally distributed with constant variance. So the conclusions based on several statistical tests are still doubtful. Under the assumption that the number of road crashes is negative binomially distributed, model *(3.2)* was obtained. This model has two advantages: the problem of overdispersion is solved and there is no reason to believe that the standardized deviance residuals are not standard normally distributed with constant variance.

In both models the exponent of L_i is almost equal to 1. For *(3.2)* the confidence interval for the parameter of log(L) even contains 1. This implies that the expected number of road crashes in three years per metre on the i-th carriageway, given by $\hat{\mu}_i/L_i$, depends almost only on the AADT:

$$\frac{\hat{\mu}_i}{L_i} \approx \begin{cases} 0.00021 \cdot AADT_i^{0.2703}, & \text{for the Poisson model,} \\ 0.00033 \cdot AADT_i^{0.3181}, & \text{for the neg. bin. model.} \end{cases} \tag{3.3}$$

This shows that the expected number of road crashes per kilometre per year, denoted by τ_i, is approximately given by

$$\tau_i \approx \begin{cases} 0.07 \cdot AADT_i^{0.2703}, & \text{for the Poisson model,} \\ 0.11 \cdot AADT_i^{0.3181}, & \text{for the neg. bin. model.} \end{cases} \tag{3.4}$$

In *Figure 3.10* the predicted number of road crashes per kilometre per year, as given in *(3.4)*, is plotted against the AADT for the Poisson based and the negative binomial based model. It follows that the negative binomial model generally gives a higher risk than the Poisson model. The different shapes of the two plots is explained by the fact that the exponent of L_i in model *(3.2)* is much closer to 1 than the one in model *(3.1)*.

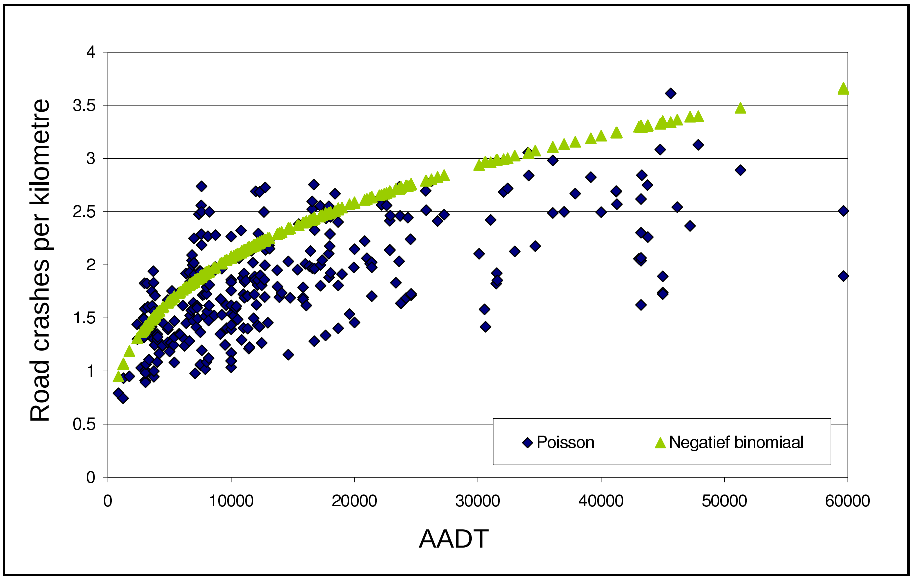

Figure 3.10. *The predicted number of road crashes per kilometre per year against the AADT for urban carriageways.*

It is possible to develop a model such that $\hat{\mu}_i/L_i$ (and hence τ_i) does not depend on the carriageway length at all, namely by defining log(L) as an offset variable. An offset variable is a variable whose parameter is set equal to 1. If $\log(L)$ is taken as an offset variable, then the resulting models for τ_i

are

$$\tau_i = \begin{cases} 0.38 \cdot AADT_i^{0.1960}, & \text{for the Poisson model,} \\ 0.11 \cdot AADT_i^{0.3186}, & \text{for the neg. bin. model.} \end{cases}$$

For the model based on the negative binomial distribution almost nothing has changed compared to *(3.3)*. The Poisson based model, however, did change considerably. The exponent of $AADT$ even lies outside the confidence interval given in *Table 3.2*. The new models are plotted in *Figure 3.11.* This plot shows that the crash rates predicted by both models do not differ much.

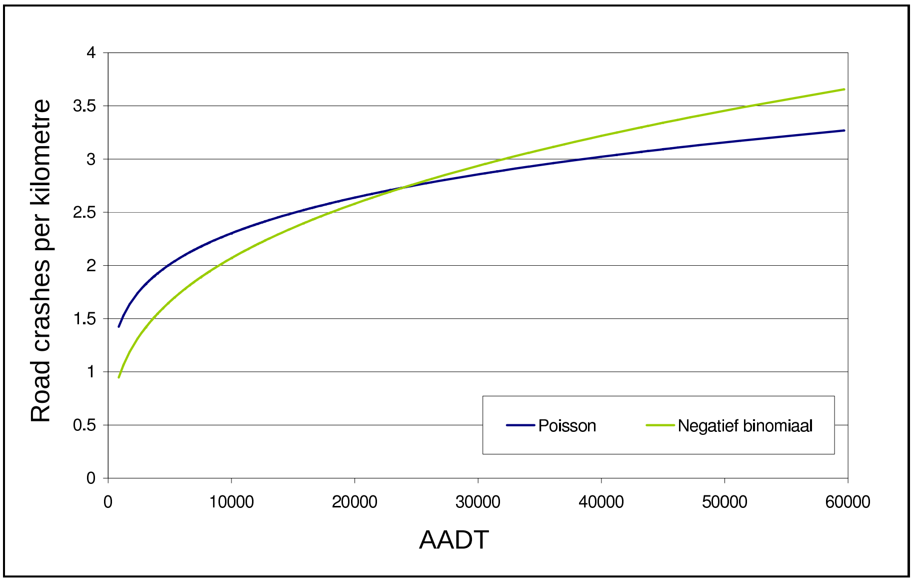

Figure 3.11. *The predicted number of road crashes per kilometre per year against the AADT for urban carriageways with log(L) as an offset variable.*

3.2. Rural carriageways

3.2.1. *The Poisson distribution*

The goodness-of-fit of the model for rural carriageways under the assumption that the number of road crashes is Poisson distributed is described in *Table 3.13*. The deviance and Pearson's χ^2 are approximately twice as large as the number of degrees of freedom. This indicates the presence of overdispersion.

Criterion	Degrees of freedom (DF)	Value	Value/DF
Deviance	95	185.9737	1.9576
Pearson's χ^2	95	191.6960	2.0179
Log likelihood		305.0167	

Table 3.13. *Criteria for assessing the goodness-of-fit of the simple model for rural carriageways, based on the Poisson distribution.*

The parameter estimates and several statistics are given in *Table 3.14*. The relation between the expected number of road crashes in three years, the carriageway length in metres and the AADT is therefore given by

$$\hat{\mu}_i = e^{-8.1194} \cdot AADT_i^{0.3028} \cdot L_i^{0.9290} = 0.00030 \cdot AADT_i^{0.3028} \cdot L_i^{0.9290}.$$

The confidence interval corresponding to log(L) includes 1, so also for this model the exponent of L is close to 1. The parameter corresponding to log(L) is statistically significant for all confidence levels higher than 0.0001. The parameter corresponding to log$(AADT)$ is only statistically significant for all confidence levels higher than 0.0093, which is a less convincing significance. The standard errors are much larger than those of the parameters of the Poisson model for urban carriageways. This could be a consequence of the lower number of available carriageways.

Parameter	Estimate	Standard error	Wald's 95% confidence interval	Wald's χ^2	p-value
Intercept	-8.1194	1.1903	(-10.4524, -5.7864)	46.53	$<$ 0.0001
log(L)	0.9290	0.0479	(0.8351, 1.0230)	375.57	$<$ 0.0001
log$(AADT)$	0.3028	0.1165	(0.0745, 0.5311)	6.76	0.0093

Table 3.14. *Analysis of the parameter estimates for the simple model for rural carriageways, based on the Poisson distribution.*

The results of the Type 1 and Type 3 analyses are summarized in *Table 3.15* and *3.16*. From the data in these tables it also follows that the parameter corresponding to log$(AADT)$ is statistically significant with far less confidence than the parameter corresponding to $\log(L)$.

Source	Scaled deviance	Difference of SD's	p-value
Intercept	678.6608		
log(L)	192.8836	485.78	$<$ 0.0001
log$(AADT)$	185.9737	6.91	0.0086

Table 3.15. *Statistics for the Type 1 analysis of the simple model for rural carriageways, based on the Poisson distribution.*

Source	Difference of scaled deviances	p-value
log(L)	483.26	$<$ 0.0001
log$(AADT)$	6.91	0.0086

Table 3.16. *Statistics for the Type 3 analysis of the simple model for rural carriageways, based on the Poisson distribution.*

The four standard graphs of the standardized deviance residuals are given in *Figures 3.12 – 3.15*. These plots show the same problems as the plots in *Section 3.1.1*. The plots of the residuals against log(L) and against the linear predictor do not have the desired pattern: there is no constant variance. Furthermore, the dots in the QQ-plot do not approximate a straight line, which leads to the conclusion that the residuals are not normally distributed. However, the dots are closer to a line with slope 1 than the dots in *Figure 3.4*. Hence from the plots it follows that the conclusions based on the performed statistical tests are questionable.

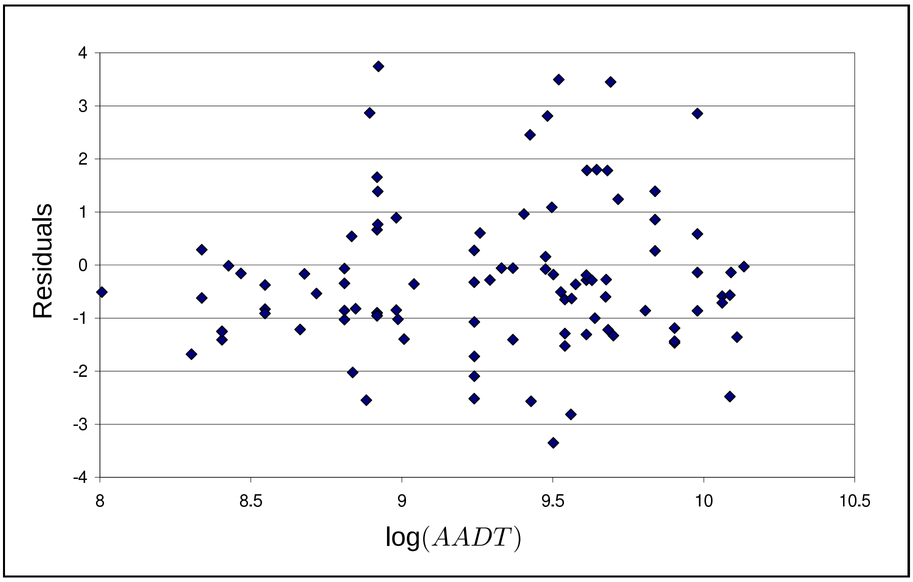

Figure 3.12. *The standardized deviance residuals of the simple model for rural carriageways, based on the Poisson distribution against log*$(AADT)$.

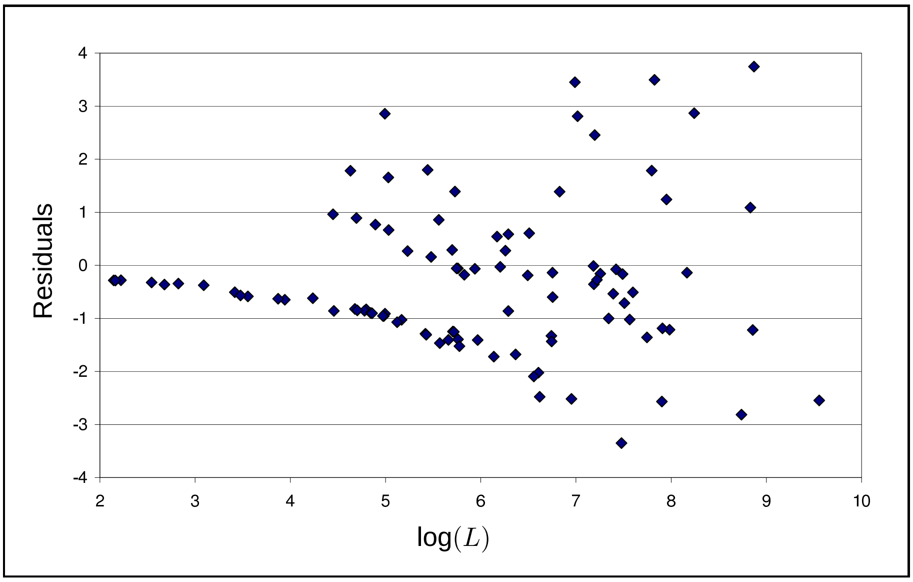

Figure 3.13. *The standardized deviance residuals of the simple model for rural carriageways, based on the Poisson distribution against log*(L).

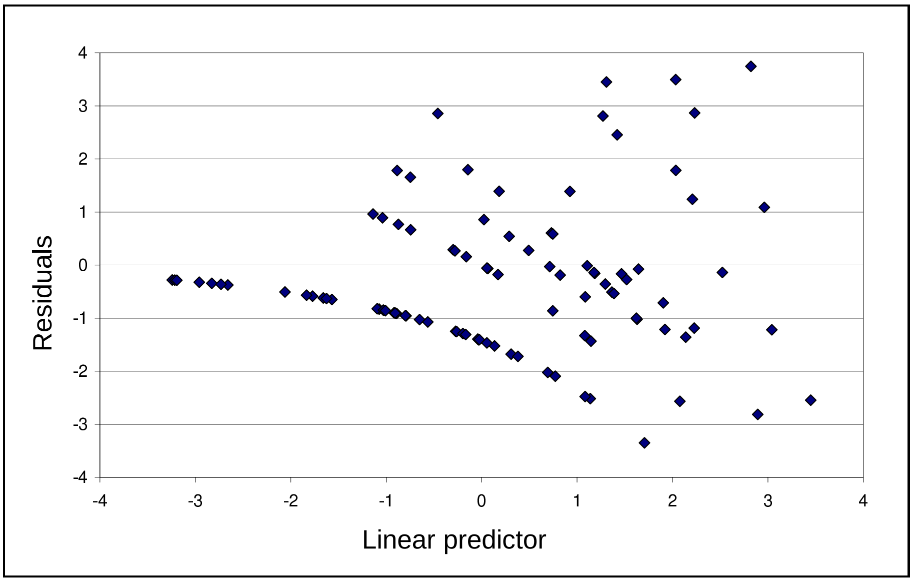

Figure 3.14. *The standardized deviance residuals of the simple model for rural carriageways, based on the Poisson distribution against the linear predictor.*

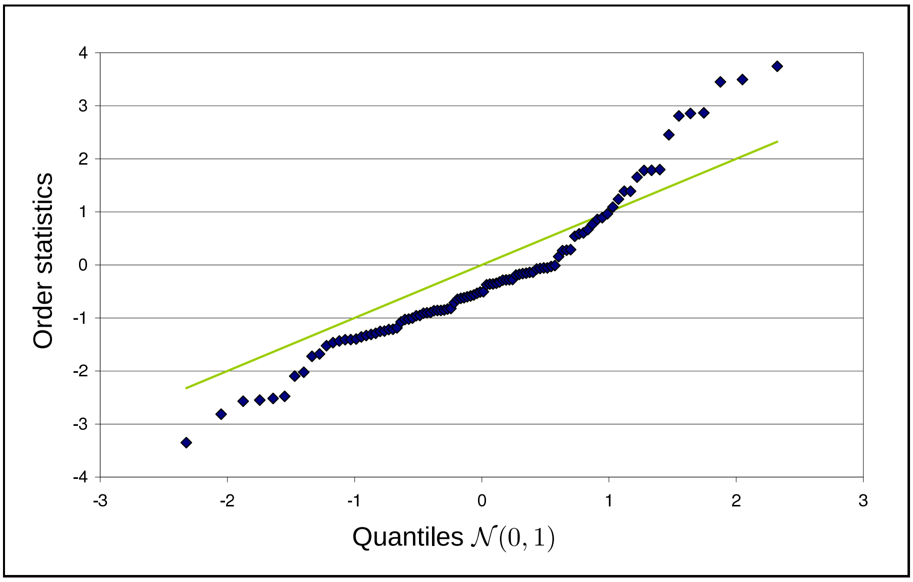

Figure 3.15. *The QQ-plot for the standardized deviance residuals of the simple model for rural carriageways, based on the Poisson distribution.*

3.2.2. *The negative binomial distribution*

The goodness-of-fit of the model based on the negative binomial distribution is described in *Table 3.17*. The deviance and Pearson's χ^2 are smaller than the number of degrees of freedom, which indicates the presence of under-dispersion. The underdispersion is however of a low level and hence not a problem.

Criterion	Degrees of freedom (DF)	Value	Value/DF
Deviance	95	94.1812	0.9914
Pearson's χ^2	95	93.1289	0.9803
Log likelihood		325.4619	

Table 3.17. *Criteria for assessing the goodness-of-fit of the simple model for rural carriageways, based on the negative binomial distribution.*

The parameter estimates and several statistics are given in *Table 3.18*. The relation between the expected number of road crashes on rural carriageways, the carriageway length and AADT in three years, is therefore

$$\hat{\mu}_i = e^{-10.1934} \cdot AADT_i^{0.4967} \cdot L_i^{0.9647} = 3.74 \cdot 10^{-5} \cdot AADT_i^{0.4967} \cdot L_i^{0.9647}.$$

The confidence interval corresponding to log(L) again contains 1. The parameter corresponding to log(L) is statistically significant with very high confidence. The parameter of log$(AADT)$ is only statistically significant for all confidence levels α such that $\alpha \geq 0.0155$.

Parameter	Estimate	Standard error	Wald's 95% confidence interval	Wald's χ^2	p-value
Intercept	-10.1934	2.0450	(-14.2016, -6.1853)	24.85	$<$ 0.0001
log(L)	0.9647	0.0826	(0.8027, 1.1266)	136.23	$<$ 0.0001
log$(AADT)$	0.4967	0.2053	(0.0944, 0.8989)	5.85	0.0155
$\frac{1}{\nu}$	0.3391	0.1190	(0.1058, 0.5723)		

Table 3.18. *Analysis of the parameter estimates for the simple model for rural carriageways, based on the negative binomial distribution.*

The results of the Type 1 and Type 3 analyses are given in *Tables 3.19* and *3.20*. The analyses indicate that the parameter corresponding to log$(AADT)$ is statistically significant for all $\alpha \geq 0.0135$.

Source	Twice the log likelihood	Difference of scaled deviances	p-value
Intercept	546.8178		
log(L)	644.8201	98.00	$<$ 0.0001
log$(AADT)$	650.9237	6.10	0.0135

Table 3.19. *Statistics for the Type 1 analysis of the simple model for rural carriageways, based on the negative binomial distribution.*

Source	Difference of scaled deviances	p-value
log(L)	102.91	$<$ 0.0001
log$(AADT)$	6.10	0.0135

Table 3.20. *Statistics for the Type 3 analysis of the simple model for rural carriageways, based on the negative binomial distribution.*

In *Figures 3.16 – 3.18* the standardized deviance residuals are plotted against the explanatory variables and the linear predictor. The first plot

(*Figure 3.16*) indicates a constant variance of the residuals. The second and third plot (*Figures 3.17* and *3.18*) on the other hand, still show an increasing variance. However, the heteroscedasticity is of a lower level than under the assumption that the number of road crashes is Poisson distributed.

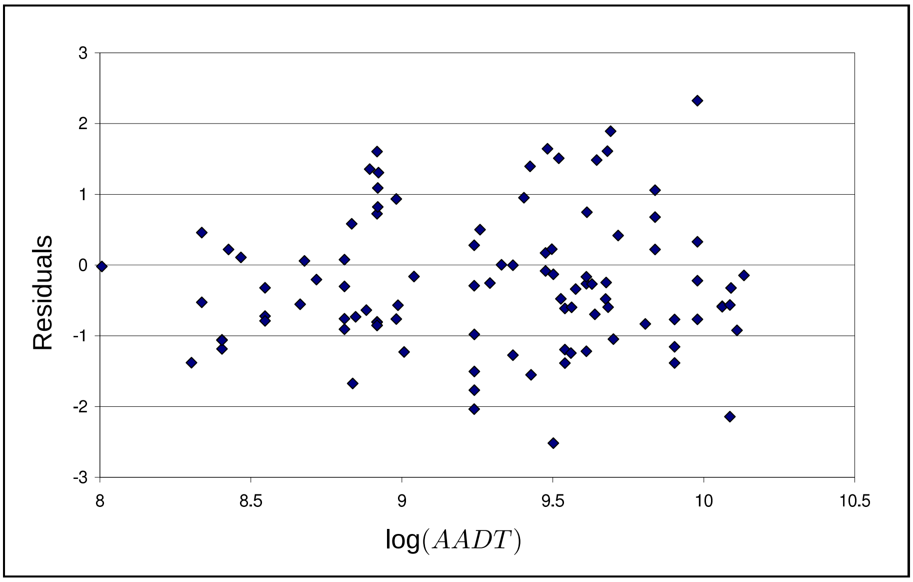

Figure 3.16. *The standardized deviance residuals of the simple model for rural carriageways, based on the negative binomial distribution against log*$(AADT)$.

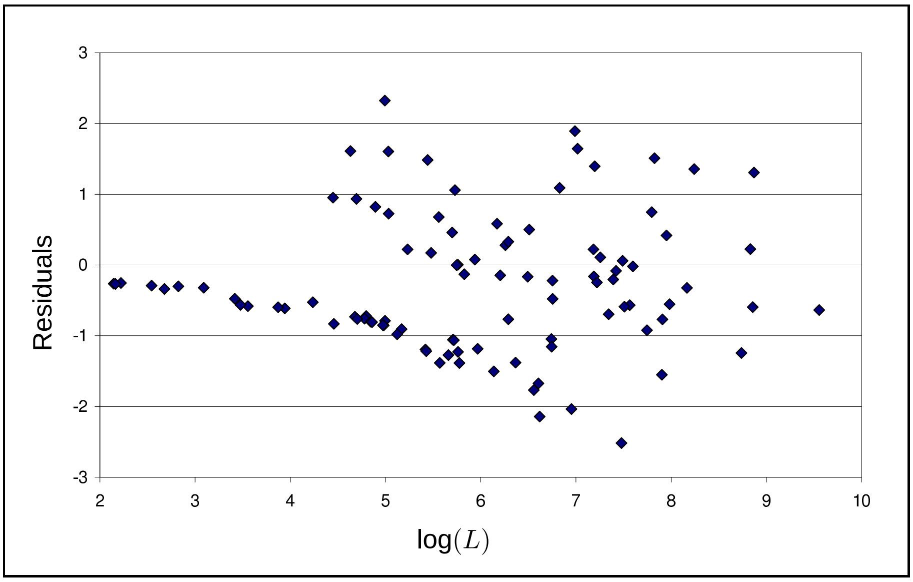

Figure 3.17. *The standardized deviance residuals of the simple model for rural carriageways, based on the negative binomial distribution against log*(L).

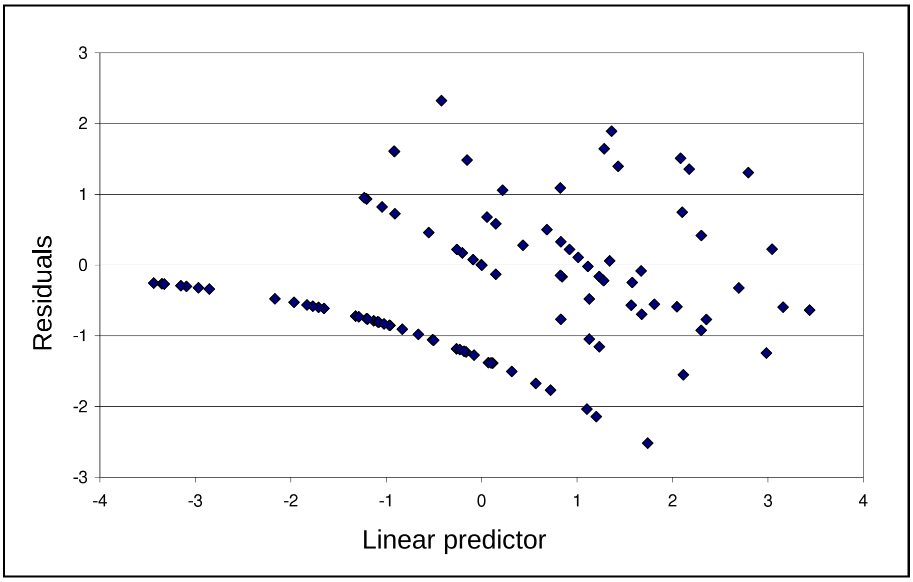

Figure 3.18. *The standardized deviance residuals of the simple model for rural carriageways, based on the negative binomial distribution against the linear predictor.*

The QQ-plot is given in *Figure 3.19*. It better resembles a straight line than the QQ-plot in *Figure 3.15* and there is no reason to believe that the standardized deviance residuals are not standard normally distributed.

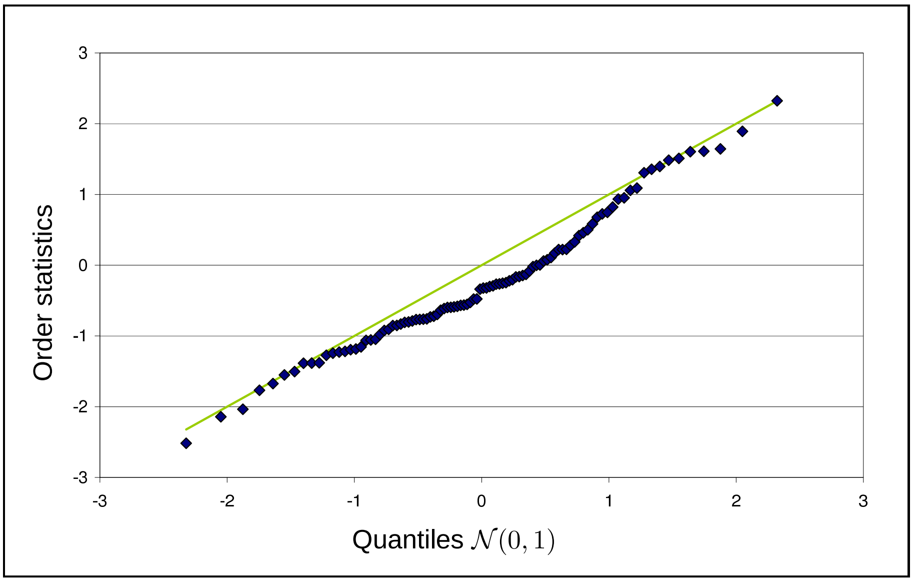

Figure 3.19. *The QQ-plot for the standardized deviance residuals of the simple model for rural carriageways, based on the negative binomial distribution.*

3.2.3. *The quasi-likelihood method*

The goodness-of-fit of the model developed with the quasi-likelihood method is described in *Table 3.21*. The quasi-likelihood parameter σ^2 is estimated by the deviance divided by the number of degrees of freedom. It follows that $\sigma^2 = \varphi = 1.9576.$

Criterion	Degrees of freedom (DF)	Value	Value/DF
Deviance	95	185.9737	1.9576
Scaled deviance	95	95.000	1.0000
Pearson's χ^2	95	191.6960	2.0179
Scaled Pearson's χ^2	95	97.9231	1.0308
Log likelihood		155.8101	

Table 3.21. *Criteria for assessing the goodness-of-fit of the simple model for rural carriageways developed using the quasi-likelihood method.*

The parameter estimates are the same as for the Poisson based model. They are stated in *Table 3.22*, together with several statistics. Because the standard errors have increased, the statistical significance of the parameters decreased. This is especially obvious for the parameter corresponding to the variable $\log(AADT)$. Its p-value increased from 0.0093 to 0.0632.

Parameter	Estimate	Standard error	Wald's 95% confidence interval	Wald's χ^2	p-value
Intercept	-8.1194	1.6654	(-11.3836, -4.8552)	23.77	$<$ 0.0001
log(L)	0.9290	0.0671	(0.7976, 1.0605)	191.85	$<$ 0.0001
log$(AADT)$	0.3028	0.1630	(-0.0167, 0.6222)	3.45	0.0632
$\sigma = \sqrt{\varphi}$	1.3991	0.0000	(1.3991, 1.3991)		

Table 3.22. *Analysis of the parameter estimates for the simple model for rural carriageways developed using the quasi-likelihood method.*

The results of the Type 1 and Type 3 analyses are summarized in *Tables 3.23* and *3.24*. They also show that the statistical significance of log$(AADT)$ decreased.

Source	Deviance	Difference of SD's	p-value
Intercept	678.6608		
log(L)	192.8836	248.15	$<$ 0.0001
log$(AADT)$	185.9737	3.53	0.0633

Table 3.23. *Statistics for the Type 1 analysis of the simple model for rural carriageways developed using the quasi-likelihood method.*

Source	Difference of scaled deviances	p-value
log(L)	246.86	$<$ 0.0001
log$(AADT)$	3.53	0.0633

Table 3.24. *Statistics for the Type 3 analysis of the simple model for rural carriageways developed using the quasi-likelihood method.*

In *Section 3.1.3* it was already stated that the QQ-plot for the standardized deviance residuals resulting from the quasi-likelihood method is different than for those following from the Poisson distribution. Therefore the QQ-plot is given in *Figure 3.20*.

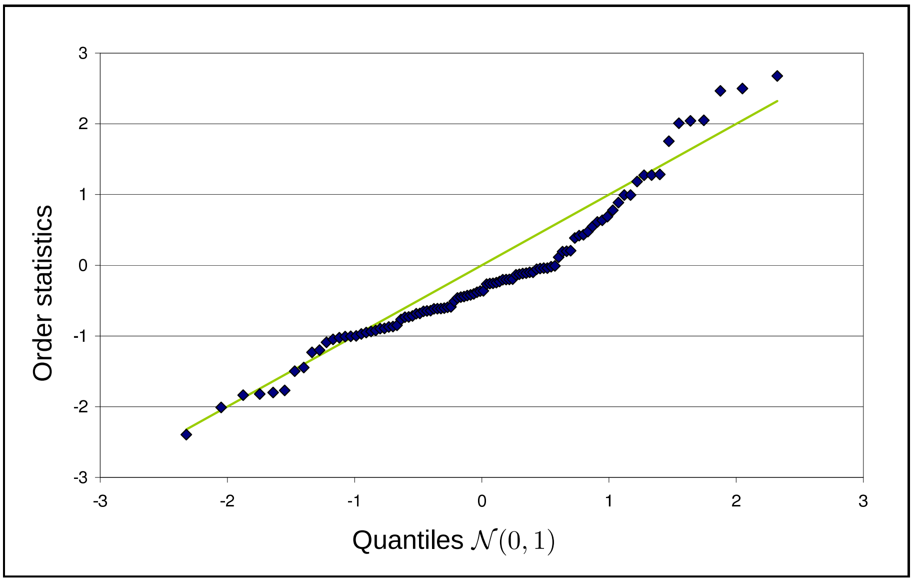

Figure 3.20. *The QQ-plot for the standardized deviance residuals of the simple model for rural carriageways developed using the quasi-likelihood method.*

3.2.4. *Discussion*

In *Sections 3.2.1 – 3.2.3* two different models are derived which describe the relation between the number of road crashes on urban carriageways in three years, the AADT and carriageway length. These models are

$$\hat{\mu}_i = 0.00030 \cdot AADT_i^{0.3028} \cdot L_i^{0.9290}, \quad (3.5)$$
$$\hat{\mu}_i = 3.74 \cdot 10^{-5} \cdot AADT_i^{0.4967} \cdot L_i^{0.9647}. \quad (3.6)$$

Model *(3.5)* was derived in two different ways: 1) by assuming that the number of road crashes follows a Poisson distribution and 2) by applying the quasi-likelihood method. As for urban carriageways the quasi-likelihood method is preferred, because it deals with overdispersion. However, the standardized deviance residuals do not follow a normal distribution with standard variance. There is no reason to believe that the residuals resulting from model *(3.6)*, which was obtained under the assumption that the number of road crashes is negative binomially distributed, are not normally distributed.

In both models the exponent of L is almost equal to 1, which was also the case in the models for urban carriageways. For all three modelling methods (based on Poisson, or negative binomial distribution, or the quasi-likelihood method) the 95%-confidence interval even contained 1. It follows that $\hat{\mu}_i/L_i$ depends almost only on the AADT:

$$\frac{\hat{\mu}_i}{L_i} \approx \begin{cases} 0.00030 \cdot AADT_i^{0.3028}, & \text{for Poisson model,} \\ 3.74 \cdot 10^{-5} \cdot AADT_i^{0.4967}, & \text{for negative binomial model.} \end{cases}$$

Hence τ_i, which stands for the number of crashes per kilometre per year, is approximately given by

$$\tau_i \approx \begin{cases} 0.10 \cdot AADT_i^{0.3028}, & \text{for Poisson model,} \\ 0.012 \cdot AADT_i^{0.4967}, & \text{for negative binomial model.} \end{cases}$$

In *Figure 3.21* τ_i is plotted against the AADT. It follows that the negative binomial model gives in general a lower risk for low AADT and a higher risk for high AADT than the Poisson model.

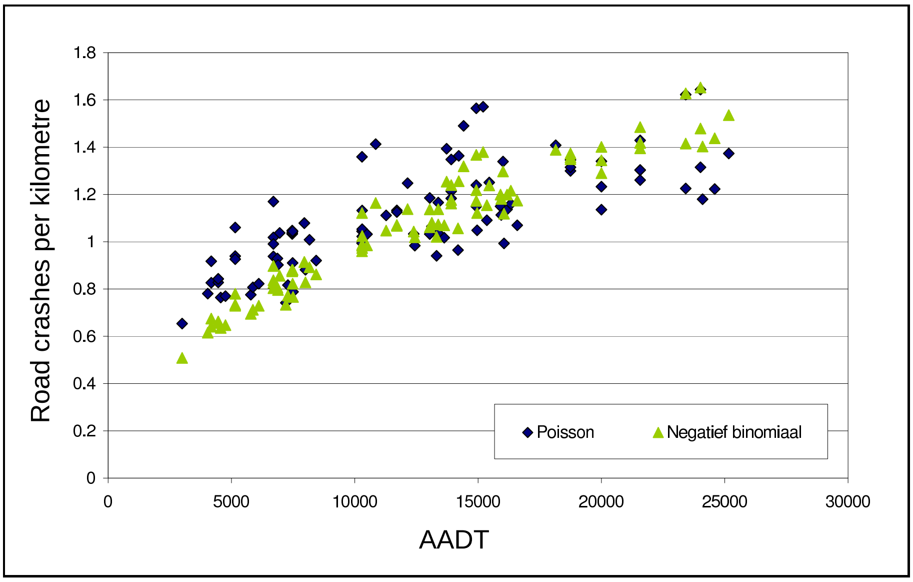

Figure 3.21. *The predicted number of road crashes per kilometre per year against the AADT for rural carriageways.*

In order to remove the dependency of $\hat{\mu}_i / L_i$ on L, $\log(L)$ is taken as an offset variable, which means that its coefficient is set equal to 1. The resulting models for τ_i are

$$\tau_i = \begin{cases} 0.047 \cdot AADT_i^{0.3223}, & \text{for the Poisson model,} \\ 0.009 \cdot AADT_i^{0.5029}, & \text{for the negative binomial model.} \end{cases}$$

These models are plotted in *Figure 3.22*. For lower AADTs both models are very close, but for higher AADTs the negative binomial model tends to predict a larger number of road crashes than the Poisson model.

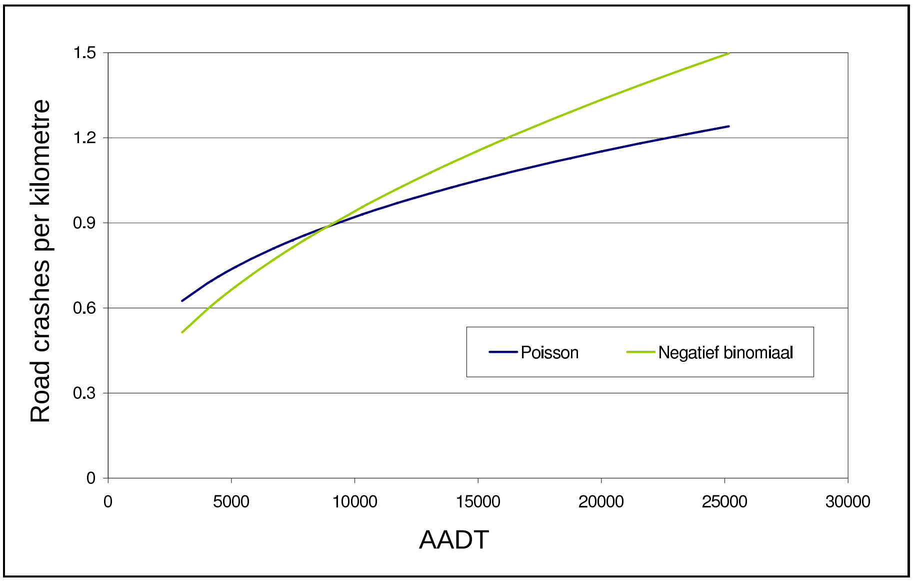

Figure 3.22. *The predicted number of road crashes per kilometre per year against the AADT for rural carriageways with log(L) as an offset variable.*

3.3. Comparison of the simple models

It is interesting to compare the models for urban carriageways to the models for rural carriageways. A first conclusion is that the derived models for urban carriageways are more reliable than the models for rural carriageways. This follows from the following two observations:

– The explanatory variables for all models for urban carriageways are statistically significant for all confidence levels higher than 0.0001. This is not the case for the models for rural carriageways. For those models, the variable log$(AADT)$ is only statistically significant for relatively high confidence levels.
– The standard errors of the parameter estimates for the models for rural carriageways are about a factor 2 to 4 higher than for the models for urban carriageways.

This is possibly a consequence of the number of available carriageways: the database contained three times more information about urban carriageways than about rural carriageways.

Secondly, for urban carriageways as well as for rural carriageways the exponent of L in the developed models is reasonable close to 1. For four of the six models 1 is even contained in the 95%-confidence interval corresponding to the variable $\log(L)$. Hence the number of crashes on urban and rural carriageways is approximately proportional to the carriageway length. By including log(L) in the model as an offset variable, its exponent is forced to be equal to 1.

The exponent of $AADT$ is different for the models for urban and rural carriageways. For the Poisson based model this exponent is 0.2703 for urban and 0.3028 for rural carriageways whereas for the negative binomial based model it is equal to 0.3181 for urban and to 0.4967 for rural carriageways. Therefore it can be concluded that the effect of the $AADT$ on the number of crashes is larger for rural carriageways than for urban carriageways. This difference in effect is especially clear for the negative binomial based model.

Finally, it is also possible to compare the modelled risk of urban and rural carriageways. For an easy comparison the obtained models for τ_i with log(L) as an offset variable are plotted in *Figure 3.23*. It follows that the modelled risk for urban carriageways is higher than the risk for rural carriageways for equal AADT.

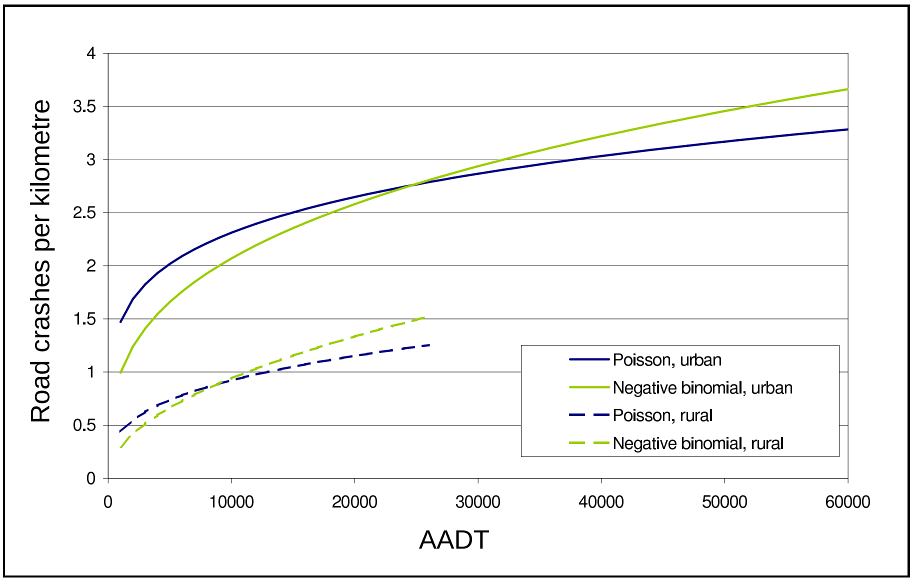

Figure 3.23. *The predicted number of road crashes per kilometre per year against the AADT.*

4. The extended model

In this chapter models of the type *(2.3)* will be discussed. Again separate models are modelled for urban carriageways (*Section 4.1*) and rural carriageways (*Section 4.2*), by using three different modelling techniques: Poisson based, negative binomial based and the quasi-likelihood method.

4.1. Urban carriageways

4.1.1. *The Poisson distribution*

The goodness-of-fit of the model based on the Poisson distribution is described in *Table 4.1*. The overdispersion is of a slightly lower level than for the Poisson based model for urban carriageways with two explanatory variables, see *Table 3.1*.

Criterion	Degrees of freedom (DF)	Value	Value/DF
Deviance	299	1132.5283	3.7877
Pearson's χ^2	299	1285.2341	4.2984
Log likelihood		7102.3012	

Table 4.1. *Criteria for assessing the goodness-of-fit of the extended model for urban carriageways, based on the Poisson distribution.*

Table 4.2 gives the parameter estimates together with several statistics. It follows that the predicted number of road crashes in three years on urban carriageways is given by:

$$\hat{\mu}_i = 4.5408 \cdot 10^{-7} \cdot L_i^{1.0915} \cdot AADT_i^{1.0406} \cdot e^{-0.0581 \cdot \frac{AADT_i}{1000}}. \qquad (4.1)$$

All variables are statistically significant for all confidence levels higher than 0.0001. Also for this model, the exponent of L is not very different from 1, although the confidence interval does not contain 1. The confidence interval corresponding to log$(AADT)$ does contain 1, but due to the presence of $AADT/1000$ this has no special meaning.

Parameter	Estimate	Standard error	Wald's 95% confidence interval	Wald's χ^2	p-value
Intercept	-14.6050	0.7085	(-15.9937, -13.2163)	424.90	$<$ 0.0001
log(L)	1.0915	0.0157	(1.0607, 1.1223)	4821.88	$<$ 0.0001
log$(AADT)$	1.0406	0.0822	(0.8795, 1.2018)	160.21	$<$ 0.0001
$AADT$/1000	-0.0581	0.0058	(-0.0695, -0.0466)	99.37	$<$ 0.0001

Table 4.2. *Analysis of the parameter estimates for the extended model for urban carriageways, based on the Poisson distribution.*

A Type 1 and Type 3 analysis are also conducted. The results are given in *Tables 4.3* and *4.4*. These results also lead to the conclusion that the variables are statistically significant for all confidence levels higher than 0.0001.

Source	Twice the log likelihood	χ^2	p-value
Intercept	8337.5883		
log(L)	1331.8325	7005.76	$<$ 0.0001
log$(AADT)$	1247.9565	83.88	$<$ 0.0001
$AADT$/1000	1132.5283	115.43	$<$ 0.0001

Table 4.3. *Statistics for the Type 1 analysis of the extended model for urban carriageways, based on the Poisson distribution.*

Source	χ^2	p-value
log(L)	6736.48	$<$ 0.0001
log$(AADT)$	183.15	$<$ 0.0001
$AADT/1000$	115.43	$<$ 0.0001

Table 4.4. *Statistics for the Type 3 analysis of the extended model for urban carriageways involving based on the Poisson distribution.*

The Type 1 analysis can also be used to decide whether or not the extended model is an improvement of the simple model. In *Section 3.1.1* it was explained that in a Type 1 analysis a sequence of models is fitted, starting with the model only containing the intercept. In each step an explanatory variable is added to the model. If the p-value of an added variable is smaller than a chosen confidence level $\alpha,$ then the null hypothesis that the parameter of this variable is equal to zero can be rejected. This means that the model with this additional variable is an improvement of the model without it.

In the Type 1 analysis for *(4.1)* first log(L) is added to the model only containing the intercept, then log$(AADT)$ and finally $AADT/1000.$
From *Table 4.3* it follows that the null hypothesis that the parameter of $AADT/1000$ is equal to zero can be rejected with high confidence. Indeed, the corresponding p-value is smaller than 0.0001. Hence, model *(4.1)* fits the data better than the model with only the intercept, log(L) and log$(AADT)$ as explanatory variables. This last model is exactly the simple model of *Section 3.1*, from which it follows that the extended model is better than the simple model.

Similar to *Chapter 3* the standardized deviance residuals will be studied by means of several plots. The plots of the standardized deviance residuals against the explanatory variables and the linear predictor are given in *Figures 4.1 – 4.4*. Specially *Figures 4.2* and *4.4* do not have the shape they should have: they show an increasing variance of the residuals. Although the shape is similar to the shape of the plots in *Figures 3.1 – 3.3*, it seems that the residuals corresponding to *(4.1)* are slightly smaller than those corresponding to the model discussed in *Section 3.1.1*.

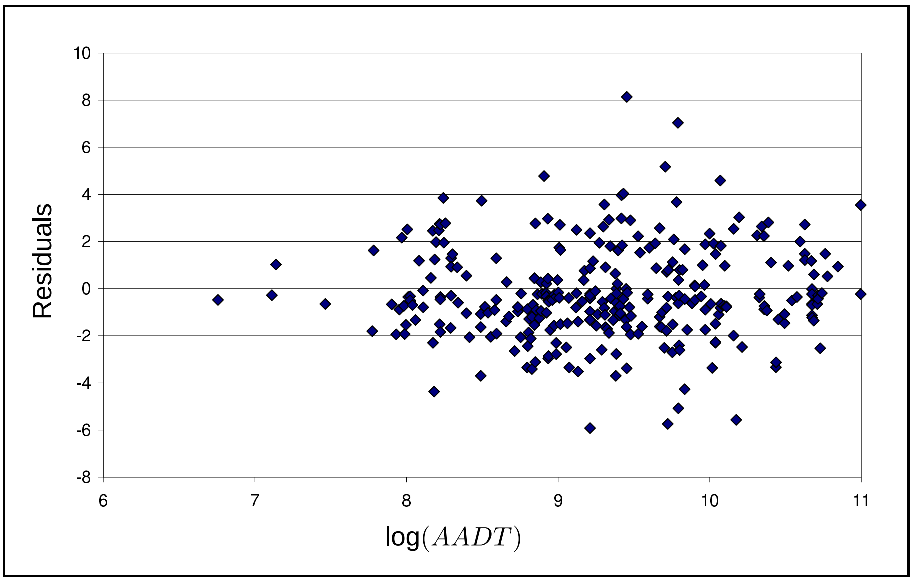

Figure 4.1. *The standardized deviance residuals of the extended model for urban carriageways, based on the Poisson distribution against log$(AADT)$.*

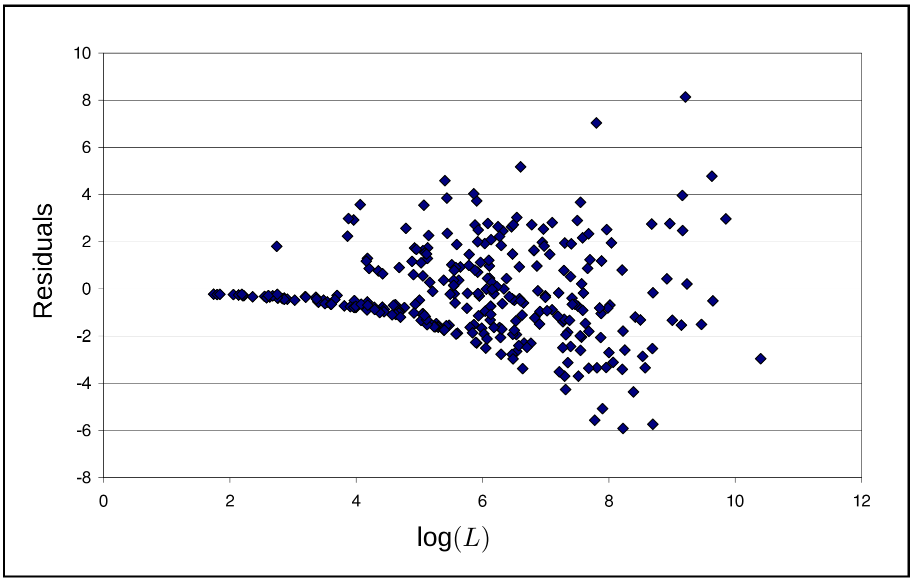

Figure 4.2. *The standardized deviance residuals of the extended model for urban carriageways, based on the Poisson distribution against log(L).*

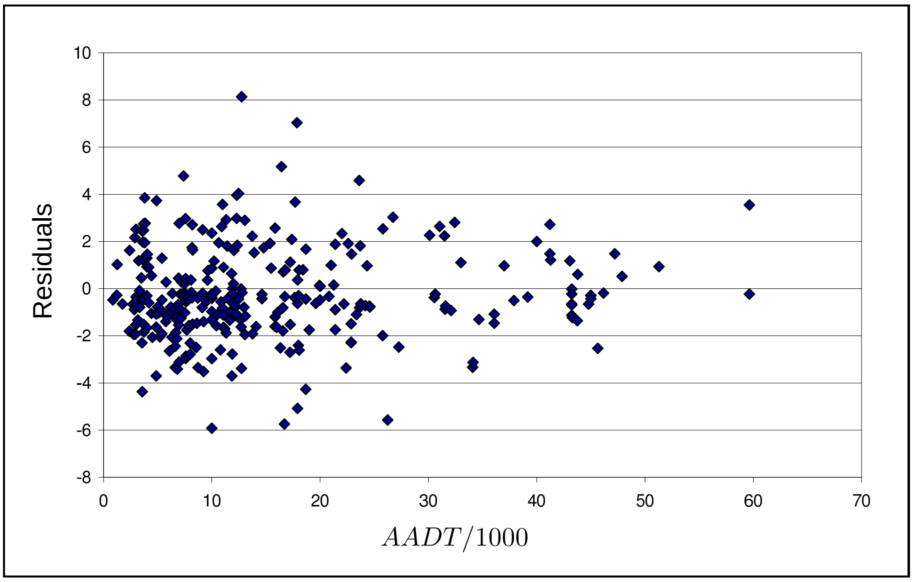

Figure 4.3. *The standardized deviance residuals of the extended model for urban carriageways, based on the Poisson distribution against* $AADT$.

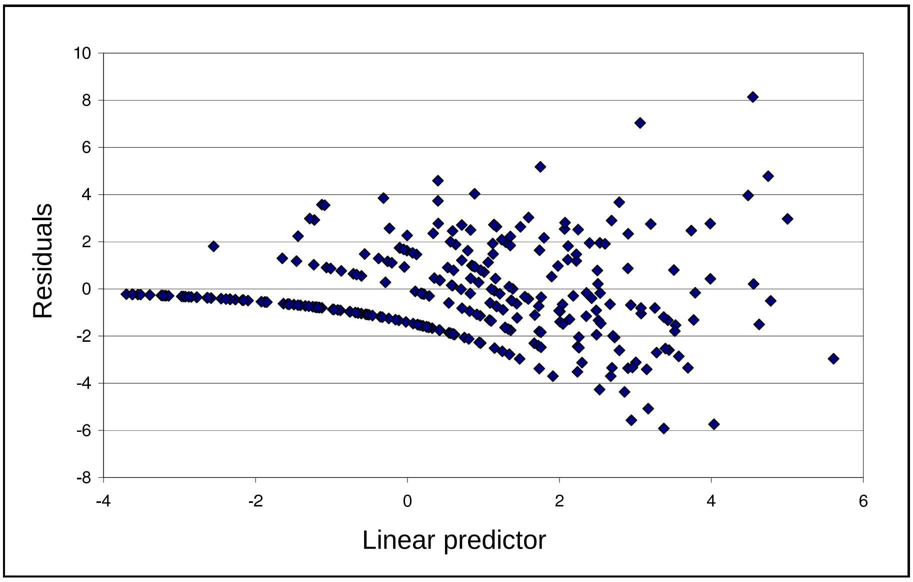

Figure 4.4. *The standardized deviance residuals of the extended model for urban carriageways, based on the Poisson distribution against the linear predictor.*

The QQ-plot is shown in *Figure 4.5*. Because the dots deviate from a straight line with slope 1, it cannot be concluded that the standardized deviance residuals are standard normally distributed.

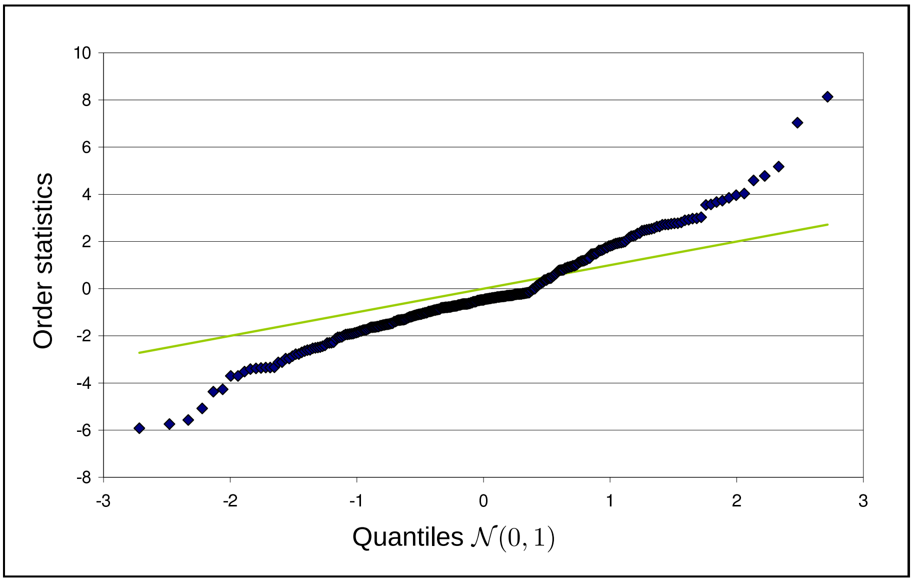

Figure 4.5. *The QQ-plot for the standardized deviance residuals of the extended model for urban carriageways, based on the Poisson distribution.*

4.1.2. *The negative binomial distribution*

The goodness-of-fit of the model based on the negative binomial distribution is described by the statistics in *Table 4.5*. It follows that overdispersion is not present in this model. Indeed, if the values of the deviance and Pearson's χ^2 are compared to their χ^2_{299} distribution, then p-values of 0.21 and 0.14 are found.

Criterion	Degrees of freedom (DF)	Value	Value/DF
Deviance	299	318.3506	1.0647
Pearson's χ^2	299	325.2484	1.0878
Log likelihood		7353.9790	

Table 4.5. *Criteria for assessing the goodness-of-fit of the extended model for urban carriageways, based on the negative binomial distribution.*

The parameter estimates and various statistics are given in *Table 4.6*. The model for the number of road crashes in three years on urban carriageways is

$$\hat{\mu}_i = 5.8880 \cdot 10^{-6} \cdot L_i^{0.9875} \cdot AADT_i^{0.8181} \cdot e^{-0.0375 \cdot \frac{AADT_i}{1000}}.$$

For this model 1 is contained in the confidence intervals corresponding to $\log(L)$. Further, the variables log(L) and log$(AADT)$ are statistically significant with higher confidence than the variable $AADT/1000$, but they are all three statistically significant for confidence level $\alpha = 0.0035$.

Parameter	Estimate	Standard error	Wald's 95% confidence interval	Wald's χ^2	p-value
Intercept	-12.0426	1.6318	(-15.2409, -8.8442)	54.46	$<$ 0.0001
log(L)	0.9875	0.0411	(0.9070, 1.0679)	578.51	$<$ 0.0001
log$(AADT)$	0.8181	0.1895	(0.4466, 1.1895)	18.63	$<$ 0.0001
$AADT$/1000	-0.0375	0.0129	(-0.0627, -0.0124)	8.53	0.0035
$\frac{1}{\nu}$	0.5463	0.0801	(0.3893, 0.7032)		

Table 4.6. *Analysis of the parameter estimates for the extended model for urban carriageways, based on the negative binomial distribution.*

The results of the Type 1 and Type 3 analyses are given in *Tables 4.7* and *4.8*. These results indicate that the statistical significance of the variables log(L) and log$(AADT)$ is of a higher level than the statistical significance of $AADT/1000$. Furthermore, the Type 1 analysis indicates that the model fitted in this section is an improvement of the model of *Section 3.1.2*.

Source	Twice the log likelihood	χ^2	p-value
Intercept	14320.4013		
log(L)	14684.6662	364.26	$<$ 0.0001
log$(AADT)$	14699.6700	15.00	0.0001
$AADT$/1000	14707.9579	8.29	0.0040

Table 4.7. *Statistics for the Type 1 analysis of the extended model for urban carriageways, based on the negative binomial distribution.*

Source	χ^2	p-value
log(L)	372.12	$<$ 0.0001
log$(AADT)$	17.87	$<$ 0.0001
$AADT$/1000	8.29	0.0040

Table 4.8. *Statistics for the Type 3 analysis of the extended model for urban carriageways, based on the negative binomial distribution.*

The plots of the standardized deviance residuals against the explanatory variables and the linear predictor are given in *Figures 4.6 – 4.9*. These scatter plots show less dependency from the residuals on the explanatory variables and linear predictor than the scatter plots for the residuals of the Poisson based model, discussed in the previous section. There is not a big difference between *Figures 4.6 – 4.9* and *Figures 3.5 – 3.7*.

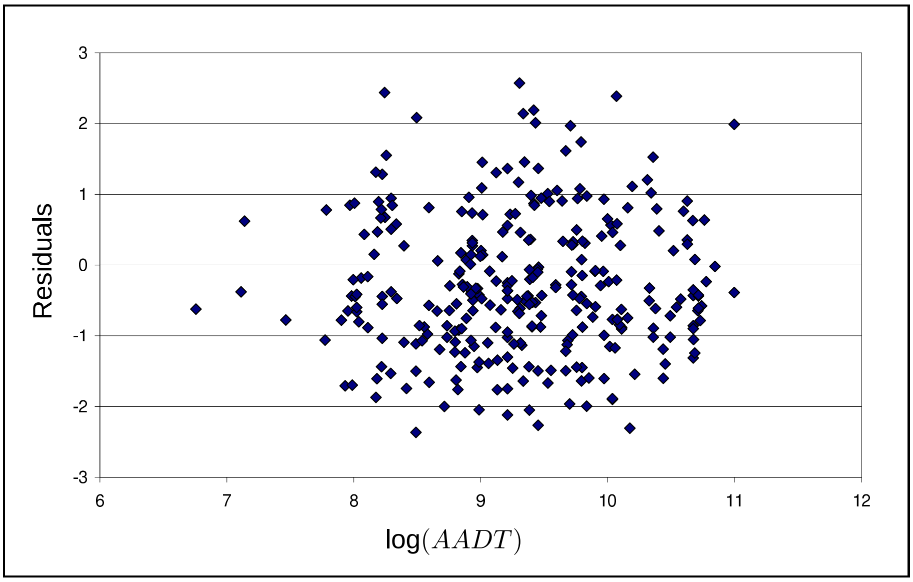

Figure 4.6. *The standardized deviance residuals of the extended model for urban carriageways, based on the negative binomial distribution against log$(AADT)$.*

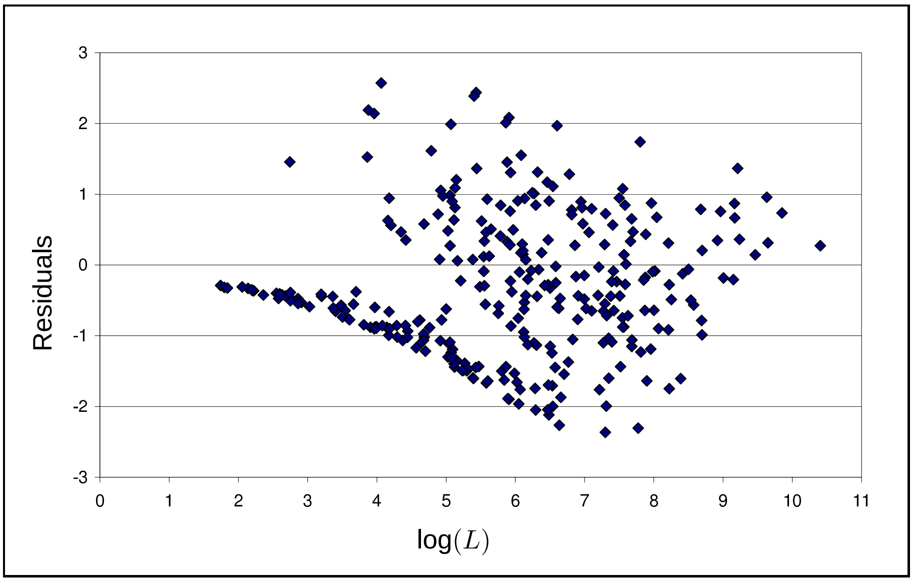

Figure 4.7. *The standardized deviance residuals of the extended model for urban carriageways, based on the negative binomial distribution against log(L).*

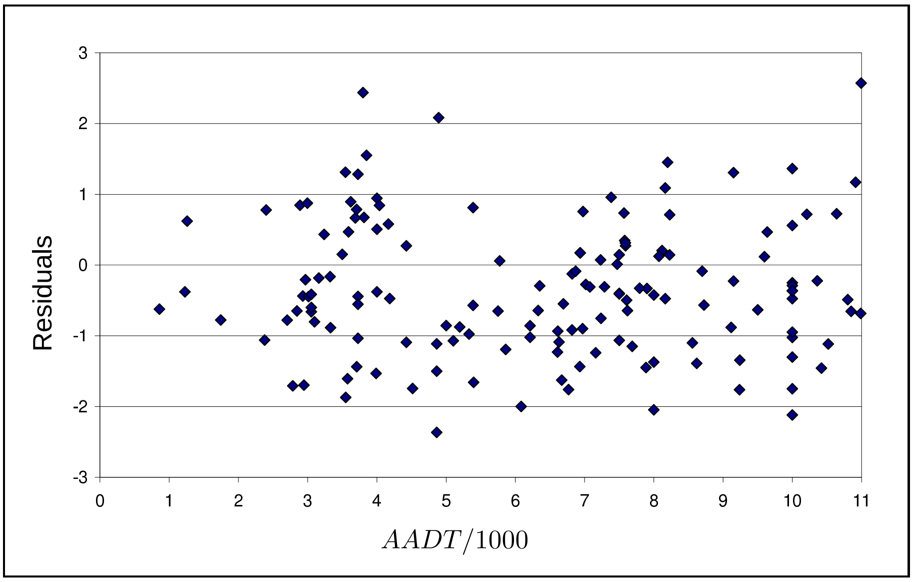

Figure 4.8. *The standardized deviance residuals of the extended model for urban carriageways, based on the negative binomial distribution against* $AADT$.

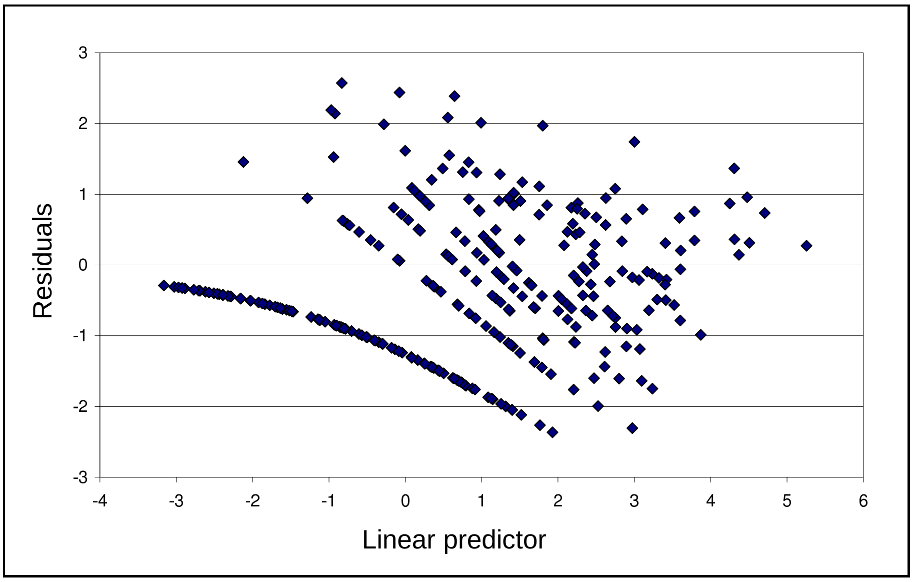

Figure 4.9. *The standardized deviance residuals of the extended model for urban carriageways, based on the negative binomial distribution against the linear predictor.*

The QQ-plot (*Figure 4.10*) strongly resembles a straight line, which indicates that the standardized deviance residuals indeed are standard normally distributed.

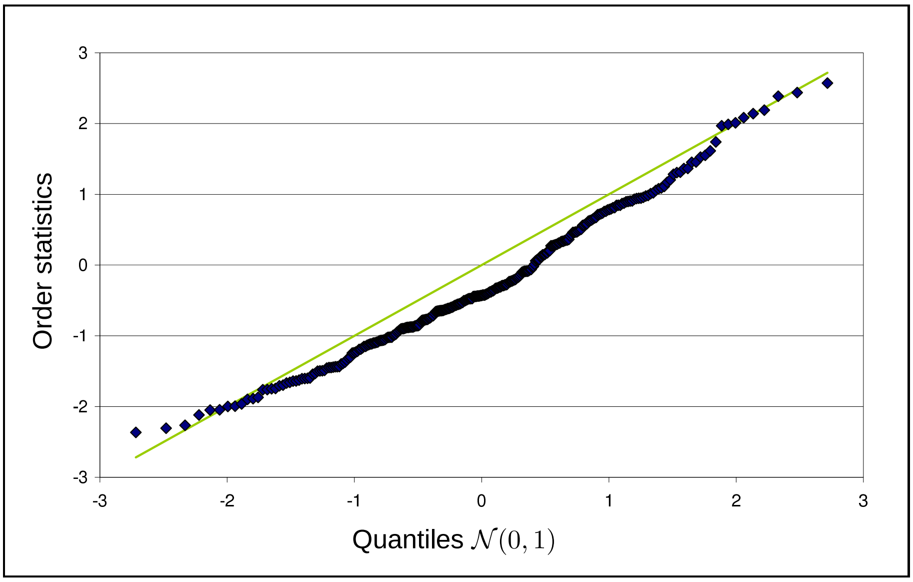

Figure 4.10. *The QQ-plot for the standardized deviance residuals of the extended model for urban carriageways, based on the negative binomial distribution.*

4.1.3. *The quasi-likelihood method*

The model based on the Poisson distribution can also be obtained by applying the quasi-likelihood method. The dispersion parameter is again estimated by the deviance divided by the number of degrees of freedom. The goodness-of-fit of the obtained model is described in *Table 4.9*.

Criterion	Degrees of freedom (DF)	Value	Value/DF
Deviance	299	1132.5283	3.7877
Scaled deviance	299	299.0000	1.0000
Pearson's χ^2	299	1285.2341	4.2984
Scaled Pearson's χ^2	299	339.3160	1.1348
Log likelihood		1875.0861	

Table 4.9. *Criteria for assessing the goodness-of-fit of the extended model for urban carriageways developed using the quasi-likelihood method.*

In *Table 4.10* the parameter estimates and the corresponding statistics are stated. The parameter estimates are equal to the ones in *Table 4.2*. Although the standard errors in *Table 4.10* are larger than the standard errors in *Table 4.2*, and hence the statistical significance of the variables is smaller, the variables still are statistically significant for all confidence levels higher than 0.0001.

Parameter	Estimate	Standard error	Wald's 95% confidence interval	χ^2	p-value
Intercept	-14.6050	1.3789	(-17.3077, -11.9023)	112.18	$<$ 0.0001
log(L)	1.0915	0.0306	(1.0315, 1.1514)	1273.03	$<$ 0.0001
log$(AADT)$	1.0406	0.1600	(0.7270, 1.3543)	42.30	$<$ 0.0001
$AADT$/1000	-0.0581	0.0113	(-0.0803, -0.0358)	26.23	$<$ 0.0001
σ	1.9462	0.0000	(1.9462, 1.9462)		

Table 4.10. *Analysis of the parameter estimates for the extended model for urban carriageways developed using the quasi-likelihood method.*

From the Type 1 and Type 3 analyses it also follows that the variables are statistically significant with high confidence, see *Tables 4.11* and *4.12*. The statistical significance of $AADT/1000$ again shows that adding this variable improves the model.

Source	Deviance	χ^2	p-value
Intercept	8337.5883		
log(L)	1331.8325	1849.60	$<$ 0.0001
log$(AADT)$	1247.9565	22.14	$<$ 0.0001
$AADT$/1000	1132.5283	30.47	$<$ 0.0001

Table 4.11. *Statistics for the Type 1 analysis of the extended model for urban carriageways developed using the quasi-likelihood method.*

Source	χ^2	p-value
log(L)	1778.51	$<$ 0.0001
log$(AADT)$	48.35	$<$ 0.0001
$AADT$/1000	30.47	$<$ 0.0001

Table 4.12. *Statistics for the Type 3 analysis of the extended model for urban carriageways developed using the quasi-likelihood method.*

The QQ-plot of the standardized deviance residuals is given in *Figure 4.11*. Except for the middle and the end, the dots in this plot are close to the line with slope 1.

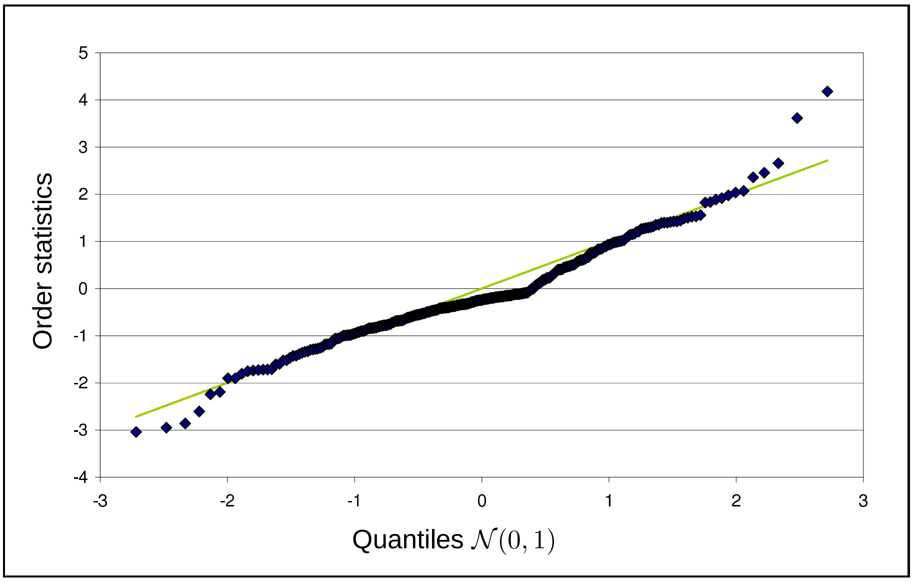

Figure 4.11. *The QQ-plot for the standardized deviance residuals of the extended model for urban carriageways developed using the quasi-likelihood method.*

4.1.4. *Discussion*

In *Sections 4.1.1 – 4.1.3* two different models are derived which describe the relation between the number of road crashes on urban carriageways over a period of three years, the AADT and carriageway length. These models are

$$\hat{\mu}_i = 4.5408 \cdot 10^{-7} \cdot L_i^{1.0915} \cdot AADT_i^{1.0406} \cdot e^{-0.0581 \cdot \frac{AADT_i}{1000}}, \qquad (4.2)$$

$$\hat{\mu}_i = 5.8880 \cdot 10^{-6} \cdot L_i^{0.9875} \cdot AADT_i^{0.8181} \cdot e^{-0.0375 \cdot \frac{AADT_i}{1000}}. \qquad (4.3)$$

Model *(4.2)* was obtained in two different ways: assuming that the number of road crashes follows a Poisson distribution and by applying the quasi-likelihood method. This last method is preferred, because it solves the overdispersion problem. Under the assumption that the number of road crashes is negative binomially distributed model *(4.3)* was obtained. This model is not affected by overdispersion either. Furthermore, the plots involving the standardized deviance residuals of this second model are better, i.e., are closer to the desired form, than those of the first model.

Similar to the simple models for urban carriageways, the exponents of L in both models are close to 1. For the negative binomial based model this value is even included in the confidence interval. Therefore the following approximation holds:

$$\frac{\hat{\mu}_i}{L_i} \approx \begin{cases} 4.5408 \cdot 10^{-7} \cdot AADT_i^{1.0406} \cdot e^{-0.0581 \cdot \frac{AADT_i}{1000}}, & \text{Poisson,} \\ 5.8880 \cdot 10^{-6} \cdot AADT_i^{0.8181} \cdot e^{-0.0375 \cdot \frac{AADT_i}{1000}}, & \text{neg. bin.} \end{cases}$$

So the models for τ_i (the number of road crashes per kilometre per year) are approximately

$$\tau_i \approx \begin{cases} 1.51 \cdot 10^{-4} \cdot AADT_i^{1.0406} \cdot e^{-0.0581 \cdot \frac{AADT_i}{1000}}, & \text{Poisson,} \\ 1.96 \cdot 10^{-3} \cdot AADT_i^{0.8181} \cdot e^{-0.0375 \cdot \frac{AADT_i}{1000}}, & \text{neg. bin.} \end{cases}$$

These models are plotted in *Figure 4.12*. In general, the model based on the negative binomial distribution predicts a higher risk for high AADT than the model based on the Poisson distribution.

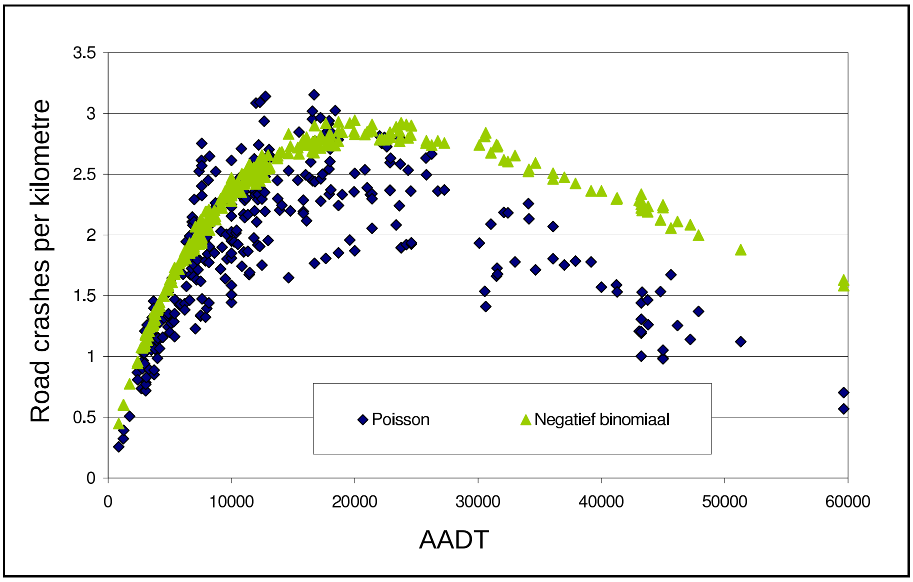

Figure 4.12. *The predicted number of road crashes per kilometre per year against the AADT for urban carriageways.*

In order to obtain models in which $\hat{\mu}_i/L_i$ does not depend on L the variable $\log(L)$ is included in the model as an offset variable. The resulting models for τ_i are:

$$\tau_i = \begin{cases} 3.30 \cdot 10^{-4} \cdot AADT_i^{1.0467} \cdot e^{-0.0637 \cdot \frac{AADT_i}{1000}}, & \text{Poisson,} \\ 1.77 \cdot 10^{-3} \cdot AADT_i^{0.8193} \cdot e^{-0.0373 \cdot \frac{AADT_i}{1000}}, & \text{neg. bin.} \end{cases}$$

These expressions are plotted in *Figure 4.13*. This plot also shows the difference between the risk predicted by the Poisson based and by the negative binomial based model for high AADT. The shape of these plots obviously is more like the shape of *Figures 2.1* and *2.2* than that of *Figure 3.22* does.

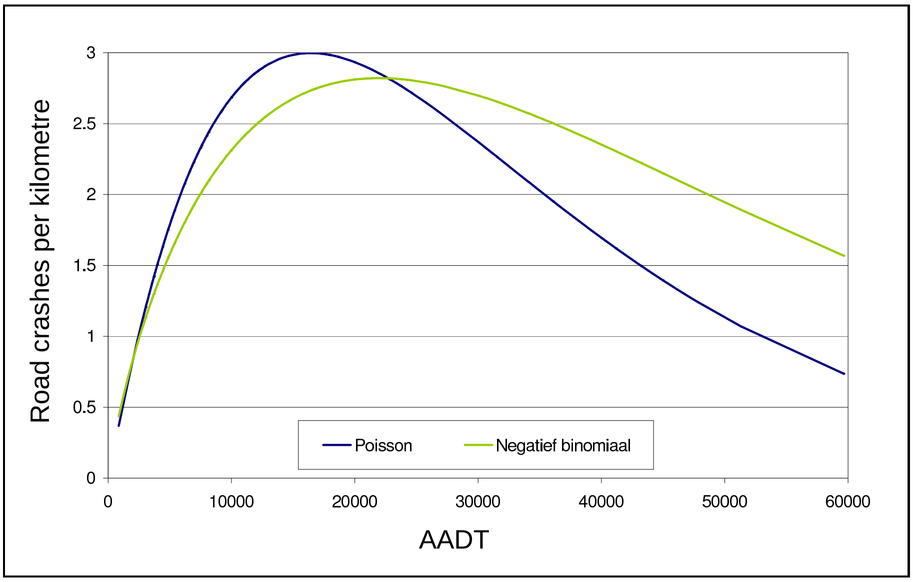

Figure 4.13. *The predicted number of road crashes per kilometre per year against the AADT for urban carriageways with log(L) as an offset variable.*

4.2. Rural carriageways

4.2.1. *The Poisson distribution*

The goodness-of-fit of the Poisson based model for rural carriageways is described by the statistics in *Table 4.13.* The overdispersion is slightly less than in the Poisson based model involving only two explanatory variables for rural carriageways.

Criterion	Degrees of freedom (DF)	Value	Value/DF
Deviance	94	180.8276	1.9237
Pearson's χ^2	94	179.7388	1.9121
Log likelihood		307.5897	

Table 4.13. *Criteria for assessing the goodness-of-fit of the extended model for rural carriageways, based on the Poisson distribution.*

Table 4.14 contains the parameter estimates together with several statistics. It follows that the model for rural carriageways is given by:

$$\hat{\mu}_i = 3.78 \cdot 10^{-8} \cdot L_i^{0.9133} \cdot AADT_i^{1.4058} \cdot e^{-0.0956 \cdot \frac{AADT_i}{1000}}.$$

The variables log(L) and log$(AADT)$ are statistically significant with relative high confidence, whereas $AADT/1000$ is statistically significant for confidence levels higher than 0.02841 Furthermore, the confidence interval for log(L) does contain 1.

Parameter	Estimate	Standard error	Wald's 95% confidence interval	Wald's χ^2	p-value
Intercept	-17.0912	4.3002	(-25.5194, -8.6631)	15.80	$<$ 0.0001
log(L)	0.9133	0.0478	(0.8196, 1.0070)	364.81	$<$ 0.0001
log$(AADT)$	1.4058	0.5189	(0.3888, 2.4228)	7.34	0.0067
$AADT$/1000	-0.0956	0.0436	(-0.1811, -0.0101)	4.81	0.0284

Table 4.14. *Analysis of the parameter estimates for the extended model for rural carriageways, based on the Poisson distribution.*

The Type 1 and Type 3 analyses give the same results, see *Tables 4.15* and *4.16*. From the Type 1 analysis it follows that if $AADT/1000$ is added to the model only containing the intercept, log(L), and log$(AADT)$ as explanatory variables, we can conclude that its parameter is not equal to zero, in other words, that adding log$(AADT)$ improves the model. Indeed, its p-value is 0.0233.

Source	Twice the log likelihood	χ^2	p-value
Intercept	678.6608		
log(L)	192.8836	485.78	$<$ 0.0001
log$(AADT)$	185.9737	6.91	0.0086
$AADT/1000$	180.8276	5.15	0.0233

Table 4.15. *Statistics for the Type 1 analysis of the extended model for rural carriageways, based on the Poisson distribution.*

Source	χ^2	p-value
log(L)	478.30	$<$ 0.0001
log$(AADT)$	8.03	0.0046
$AADT/1000$	5.15	0.0233

Table 4.16. *Statistics for the Type 3 analysis of the extended model for rural carriageways, based on the Poisson distribution.*

The standardized deviance residuals are plotted against the explanatory variables and the linear predictor in *Figures 4.14 – 4.17*. They are very much similar to *Figures 3.12 – 3.14*. Also the QQ-plot did not change considerably, compare *Figures 4.18* and *3.15*. Therefore, adding the extra term to the model did not improve the behaviour of the residuals.

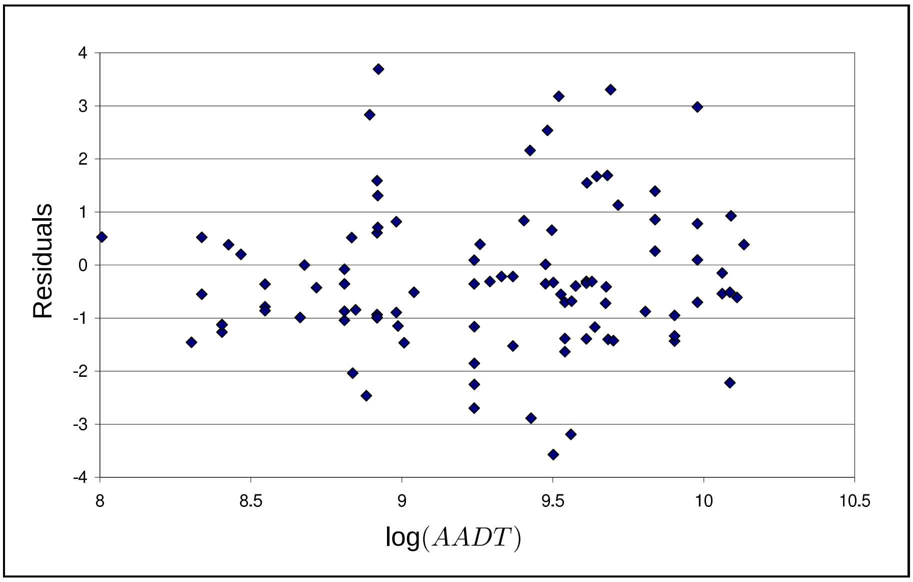

Figure 4.14. *The standardized deviance residuals of the extended model for rural carriageways, based on the Poisson distribution against log*$(AADT)$*.*

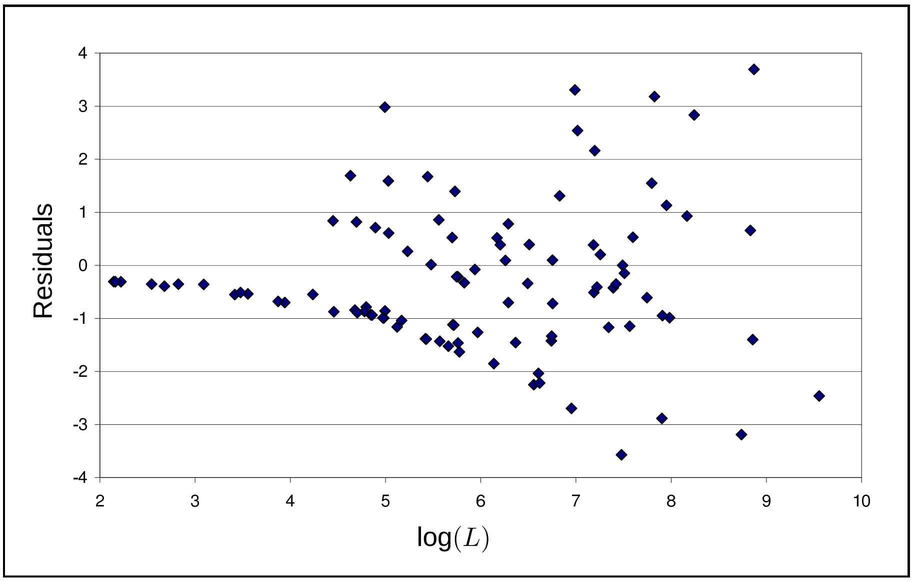

Figure 4.15. *The standardized deviance residuals of the extended model for rural carriageways, based on the Poisson distribution against log*(L)*.*

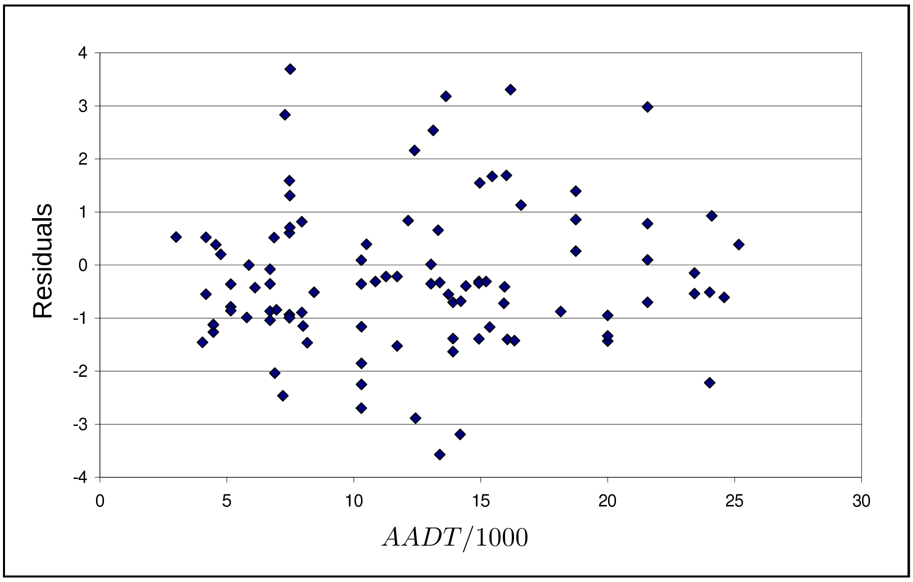

Figure 4.16. *The standardized deviance residuals of the extended model for rural carriageways, based on the Poisson distribution against $AADT$.*

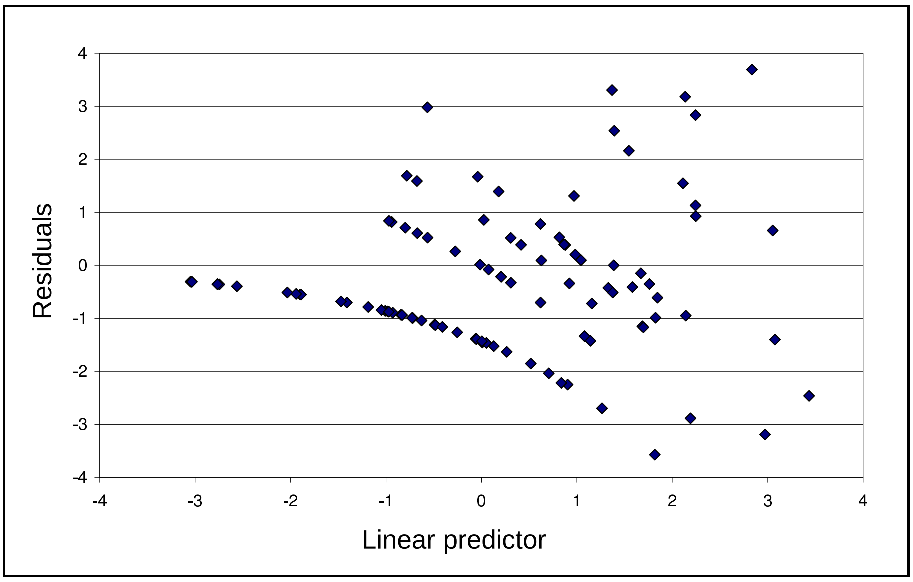

Figure 4.17. *The standardized deviance residuals of the extended model for rural carriageways, based on the Poisson distribution against the linear predictor.*

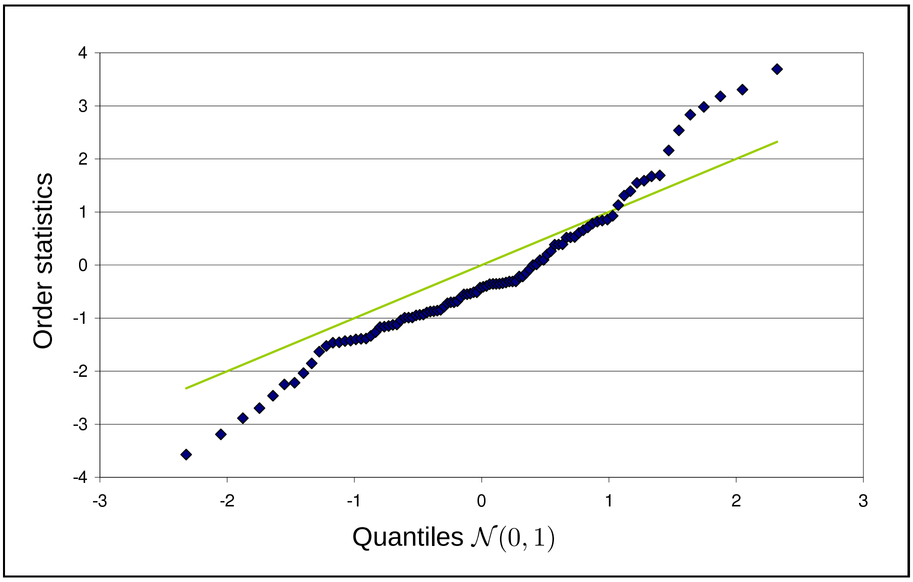

Figure 4.18. *The QQ-plot for the standardized deviance residuals of the extended model for rural carriageways, based on the Poisson distribution.*

4.2.2. *The negative binomial distribution*

In this section the results for the model based on the negative binomial distribution are given. From the deviance and Pearson's χ^2 it follows that there is no overdispersion, Pearson's χ^2 even indicates underdispersion, see *Table 4.17*. Comparing the value of the deviance to its χ^2_{94} distribution gives a p-value of 0.4531, showing that the null hypothesis that the considered model is the right model, cannot be rejected.

Criterion	Degrees of freedom (DF)	Value	Value/DF
Deviance	94	94.9517	1.0101
Pearson's χ^2	94	92.1315	0.9801
Log likelihood		326.4145	

Table 4.17. *Criteria for assessing the goodness-of-fit of the extended model for rural carriageways, based on the negative binomial distribution.*

The parameter estimates and the corresponding statistics are given in *Table 4.18*. So the model for rural carriageways is given by:

$$\hat{\mu}_i = 7.52 \cdot 10^{-9} \cdot L_i^{0.9588} \cdot AADT_i^{1.5407} \cdot e^{-0.0940 \cdot \frac{AADT_i}{1000}}$$

Again the exponent of L is close to 1, this value is even included in the 95%-confidence interval corresponding to $\log(L)$. The variables log(L) and log$(AADT)$ are statistically significant with reasonably high confidence.

Parameter	Estimate	Standard error	Wald's 95% confidence interval	Wald's χ^2	p-value
Intercept	-18.7059	6.5065	(-31.4583, -5.9534)	8.27	0.0040
log(L)	0.9588	0.0812	(0.7997, 1.1179)	139.49	$<$ 0.0001
log$(AADT)$	1.5407	0.7827	(0.0067, 3.0747)	3.88	0.0490
$AADT$/1000	-0.0940	0.0673	(-0.2259, 0.0378)	1.95	0.1621
$\frac{1}{\nu}$	0.3112	0.1132	(0.0892, 0.5331)		

Table 4.18. *Analysis of the parameter estimates for the extended model for rural carriageways, based on the negative binomial distribution.*

The same conclusion can be drawn from the results of the Type 1 and Type 3 analysis, see *Tables 4.19* and *4.20.* The Type 1 analysis indicates that the confidence of $AADT/1000$ is far less than that of the other two variables. The null hypothesis that adding $AADT/1000$ to the model is not an improvement, can not even be rejected.

Source	Twice the log likelihood	χ^2	p-value
Intercept	546.8178		
log(L)	644.8201	98.00	$<$ 0.0001
log$(AADT)$	650.9237	6.10	0.0135
$AADT$/1000	652.8290	1.91	0.1675

Table 4.19. *Statistics for the Type 1 analysis of the extended model for rural carriageways, based on the negative binomial distribution.*

Source	χ^2	p-value
log(L)	103.59	$<$ 0.0001
log$(AADT)$	3.82	0.0507
$AADT$/1000	1.91	0.1675

Table 4.20. *Statistics for the Type 3 analysis of the extended model for rural carriageways, based on the negative binomial distribution.*

Also the scatter plots of the residuals against the explanatory variables and the linear predictor are given, see *Figures 4.19 – 4.22.*

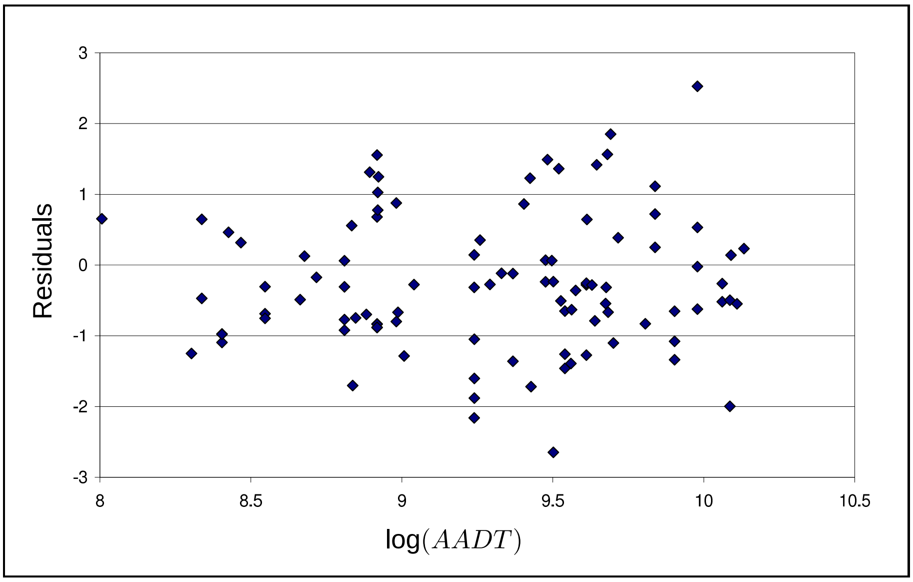

Figure 4.19. *The standardized deviance residuals of the extended model for rural carriageways, based on the negative binomial distribution against log$(AADT)$.*

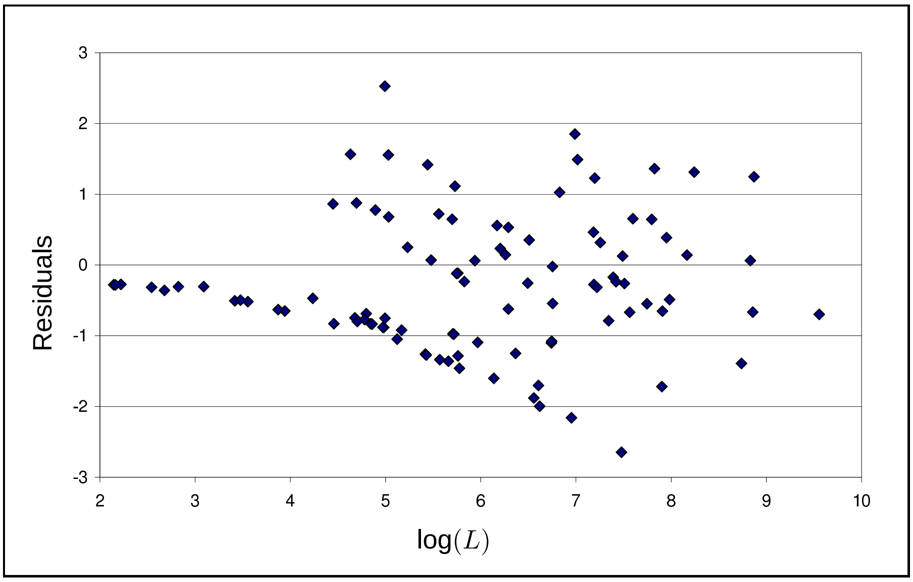

Figure 4.20. *The standardized deviance residuals of the extended model for rural carriageways, based on the negative binomial distribution against log(L).*

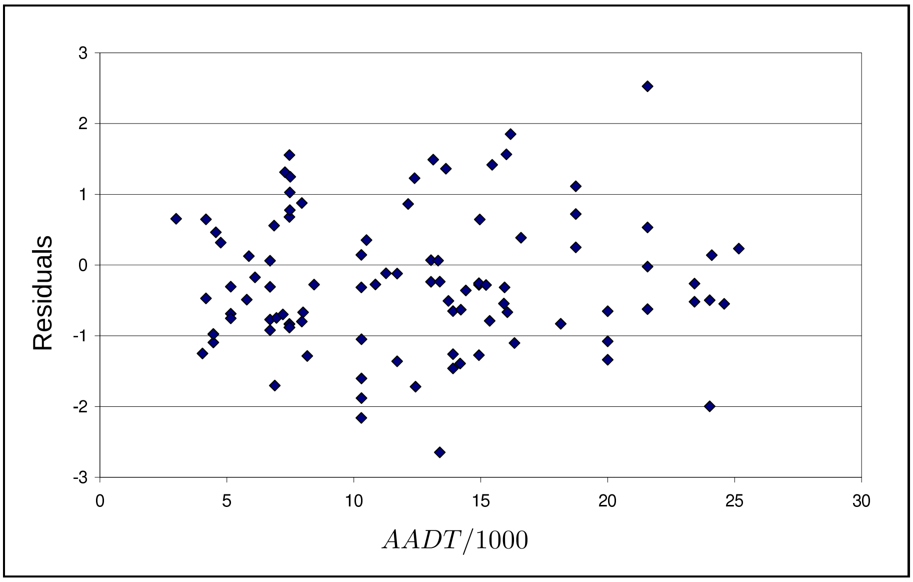

Figure 4.21. *The standardized deviance residuals of the extended model for rural carriageways, based on the negative binomial distribution against $AADT$.*

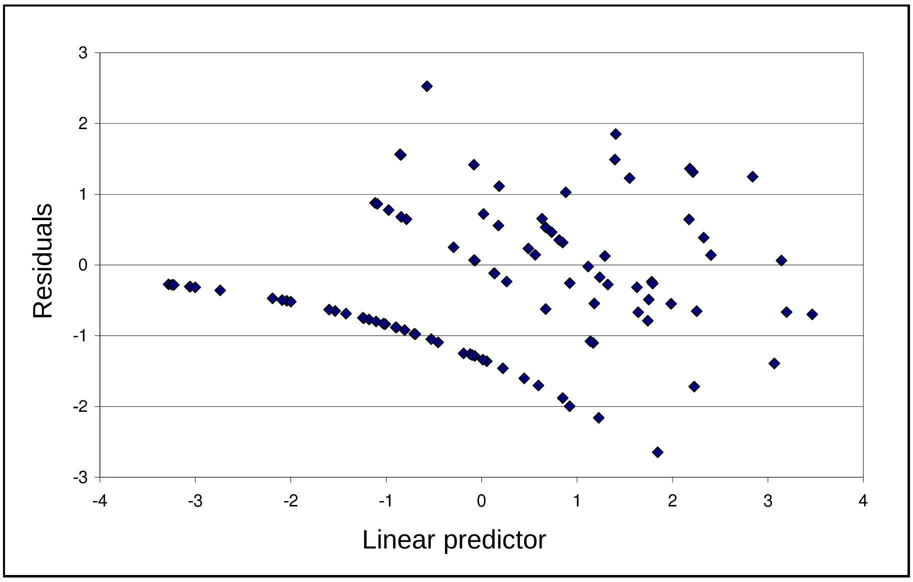

Figure 4.22. *The standardized deviance residuals of the extended model for rural carriageways, based on the negative binomial distribution against the linear predictor.*

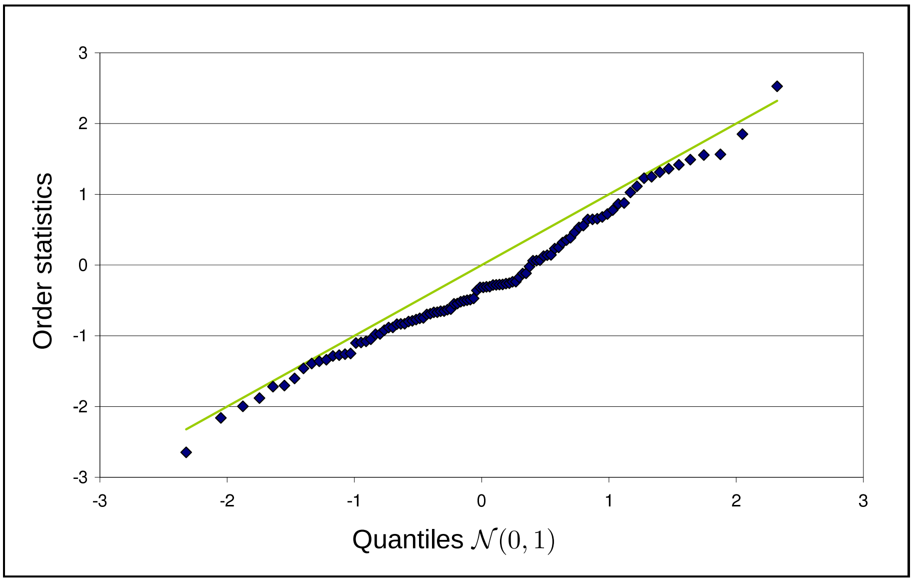

Figure 4.23. *The QQ-plot for the standardized deviance residuals of the extended model for rural carriageways, based on the negative binomial distribution.*

4.2.3. *The quasi-likelihood method*

The last model to be discussed in this chapter is the one obtained by applying the quasi-likelihood method. As before, the dispersion parameter is estimated by the deviance divided by the number of degrees of freedom. The goodness-of-fit statistics are given in *Table 4.21.*

Criterion	Degrees of freedom (DF)	Value	Value/DF
Deviance	94	180.8276	1.9237
Scaled deviance	94	94.0000	1.0000
Pearson's χ^2	94	179.7388	1.9121
Scaled Pearson's χ^2	94	93.4340	0.9940
Log likelihood		159.8951	

Table 4.21. *Criteria for assessing the goodness-of-fit of the extended model for rural carriageways developed using the quasi-likelihood method.*

The parameter estimates are the same as for the model based on the Poisson distribution. They are given, together with the changed statistics, in *Table 4.22.* The statistical significance of the variables decreased, which is easy to see for log$(AADT)$ and $AADT/1000$. The variable log(L) is still statistically significant for all confidence levels higher than 0.0001.

Parameter	Estimate	Standard error	Wald's 95% confidence interval	Wald's χ^2	p-value
Intercept	-17.0912	5.9642	(-28.7808, -5.4016)	8.21	0.0042
log(L)	0.9133	0.0663	(0.7833, 1.0433)	189.64	$<$ 0.0001
log$(AADT)$	1.4058	0.7197	(-0.0048, 2.8163)	3.82	0.0508
$AADT$/1000	-0.0956	0.0605	(-0.2142, -0.0230)	2.50	0.1140
σ	1.3870	0.0000	(1.3870, 1.3870)		

Table 4.22. *Analysis of the parameter estimates for the extended model for rural carriageways developed using the quasi-likelihood method.*

The Type 1 and Type 3 analyses show that adding the variables log$(AADT)$ and $AADT/1000$ to the model, does not statistically significantly improve the fit of the model.

Source	Deviance	χ^2	p-value χ^2
Intercept	678.6608		
log(L)	192.8836	252.52	$<$ 0.0001
log$(AADT)$	185.9737	3.59	0.0611
$AADT$/1000	180.8276	2.68	0.1053

Table 4.23. *Statistics for the Type 1 analysis of the extended model for rural carriageways developed using the quasi-likelihood method.*

Source	χ^2	p-value
log(L)	248.64	$<$ 0.0001
log$(AADT)$	4.18	0.0438
$AADT$/1000	2.68	0.1053

Table 4.24. *Statistics for the Type 3 analysis of the extended model for rural carriageways developed using the quasi-likelihood method.*

The QQ-plot of the standardized deviance residuals is given in *Figure 4.24*.

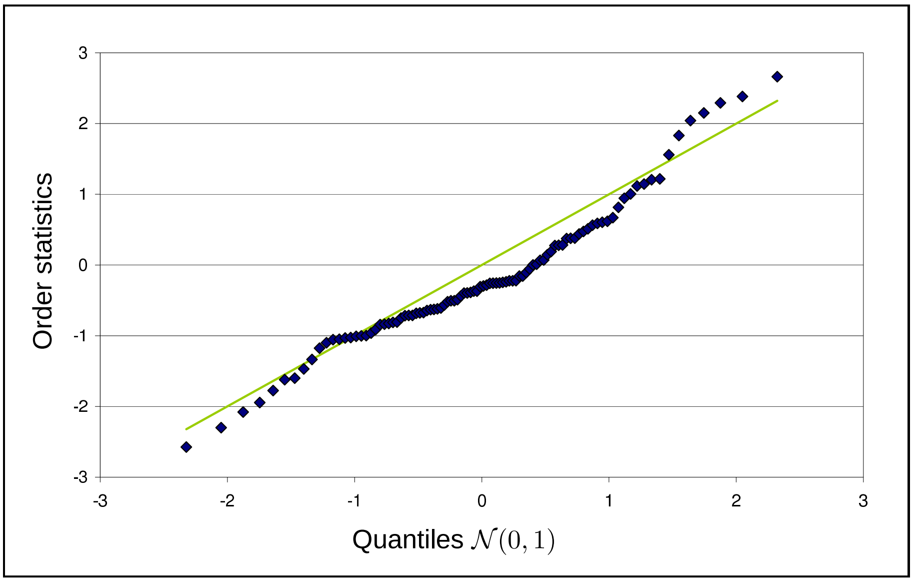

Figure 4.24. *The QQ-plot for the standardized deviance residuals of the extended model for rural carriageways developed using the quasi-likelihood method.*

4.2.4. *Discussion*

In *Sections 4.2.1 – 4.2.3* two different models are derived which describe the relation between the number of road crashes on urban carriageways in three years, the AADT, and carriageway length. These models are

$$\hat{\mu}_i = 3.78 \cdot 10^{-8} \cdot L_i^{0.9133} \cdot AADT_i^{1.4058} \cdot e^{-0.0956 \cdot \frac{AADT_i}{1000}}, \tag{4.4}$$

$$\hat{\mu}_i = 7.52 \cdot 10^{-9} \cdot L_i^{0.9588} \cdot AADT_i^{1.5407} \cdot e^{-0.0940 \cdot \frac{AADT_i}{1000}}. \tag{4.5}$$

Model *(4.4)* was derived in two different ways: 1) by assuming that the number of road crashes follows a Poisson distribution and 2) by applying the quasi-likelihood method. Under the assumption that the number of road crashes is negative binomially distributed, model *(4.5)* was obtained.

Like before, the exponents of L in both models are close to 1. For the negative binomial and quasi-likelihood based model 1 is even contained in the confidence interval. Hence

$$\frac{\hat{\mu}_i}{L_i} \approx \begin{cases} 3.78 \cdot 10^{-8} \cdot AADT_i^{1.4058} \cdot e^{-0.0956 \cdot \frac{AADT_i}{1000}}, & \text{Poisson,} \\ 7.52 \cdot 10^{-9} \cdot AADT_i^{1.5407} \cdot e^{-0.0940 \cdot \frac{AADT_i}{1000}}, & \text{neg. bin.} \end{cases}$$

and

$$\tau_i \approx \begin{cases} 1.26 \cdot 10^{-5} \cdot AADT_i^{1.4058} \cdot e^{-0.0956 \cdot \frac{AADT_i}{1000}}, & \text{Poisson,} \\ 2.50 \cdot 10^{-6} \cdot AADT_i^{1.5407} \cdot e^{-0.0940 \cdot \frac{AADT_i}{1000}}, & \text{neg. bin.} \end{cases}$$

In *Figure 4.25* the values of τ_i are plotted against the AADT.

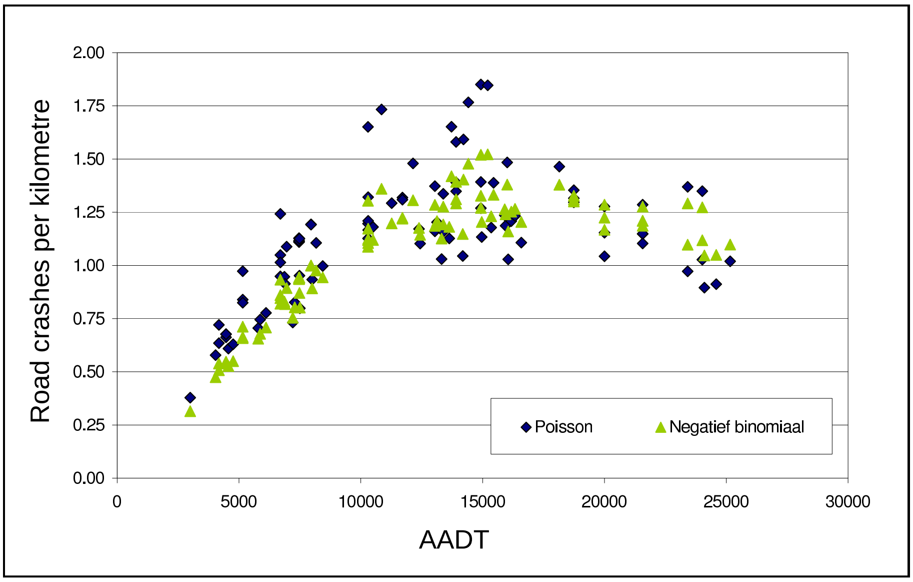

Figure 4.25. *The predicted number of road crashes per kilometre per year against the AADT for rural carriageways.*

In order to obtain models in which $\hat{\mu}_i/L_i$ does not depend on L the variable log(L) is included in the model as an offset variable. The resulting models for τ_i are:

$$\tau_i = \begin{cases} 1.33 \cdot 10^{-5} \cdot AADT_i^{1.3103} \cdot e^{-0.0852 \cdot \frac{AADT_i}{1000}}, & \text{Poisson,} \\ 1.94 \cdot 10^{-6} \cdot AADT_i^{1.5353} \cdot e^{-0.0930 \cdot \frac{AADT_i}{1000}}, & \text{neg. bin.} \end{cases}$$

These expressions are plotted in *Figure 4.26*.

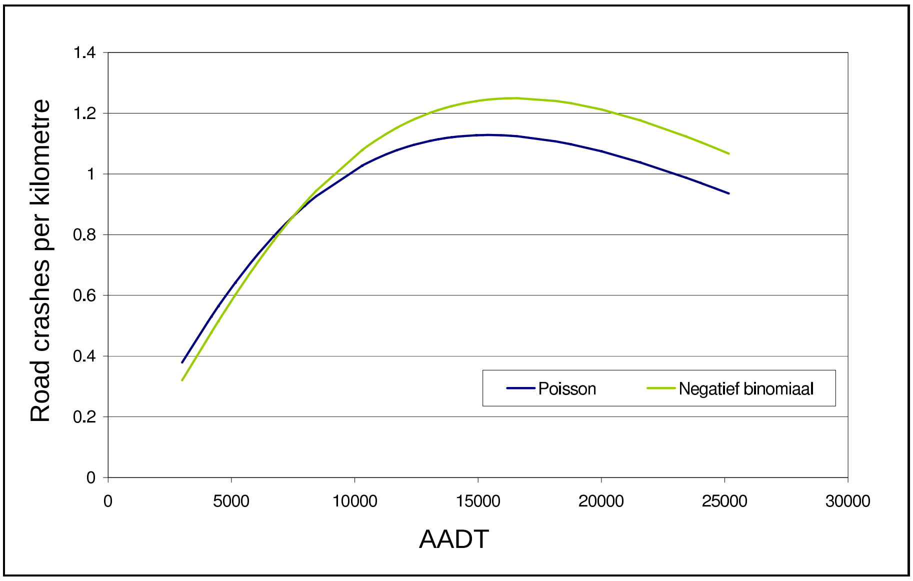

Figure 4.26. *The predicted number of road crashes per kilometre per year against the AADT for rural carriageways with log*(L) *as an offset variable.*

4.3. Comparison of the extended models

It is interesting to compare the models for urban carriageways to the models for rural carriageways. A first conclusion is that the derived models for urban carriageways are more reliable than the models for rural carriageways. This follows from the following two observations:

- For all models, the explanatory variables for urban carriageways are statistically significant for all confidence levels higher than 0.0001, except for $AADT/1000$ in the negative binomial model. For the models for rural carriageways only log(L) is statistically significant for all confidence levels higher than 0.0001.
- The standard errors of the parameter estimates for the models for rural carriageways are generally much larger than for the models for urban carriageways.

This is possibly a consequence of the number of available carriageways: the database contained information of almost three times more urban carriageways than rural carriageways.

Secondly, for urban carriageways as well as for rural carriageways, the exponent of L in the developed models is reasonable close to 1. However, this value is included in the confidence interval corresponding to log(L) for only three of the six models. Hence the number of crashes on urban and rural carriageways is approximately proportional to the carriageway length. By including log(L) in the model as an offset variable its exponent is forced to be equal to 1.

Finally, it is also possible to compare the modelled risk of urban and rural carriageways. For an easy comparison the obtained models for τ_i with log(L) as an offset variable are plotted in *Figure 3.23*. It follows that the modelled risk for urban carriageways is higher than the risk for rural carriageways for equal AADT. This was already well-known.

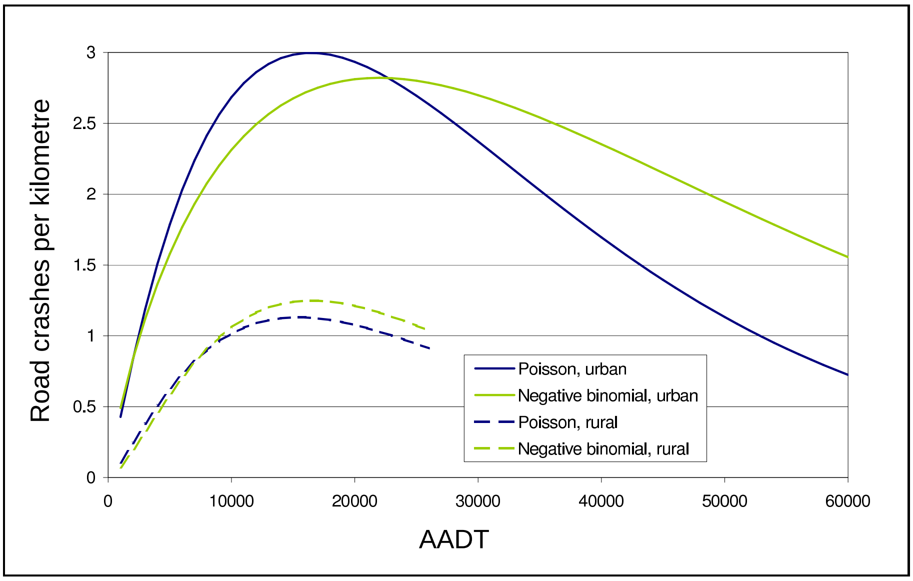

Figure 4.27. *The predicted number of road crashes per kilometre per year against the AADT.*

5. Models for the other road types

5.1. Introduction

In *Table 2.1* the crash rates for different carriageway types in Haaglanden were given. *Chapters 3* and *4* showed that for urban and rural carriageways crash rate heavily depends on the AADT: the crash rate decreases when the AADT increases. This is possibly also the case for further disaggregations of the selection of carriageways. Therefore it would be informative to develop accident prediction models for each carriageway type given in *Table 2.1*.
A problem is that the database does not contain sufficient data for each type of carriageway to make fitting reliable accident prediction models possible. In this chapter, models will be fitted for the types for which sufficient carriageways are available and some plots will be made for some of the other carriageway types to indicate the differences in crash rate for the different carriageway types.

5.2. Carriageways of urban dual carriageway roads with a speed limit of 50 km/h

In this section a model will be fitted for carriageways of urban dual carriageway distributor roads, with a speed limit of 50 km/h, one lane and one driving direction. The database contains 138 carriageways satisfying these conditions. Based on the results in *Chapters 3* and *4* and on *Figure 5.1* it was decided to use the negative binomial distribution and the extended model form.

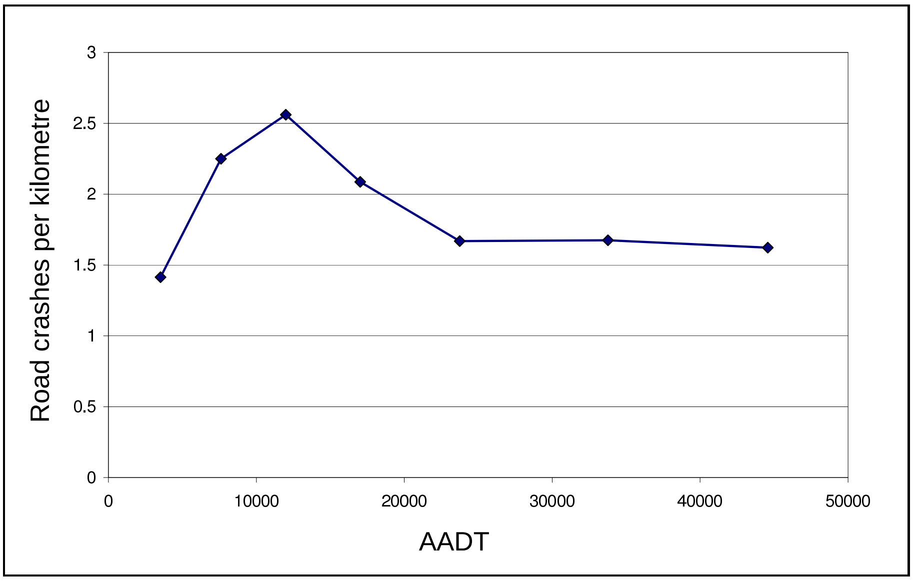

Figure 5.1. *The number of road crashes per kilometre per year in each AADT class for carriageways of urban dual carriageway roads, with a speed limit of 50 km/h, one lane and one driving direction.*

The modelling results are given in *Tables 5.1 – 5.4*. The deviance indicates that the hypothesis that the fitted model is the correct model cannot be rejected on basis of the confidence level $\alpha = 0.05$, because its p-value is

0.0771. The p-value corresponding to Pearson's χ^2, however, is equal to 0.0009, from which this conclusion can not be drawn.

Criterion	Degrees of freedom (DF)	Value	Value/DF
Deviance	133	156.8821	1.1796
Pearson's χ^2	133	189.9423	1.4281
Log likelihood		4049.3505	

Table 5.1. *Criteria for assessing the goodness-of-fit of the model for carriageways of urban dual carriageway roads, with a speed limit of 50 km/h, one lane and one driving direction.*

From *Tables 5.2 – 5.4* it follows that the parameter estimates are reasonably statistically significant, except maybe the parameter corresponding with $AADT/1000$. A model was also fitted without this variable. The deviance and Pearson's χ^2 were slightly higher for this model, indicating a less adequate fit. Therefore the variable was kept in the model.

Parameter	Estimate	Standard error	Wald's 95% confidence interval	Wald's χ^2	p-value
Intercept	-10.9144	2.3271	(-15.4753, -6.3534)	22.00	$<$ 0.0001
log(L)	1.0067	0.0584	(0.8922, 1.1212)	296.90	$<$ 0.0001
log$(AADT)$	0.6523	0.2677	(0.1276, 1.1770)	5.94	0.0148
$AADT$/1000	-0.0279	0.0170	(-0.0611, 0.0054)	2.70	0.1005
$\frac{1}{\nu}$	0.4719	0.1091	(0.2582, 0.6857)		

Table 5.2. *Analysis of the parameter estimates for the model for urban carriageways of of dual carriageway roads, with a speed limit of 50 km/h, one land and one driving direction.*

Source	Deviance	χ^2	p-value χ^2
Intercept	7918.9947		
log(L)	8091.2635	172.27	$<$ 0.0001
log$(AADT)$	8096.0405	4.78	0.0288
$AADT$/1000	8098.7010	2.66	0.1029

Table 5.3. *Statistics for the Type 1 analysis of the model for carriageways of urban dual carriageway roads, with a speed limit of 50 km/h, one lane and one driving direction.*

Source	χ^2	p-value
log(L)	170.53	$<$ 0.0001
log$(AADT)$	5.71	0.0169
$AADT$/1000	2.66	0.1029

Table 5.4. *Statistics for the Type 3 analysis of the model for carriageways of urban dual carriageway roads, with a speed limit of 50 km/h, one lane and one driving direction.*

Because the parameter of log(L) is very close to 1, the number of road crashes per kilometre per year, denoted by τ_i is approximately given by

$$\tau_i \approx 6.06 \cdot 10^{-3} \cdot AADT_i^{0.6523} \cdot e^{-0.0279 \cdot \frac{AADT_i}{1000}}.$$

The right hand-side of this expression is plotted in *Figure 5.2*.

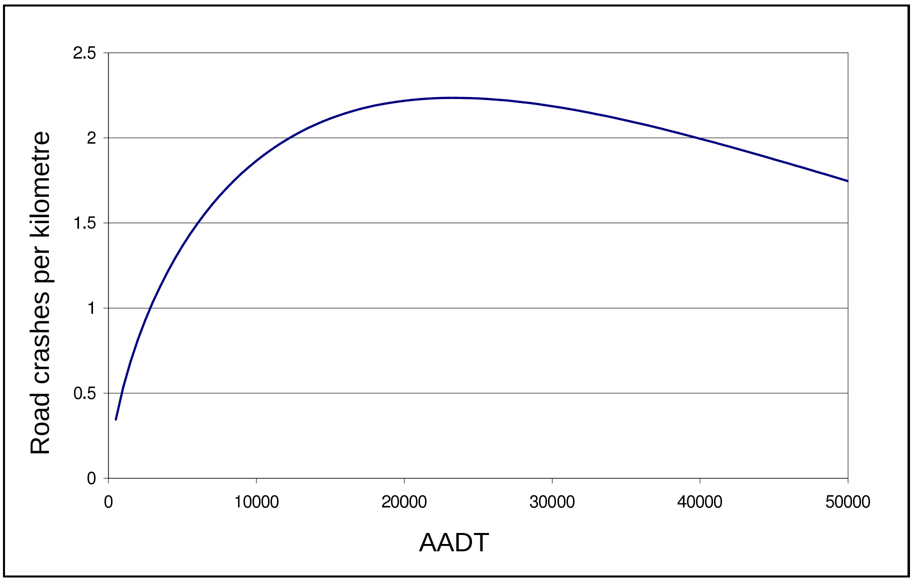

Figure 5.2. *The number of road crashes per kilometre per year for carriageways of urban dual carriageway roads, with a speed limit of 50 km/h, one lane and one driving direction.*

5.3. Carriageways of urban single carriageway roads with a speed limit of 50 km/h

In this section a model will be fitted for carriageways of urban single carriageway distributor roads, with a speed limit of 50 km/h, two lanes and two driving directions. The database contains 122 carriageways satisfying these conditions. According to the results in *Chapters 3* and *4* and to *Figure 5.3* the negative binomial distribution and the extended model form were used.

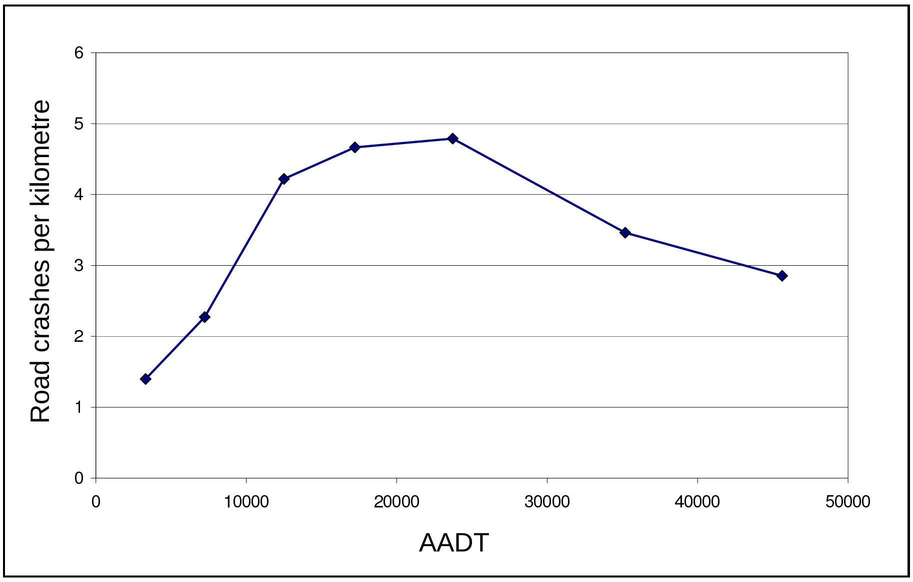

Figure 5.3. *The number of road crashes per kilometre per year in each AADT class for carriageways of urban single carriageway roads, with a speed limit of 50 km/h, two lanes and two driving directions.*

The variable $AADT/1000$ was however not statistically significant at all ($p = 0.8854$). An explanation might be that the decreasing part of *Figure 5.3* is only based on nine carriageways and hence not very reliable. Therefore the model is fitted on the 113 carriageways with an AADT smaller than 30,000. The results are given in *Tables 5.5 – 5.8*. Both the deviance and Pearson's χ^2 indicate that the hypothesis that the fitted model is the correct model with high confidence. Their corresponding p-values is 0.18.

Criterion	Degrees of freedom (DF)	Value	Value/DF
Deviance	110	123.3214	1.1211
Pearson's χ^2	110	123.1570	1.1196
Log likelihood		3266.3493	

Table 5.5. *Criteria for assessing the goodness-of-fit of the model for carriageways of urban single carriageway roads, with a speed limit of 50 km/h, two lanes and two driving directions.*

From *Table 5.6 – 5.8* it follows that all the parameter estimates are highly statistically significant. The parameter of log(L) is again close to 1.

Parameter	Estimate	Standard error	Wald's 95% confidence interval	Wald's χ^2	p-value
Intercept	-11.4869	1.2514	(-13.9397, -9.0342)	84.25	$<$ 0.0001
log(L)	1.0538	0.0601	(0.9361, 1.1715)	307.87	$<$ 0.0001
log$(AADT)$	0.6808	0.1215	(0.4427, 0.9188)	31.42	$<$ 0.0001
$\frac{1}{\nu}$	0.3509	0.0961	(0.1626, 0.5392)		

Table 5.6. *Analysis of the parameter estimates for the model for carriageways of urban single carriageway roads, with a speed limit of 50 km/h, two lanes and two driving directions.*

Source	Deviance	χ^2	p-value χ^2
Intercept	6362.6223		
log(L)	6506.0608	143.44	$<$ 0.0001
log$(AADT)$	6532.6986	26.64	$<$ 0.0001

Table 5.7. *Statistics for the Type 1 analysis of the model for carriageways of urban single carriageway roads, with a speed limit of 50 km/h, two lanes and two driving directions.*

Source	χ^2	p-value
log(L)	167.14	$<$ 0.0001
log$(AADT)$	26.64	$<$ 0.0001

Table 5.8. *Statistics for the Type 3 analysis of the model for carriageways of urban single carriageway roads, with a speed limit of 50 km/h, two lanes and two driving directions.*

Because the parameter of log(L) is very close to 1, the number of road crashes per kilometre per year, denoted by τ_i is approximated by

$$\tau_i \approx 3.42 \cdot 10^{-3} \cdot AADT_i^{0.6808}.$$

The right hand-side of this expression is plotted in *Figure 5.4*.

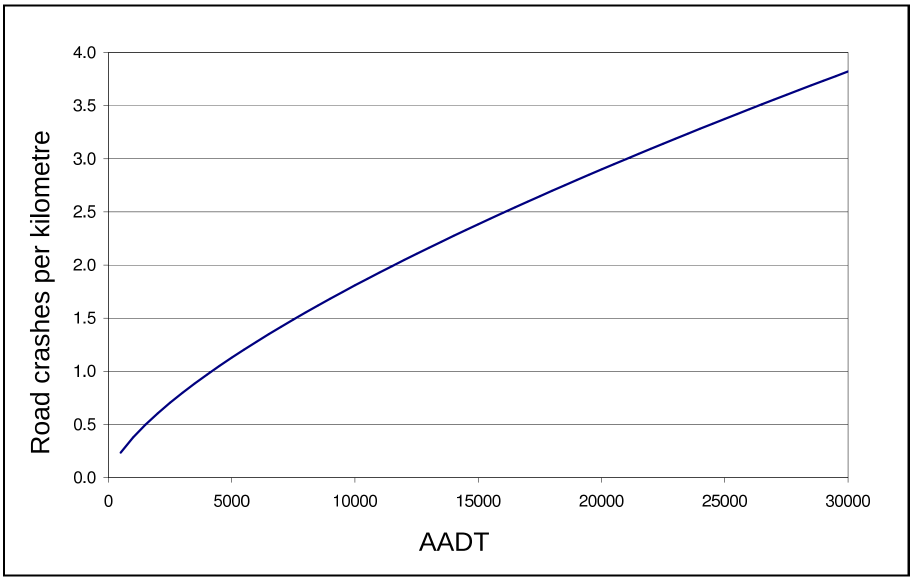

Figure 5.4. *The number of road crashes per kilometre per year for carriageways of urban single carriageway roads, with a speed limit of 50 km/h, two lanes and two driving directions.*

5.4. Rural carriageways with a speed limit of 80 km/h and one driving direction

The database contains only 16 rural carriageways with a speed limit of 80 km/h and one driving direction. This is not enough to fit a reliable model. Only the number of crashes per kilometre per year is computed, for the AADT classes introduced in *Chapter 2*. Three of these classes are empty. The result is given in *Figure 5.5*.

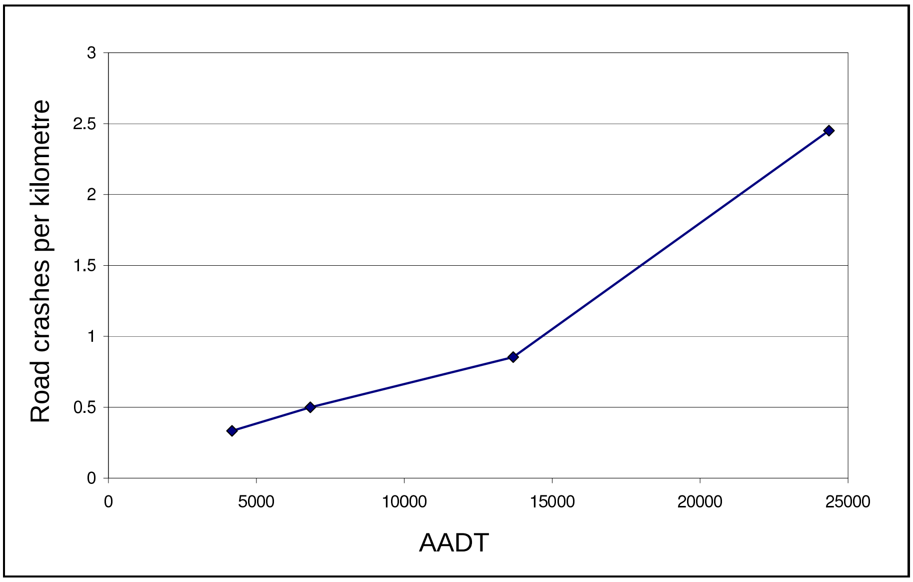

Figure 5.5. *The number of road crashes per kilometre per year in each AADT class for rural carriageways with a speed limit of 80 km/h and one driving direction.*

5.5. Rural carriageways with a speed limit of 80 km/h and two driving directions

The database only contains 38 rural carriageways with a speed limit of 80 km/h and two driving directions. This is not enough to fit a reliable model. Only the number of crashes per kilometre per year is computed, for the AADT classes introduced in *Chapter 2*. The three highest classes are empty. The result is given in *Figure 5.6*.

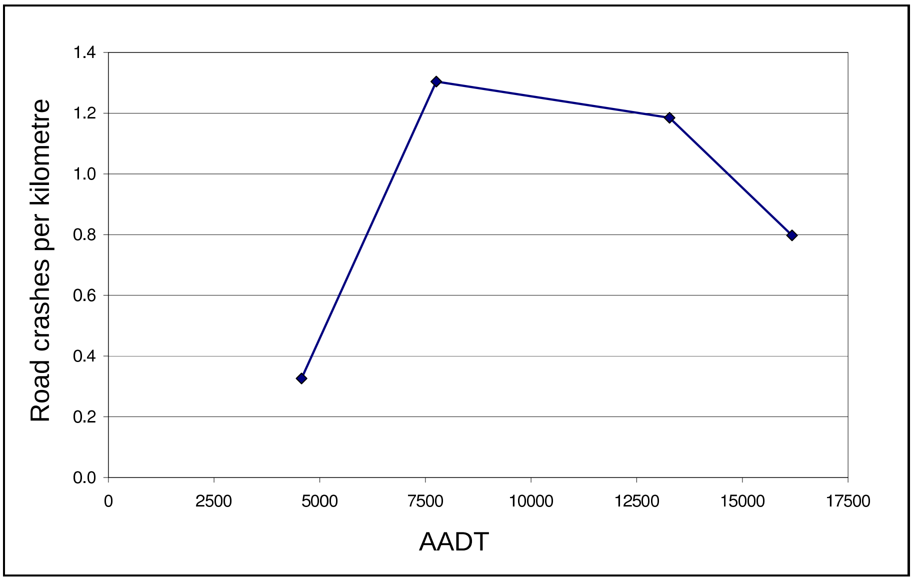

Figure 5.6. *The number of road crashes per kilometre per year in each AADT class for rural carriageways with a speed limit of 80 km/h and two driving directions.*

5.6. Discussion

In *Figure 5.7* the graphs of the models in this chapter are combined to make comparisons easier. For the carriageways for which a model was not developed, the corresponding values in *Table 2.1* are added to the graph. From this graph it follows that for low AADTs the crash rate of carriageways with a speed limit of 50 km/h is the same for both the single and dual carriageway types. For high AADTs the carriageways with only one driving direction are much safer than carriageways with two driving directions. For rural carriageways with a speed limit of 80 km/h also carriageways with only one driving direction are safer than carriageways with two directions. It can also be seen that roads with a speed limit of 80 km/h with a large traffic flow (above 18,000 motorverhicles per day) have two carriageways with one driving directions each. Based on the points for carriageways with a 70 km/h and 60 km/h it can be concluded that inside urban areas carriageways with a speed limit of 70 km/h are safer than those with a speed limit of 50 km/h and that outside urban areas the crash rates of carriageways with a speed limit of 60 km/h and 80 km/h with two driving directions are not very different.

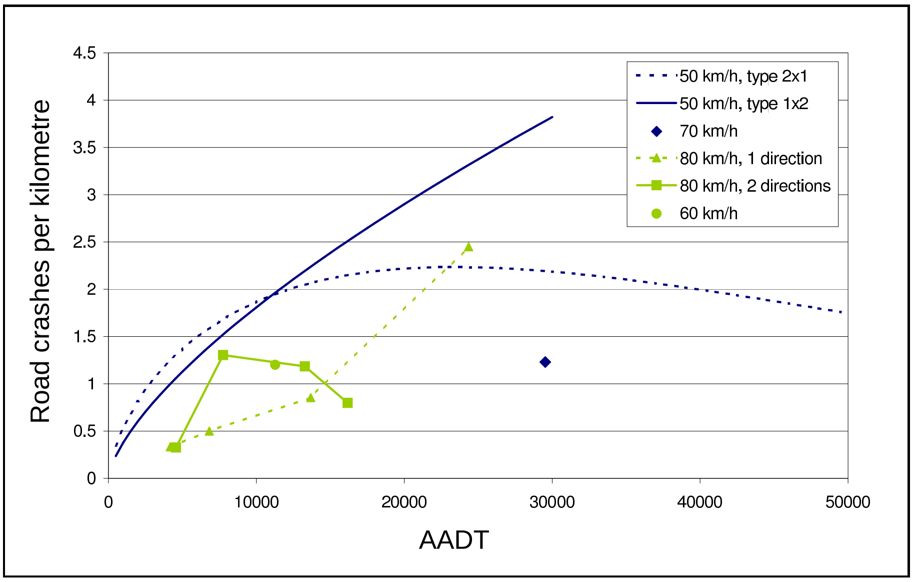

Figure 5.7. *The number of road crashes per kilometre per year for four different carriageway types.*

6. Conclusions and recommendations

6.1. The structure of the models

In this study two model structures were tried, namely

$$\mu_i = e^{\alpha} \cdot AADT_i^{\beta_1} \cdot L_i^{\beta_2} \tag{6.1}$$

and

$$\mu_i = \beta_0 \cdot AADT_i^{\beta_1} \cdot L_i^{\beta_2} \cdot e^{\beta_3 \cdot \frac{AADT_i}{1000}}. \tag{6.2}$$

Model *(6.1)* was directly based on APMs found in the literature. However, in *Chapter 2* it was showed that this form is not appropriate for the datasets containing all urban or all rural carriageways. Hence also models of the form *(6.2)* were developed for these sets. It turned out that these models did not only have the desired structure, but adding the variable $AADT/1000$ was indeed an improvement of the models for urban carriageways. In *Chapter 5* other road types were considered and it followed that for the carriageways of single carriageway roads, with a speed limit of 50 km/h the simple model type *6.1* was suitable.

This illustrates that it should always be checked if a particular model type is appropriate for the data, instead of just fitting a model as described in the literature. Therefore plots of the form of *Figures 2.1* and *2.2* are advized to get an idea of the most appropriate model type.

6.2. Modelling technique

In this report generalized linear modelling techniques are used to develop the accident prediction models. This what is called GLM is widely accepted for application in road safety modelling. At first a model can be fitted based on the assumption that the number of road crashes is Poisson distributed, which is a reasonable assumption. However, in practice it is often the case that such a model is affected by overdispersion, meaning that the variance exceeds the mean. If this is the case it is better to fit another model, now based on the assumption that the number of road crashes is negative binomially distributed. Another solution to the overdispersion problem is to use the quasi-likelihood method.

The technique based on the negative binomial distribution did have a positive side-effect. Besides solving the overdispersion problem, it also improved the behaviour of the standardized deviance residuals. These should be normally distributed, because the statistical tests performed on the parameter estimates are only reliable if this is the case.

6.3. Practical use

Road authorities can use accident prediction models to investigate the safety level of their roads. If they know the values of the explanatory variables (in this report the AADT and the length) for a particular carriageway, the APM can be used to compute the expected number of crashes on that carriageway. This computed value can be considered as an average number of crashes on a selection of roads for which the values of the explanatory variables are equal to those of the carriageway under consideration. So if this computed number is lower than the actual number of road crashes on that carriageway, the carriageway can be considered being too unsafe. To exclude the possibility that the numbers are different by coincidence, a statistical test should be used to test the significance of the difference. Wood (2005) gives 95%-confidence intervals for predictions made by a generalized linear model. For the prediction $\hat{y}_i$ the interval is given by

$$\left[0, \left\lfloor \hat{\mu}_i + \sqrt{19}\sqrt{\hat{\mu}_i^2 \text{Var}(\hat{\eta}_i) + \frac{\hat{\mu}_i^2 \text{Var}(\hat{\eta}_i) + \hat{\mu}_i^2}{\hat{\nu}} + \hat{\mu}_i} \right\rfloor \right],$$

where $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x. If an observed value of the number of crashes on a certain carriageway is outside this interval, then it deviates statistically significant from the predicted value. This means that the carriageway under investigation has a higher risk than expected on basis of the model. An explanation for this high risk can be found by using detailed information on road and traffic characteristics and road user behaviour. Because the left boundary of the confidence interval is 0, the conclusion can never be that a certain carriageway has a lower risk than expected.

Accident prediction models can also be used to compare different road types, as was done in this report. The following conclusions may be drawn from these comparisons:

- for $AADT \leq \pm 25000$ carriageways inside urban areas generally have a higher crash rate than carriageways outside urban areas, see *Figure 4.27*;
- carriageways with a speed limit of 50 km/h or 80 km/h and one driving direction have a lower crash rate than carriageways with a the same speed limit but with two driving directions;
- the average risk of urban carriageways with a speed limit of 70 km/h is lower than the crash rate of carriageways with a speed limit of 50 km/h and the average crash rate of rural carriageways with a speed limit of 60 km/h is almost the same as the crash rate of rural carriageways with a speed limit of 80 km/h and two driving directions.

The higher crash rate for carriageways inside urban areas can possibly be explained by the precense of pedestians and bicycles. However, in general it is known that roads inside urban areas have a lower crash rate than roads outside urban areas. The last conclusion also seems to be counterintuitive, because decreasing speed limits are expected to decrease the crash rate. However, this last conclusion only says that the carriageways in Haaglanden with lower speed limits do not necessarily have lower crash rates than carriageways with higher speed limits. It does not suggest that decreasing the speed limit increases the crash rate. For this type of conclusion before and afters studies are necessary. A perfectly logical explanation for this lower crash rate at higher speed limits is that the carriageways in the data collection only have a higher speed limit if it is safe to have such a limit. In other words, they are designed to be safe at a higher speed limit.

6.4. Further research

The research described in this report is a first attempt to develop accident prediction models for different road types. It was only possible to develop statistically significant models for all urban and all rural carriageways and for urban carriageways with a speed limit of 50 km/h and one or two driving directions. It is desirable to develop models for other, further disaggregated, road types. For this purpose more data is needed. Collecting the necessary data is very time consuming. It especially takes a long time to collect data about traffic volume, because for this purpose traffic has to be counted for a lengthy period of time. Some of the Dutch provinces have this data (due to permanent counting stations on their roads) and were prepared to share this data with the SWOV. At present, models for two different provincial road types (single and dual carriageways) are developed for two of the Dutch provinces: Noord-Holland and Gelderland. The models for Noord-Holland are of a different type than those for Haaglanden and Gelderland. Instead of taking the AADT as an explanatory variable, the amount of motorized traffic per hour is considered. The results for the two provinces will be reported by Janssen & Reurings (2007) and Reurings & Janssen (2007).

Future research can be conducted on datasets of other regions like, for example, provinces or municipalities. An interesting question which arises is whether or not similar models will be found for different regions. If this is the case, then models can be developed for all regions together. An advantage of this is that more data is available that can be used to developed the models on, so more road types can be considered. If no similar models are found, explanations must be found for the differences between the regions.

It is also interesting to develop accident prediction models for intersections. In this report intersections were not considered separately, they were considered to be a part of the carriageways. The crashes which happened on intersections were included in the number of crashes on the carriageway. By developing APMs for intersections of different types, the crash rates of these types can be compared. To make this possible, data is needed about both the major and minor traffic flows on each intersection and about several characteristics of the intersections. Especially this last information was not included in the Haaglanden database.

In the SWOV-project Infrastructure and Road Safety it was decided to develop models for different road types separately instead of developing one model for all roads together, including a lot of explanatory variables. The reason for this is that a large number of explanatory variables decreases the meaning and understanding of their estimated parameters, because the variables can be correlated. As a consequence, parameters may be in the opposite direction from the one safety engineers normally presume for those variables. This problem can be solved by developing models with only two variables (length and AADT) for different road types, as was done in this report. This has the disadvantage that a lot of data is needed. Harwood et al. (2000) solved the problem in a different way, they follow several steps:

- First a model is developed based on the extended data base HSIS, including several explanatory variables.
- Then fixed values for the explanatory variables (except the exposition) are entered in the model which results in the what is called base model.
- For other values of the explanatory variables the base model must be multiplied by the so-called accident modification factors. They represent the incremental effects of individual characteristics of roads and were developed by two expert panels based on their expert judgement and literature reviewing.
- The number of crashes on a road segment found in the previous step are entered, together with the crash site-specific crash history, in an Empirical Bayes procedure to get the predicted number of crashes on that road segment.

It may be interesting to investigate the usefulness of this procedure and especially the accident modification factors in the Netherlands.

References

Abbess, C., Jarrett, D. & Wright, C. (1981). *Accident at blackspots: estimating the effectiveness of remedial treatment, with special reference to the 'regression-to-mean' effect*. In: Traffic Engineering and Control, 22(10), p. 535–543.

Arbous, A. & Kerrich, J. (1951). *Accident statistics and the concept of accident-proneness*. In: Biometrics, 7(4), p. 340–432.

Harwood, D., Council, F., Hauer, E., Hughes, W. & Vogt, A. (2000). *Prediction of the expected safety performance of rural two-lane highways*. FHWA-RD-99-207, Office of Safety Research and Development, Federal Highway Administration, McLean, Virgina, USA.

Janssen, T. & Reurings, M. (2007). *Ongevallen en intensiteiten op provinciale wegen in Noord-Holland*. R-2006-20, SWOV, Leidschendam, The Netherlands.

Maycock, G. (s.a.). *Generalized linear models in the analysis of road accidents*. Transport and Road Research Laboratory, Crowthorne.

McCullagh, P. & Nelder, J. (1983). *Generalized linear models*. Chapman and Hall, Londen.

Newbold, E. (1927). *Practical applications of the statistics of repeated events particulary to industrial accidents*. In: Journal of the Royal Statistical Society, 90(3), p. 487–547.

Pierce, D. & Schafer, D. (1986). *Residuals in general linear models*. In: Journal of the American Statistical Association, 81(396), p. 977–986.

Pregibon, D. (1981). *Logistic regression diagnostics*. In: The Annals of Statistics, 9(4), p. 705–724.

Reurings, M. & Janssen, T. (2007). *Intensiteiten en ongevallen op provinciale wegen in Gelderland*. R-2006-21, SWOV, Leidschendam, The Netherlands.

Reurings, M., Janssen, T., Eenink, R., Elvik, R., Cardoso, J. & Stefan, C. (2005). *Accident prediction models and road safety impact assessment: a state-of-the-art*. First deliverable of WP2 of RIPCORD-ISEREST.

Wood, G. (2005). *Confidence and prediction intervals for generalised linear accident models*. In: Accident Analysis and Prevention, 37, p. 267–273.

Appendix Generalized linear modelling

General theory

Before going into the theoretical background of generalized linear modelling, the used notation will be introduced. The capitals $U, X, Y, Z,$ with or without indices, are always stochastic variables. Their realisations will be denoted by $u, x, y, z,$ also possibly with indices. If a symbol is typeset in boldface then it is a vector, for example $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ is a stochastic vector consisting of n components. Finally, a circumflex indicates an estimate or a prediction, i.e., $\hat{y}$ is the predicted value of y.

Let $\mathbf{Y}$ be an n-dimensional stochastic vector with independently distributed components $Y_1, \dots, Y_n$ and let $\mathbf{y}$ be the vector with the observed values of $\mathbf{Y}$. The mean of $\mathbf{Y}$ is given by $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ where $\mu_i = \mathbb{E}(Y_i), i = 1, \dots, n.$ The mean $\boldsymbol{\mu}$ can be specified in terms of p variables of which the values are given by $\mathbf{x}_1, \dots, \mathbf{x}_p,$ i.e., the i-th element of $\mathbf{x}_j$ is the value of the j-th variable corresponding to μ_i. In case of ordinary linear regression Y_i is assumed to be normally distributed with mean μ_i and constant variance σ^2. Further, μ_i is supposed to be a linear combination of $x_{i1}, \dots, x_{ip}$:

$$\mu_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j, \quad i = 1, \dots, n, \tag{A.1}$$

for unknown parameters $\beta_0, \dots, \beta_p$. By allowing $\mathbf{x}_0$ be a vector with each entry equal to 1, *(A.1)* can be rewritten as

$$\boldsymbol{\mu} = \sum_{j=0}^{p} \mathbf{x}_j\beta_j \quad \text{or} \quad \mu_i = \sum_{j=0}^{p} x_{ij}\beta_j, \quad i = 1, \dots, n. \tag{A.2}$$

The parameters are estimated by means of the ordinary least squares method. From *(A.2)* it follows that

$$Y_i = \sum_{j=0}^{p} x_{ij}\beta_j + e_i,$$

where e_i is a normally distributed stochastic variable with mean 0 and variance σ^2 for $i = 1, \dots, n.$ The e_i's are called the error terms and can be estimated by the residuals, i.e.,

$$\hat{e}_i = y_i - \hat{y}_i,$$

where $\hat{y}_i$ is the value of Y_i predicted by the model.

If the components of $\mathbf{Y}$ are not normally distributed with constant variance or if the relation between the mean $\boldsymbol{\mu}$ and the explanatory variables is not a linear one, then use can be made of generalized linear modelling. Like the traditional linear models, these models consist of a linear predictor $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = \sum_{j=0}^{p} \mathbf{x}_j\beta_j.$$

However, the relation between the mean $\boldsymbol{\mu}$ of the dependent variable $\mathbf{Y}$ and the linear predictor $\boldsymbol{\eta}$ is not necessarily given by the equality $\boldsymbol{\mu} = \boldsymbol{\eta},$

but by $g(\mu_i) = \eta_i$ for $i = 1, \ldots, n$, where g is a monotone and differentiable map called the link function.

The parameters $\beta_0, \ldots, \beta_p$ of a generalized linear model are estimated with the maximum likelihood method, which coincides with the ordinary least squares method in case of normally distributed error terms. The maximum likelihood method involves maximizing the log likelihood function $l(\boldsymbol{\mu}; \mathbf{y})$ over $\beta_0, \ldots, \beta_p$ to get estimates for the parameters $\beta_0, \ldots, \beta_p$. The log likelihood function is given by

$$l(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} l_i(y_i; \mu_i) = \sum_{i=1}^{n} \log f_i(y_i; \mu_i),$$

where $f_i(y_i; \mu_i)$ is the distribution of Y_i given the parameter μ_i.

Generalized linear models require the distribution of Y_i to originate from an exponential family. The distribution f of a stochastic variable X is said to come from an exponential family if it has the following form:

$$f(x) = e^{\frac{x\theta - b(\theta)}{a(\varphi)} + c(x,\varphi)}.$$

Here a, b and c are the functions which determine the specific distribution. The function a is usually of the form

$$a(\varphi) = \frac{\varphi}{w},$$

where φ is constant over the observations and w is a known weight which can vary from observation to observation. In general $w = 1$. The mean and variance of a stochastic variable X which has f as its distribution function are

$$\mathbb{E}(X) = \mu = b'(\theta), \quad \mathbb{V}\text{ar}(X) = \frac{b''(\theta)\varphi}{w}.$$

The variance of X can now be expressed in μ through the variance function $V(\cdot)$:

$$\mathbb{V}\text{ar}(X) = \frac{V(\mu)\varphi}{w}.$$

The parameter φ is called the dispersion parameter of the distribution.

The goodness-of-fit of a model can be measured in several ways. One way is by the scaled deviance, $D^*(\mathbf{y}, \hat{\boldsymbol{\mu}})$, where $\hat{\boldsymbol{\mu}}$ is the value of $\boldsymbol{\mu}$ predicted by the model. The scaled deviance is equal to twice the difference between the maximum achievable value of the log likelihood function and the achieved value of this function by the model under consideration. So the scaled deviance is

$$D^*(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2(l(\hat{\boldsymbol{\mu}}; \mathbf{y}) - l(\mathbf{y}; \mathbf{y})).$$

$D^*(\mathbf{y}, \hat{\boldsymbol{\mu}})$ is called the scaled deviance because it is equal to the deviance $D(\mathbf{y}, \hat{\boldsymbol{\mu}})$ divided by the dispersion parameter. In formula

$$D^*(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \frac{D(\mathbf{y}, \hat{\boldsymbol{\mu}})}{\varphi}.$$

Another goodness-of-fit measure is Pearson's χ^2, which is defined as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

Analogously to the deviance the scaled Pearson's χ^2 is defined as Pearson's χ^2 divided by φ.

Under the assumption that the fitted model is true, the scaled versions of both the deviance and Pearson's χ^2 are approximately χ^2_{n-p-1} distributed. For this reason the scaled deviance and the scaled Pearson's χ^2 can be used to estimate the dispersion parameter φ in case its value is unknown. Indeed, the mean of D^* is equal to $n-p-1$ because D^* is χ^2_{n-p-1} distributed. Setting $D^* = n-p-1$ and solving $D^* = D/\varphi$ for φ, leads to

$$\hat{\varphi} = \frac{D}{n-p-1}.$$

A similar expression can be deduced for Pearson's χ^2. If many of the predicted expected values of $\mathbf{Y}$ are smaller than 1, then it is possible that the scaled deviance is much smaller than the number of degrees of freedom. Hence the scaled deviance can not be used as a measure for the goodness-of-fit of the model. In this case Pearson's χ^2 can be used. Pierce & Schafer (1986) tried to answer the question why Pearson's χ^2 often has a distribution closer to χ^2 than the deviance, although the deviance residuals, which will be discussed later, perform excellently. They pointed out that having a more nearly χ^2 distribution is not directly connected to being the better measure of overall goodness-of-fit. Hence they suggest that the deviance should provide a better basis for goodness-of-fit tests than Pearson's χ^2 in spite of common assertions to the contrary.

According to McCullagh & Nelder (1983) the difference between the deviances of nested models can be approximated by a χ^2distribution better than the deviances themselves, even if $\hat{\mu}_i < 1$ for many values of i. This can be used to test if an extended model has a better fit than the more simple model. To show this, suppose that M_1 is a model with p explanatory variables and that M_2 is an extended model with $p+q$ explanatory variables. The question is whether or not M_2 has a better fit than M_1. The difference of deviances $D_1 - D_2$, where D_i is the deviance of model M_i, is approximately χ^2_q distributed under the assumption that M_1 is at least as good as M_2. Hence if the probability

$$\mathbb{P}(X \geq D_1 - D_2),$$

under the assumption that M_1 is at least as good as M_2, is smaller than the confidence level α (equivalent, if $D_1 - D_2$ is larger than the critical value corresponding to α), the null hypothesis can be rejected and hence the conclusion is that M_2 is significantly better than M_1.

In ordinary linear regression the residuals can be used to check the assumptions of the model. They should be independent normally distributed variables with constant variance. If these conditions are not satisfied, the conclusions about the statistical significance of the estimated parameters might be too optimistic. If generalized linear modelling is applied to the data, the residuals will not meet the three conditions mentioned above. However, in this case there is another type of residuals which behaves like the residuals in ordinary linear regression and can therefore be used to check the adequacy of the fit of the model. This type of residuals consists of the standardized deviance residuals which are defined as follows:

$$DR_i = \frac{\sqrt{d_i}(\text{sign}(y_i - \hat{\mu}_i))}{\sqrt{\hat{\varphi}(1-h_i)}}, \tag{A.3}$$

where d_i is the contribution of the i-th observation to the deviance and h_i is the i-th diagonal entry of the matrix

$$W^{\frac{1}{2}}X(X^TWX)^{-1}X^TW^{\frac{1}{2}},$$

with $W = \text{diag}(\hat{\mu}_1, \ldots, \hat{\mu}_n)$ and X is the design matrix. The numerators of *(A.3)* are called the deviance residuals. The division by $\sqrt{\hat{\varphi}(1 - h_i)}$ causes the resulting $DR_1, \ldots, DR_n$ to have constant variance. If the fitted model is correct, the (standardized) deviance residuals are approximately normally distributed. An example of the application of deviance residuals in generalized linear models is given by Pregibon (1981).

In the next section the application of generalized linear modelling to road crashes is described.

The application to road crashes

The Poisson distribution

In general, road crashes are considered to be Poisson distributed, which means that the number of road crashes is described by a Poisson distributed stochastic variable. The idea behind this thought is as follows. If for example a car enters an intersection, there are two possibilities: either a crash does occur or it does not. This can be described by the stochastic variable X which satisfies

$$X = \begin{cases} 1 \text{ (an accident occurs),} & \text{with probability } p, \\ 0 \text{ (an accident does not occur),} & \text{with probability } 1 - p. \end{cases}$$

The variable Z, defined as the number of 1's in N samples, is binomially distributed, so

$$\mathbb{P}(Z = z) = \binom{N}{z} p^z (1-p)^{N-z}, \quad z = 0, 1, \ldots, N.$$

If N is very large and p is very small, which holds for road crashes, the binomial distribution is approximately a Poisson distribution with parameter $\lambda = N \cdot p$. In formula,

$$\mathbb{P}(Z = z) = \binom{N}{z} \left(\frac{\lambda}{N}\right)^z \left(1 - \frac{\lambda}{N}\right)^{N-z} \longrightarrow \frac{\lambda^z}{z!} e^{-\lambda} \text{ if } N \to \infty.$$

This is indeed the density function of the Poisson distribution. The Poisson distribution comes from the exponential family with $\varphi = 1$, $\theta = \log \lambda$, $b(\theta) = e^\theta$, $c(z, \theta) = -\log(z!)$ and $V(\mu) = \mu$. Because $\varphi = 1$, the deviance and Pearson's χ^2 corresponding to a model based on Poisson distributed variables are equal to their scaled versions.

Now consider a collection of n road segments or n intersections. The number of road crashes on the i-th segment or intersection in a certain fixed period is described by the Poisson distributed stochastic variable Y_i. The relation between μ_i, the expected value of Y_i, and the explanatory variables is

$$\log(\mu_i) = \sum_{i=0}^{p} x_{ij}\beta_j, \quad i = 1, \ldots, n.$$

The parameters $\beta_0, \dots, \beta_p$ are estimated with the maximum likelihood estimates, which are determined by maximizing the following log likelihood function over $\beta_0, \dots, \beta_p$:

$$l(\mathbf{y}; \boldsymbol{\mu}) = \sum_{i=1}^{n} (y_i \cdot \log(\mu_i) - \mu_i).$$

As mentioned earlier, $\varphi = 1$ for the Poisson distribution. However, in practice it is often the case that the deviance and Pearson's χ^2, under the assumption that the number of road crashes is Poisson distributed, are much larger than the number of degrees of freedom, which means that $\varphi > 1$. This phenomenon is called overdispersion, because in this case the variance is larger than expected on the basis of the chosen distribution. Hence, the hypothesis that the estimated model (based on the Poisson distribution) is the right one, is rejected if one of the probabilities

$$\mathbb{P}(X \geq D(\mathbf{y}, \hat{\boldsymbol{\mu}})), \quad \mathbb{P}(X \geq \text{Pearson's } \chi^2),$$

where X is a χ^2_{n-p} distributed stochastic variable, is smaller than the chosen confidence level α.

It is also possible that the deviance (or Pearson's χ^2) is smaller than the number of degrees of freedom. This is called underdispersion. This is much less a problem than overdispersion, because it is not really a problem if there is less variance than expected. Therefore, a statistical test is not necessary. However, if the underdispersion is very high one should be careful, because then there is probably something wrong with the data.

The assumption that the number of road crashes is Poisson distributed is questionable if overdispersion is the case. In the next sections two alternatives will be discussed.

The Gamma en negative binomial distribution

One explanation for overdispersion is that the components of $\mathbf{Y}$ are indeed Poisson distributed, but that the means $\mu_1, \dots, \mu_n$ are not constant: they vary between the different road segments and/or intersections in an (imaginary) selection for which $\mathbf{x}_0, \dots, \mathbf{x}_p$ have the same values. Hence it can be assumed that the means of the Poisson distributions are stochastic variables themselves, e.g. $\Lambda_1, \dots, \Lambda_n$.

Generally nothing is known about the distribution of $\Lambda_1, \dots, \Lambda_n$. However, it is often assumed that they are Gamma distributed (Arbous & Kerrich, 1951; Newbold, 1927). Some theoretical and psychological background for this assumption is provided by Abbess, Jarrett & Wright (1981) and Maycock (s.a.). Under the assumption that Λ_i is Gamma distributed with $\mathbb{E}(\Lambda_i) = \mu_i$ and $\mathbb{V}\mathrm{ar}(\Lambda_i) = \nu$, the variable Y_i has a negative binomial distribution, so

$$\mathbb{P}(Y_i = y_i) = \frac{\Gamma(\nu + y_i)}{\Gamma(\nu) y_i!} \left(\frac{\nu}{\nu + \mu_i} \right)^{\nu} \left(\frac{\mu_i}{\nu + \mu_i} \right)^{y_i}$$

and

$$\mathbb{E}(Y_i) = \mu_i, \quad \mathbb{V}\mathrm{ar}(Y_i) = \mu_i + \frac{\mu_i^2}{\nu}.$$

This distribution also originates from the exponential family with $b(\theta) = -\nu \log(1 - e^\theta)$, $\theta = \log(\mu_i/(\nu + \mu_i)$ and $\alpha(\varphi) = 1$. The deviance and Pearson's χ^2 are equal to the scaled versions, similar to the Poisson distribution.

The model to be fitted is of the same type as the model based on the Poisson distribution. In other words, the link function for a model based on the negative binomial distribution is $g(x) = \log(x)$. The parameters $\beta_0, \ldots, \beta_p$ are estimated with the maximum likelihood method. The log likelihood function corresponding to the negative binomial distribution is

$$l(\mathbf{y}; \boldsymbol{\mu}) = \sum_{i=1}^{n} \log\left(\frac{\Gamma(\nu + y_i)}{\Gamma(\nu) y_i!}\right) - (y_i + \nu) \log\left(\frac{\nu + \mu_i}{\nu}\right) + y_i \log\left(\frac{\mu_i}{\nu}\right).$$

In the case of negative binomial distribution, not only the parameters $\beta_0, \ldots, \beta_p$ have to be estimated, but also the parameter ν. In the literature this is done by first developing a model based on the Poisson distribution and then estimating ν using the residuals of this model. The derived value of ν is then used to fit a model based on the negative binomial distribution. Now ν is again estimated, using the residuals of the new model, and again a new model is fitted. This procedures is repeated until the dispersion parameter is close enough to 1. In the GENMOD procedure of SAS the parameter ν can be estimated directly by maximising the log likelihood function and the described iteration is not necessary.

The assumption that the stochastic variables Λ_i is Gamma distributed, is quite strong. Although it leads to a nice distribution for Y_i it is not based on any theoretical or practical knowledge. A probably more insightful way to tackle the overdispersion problem is the quasi-likelihood method, which is discussed in the next section.

The quasi-likelihood method

An advantage of using the quasi-likelihood method as opposed to using the Gamma distribution is that it only requires an assumption about the variance of Y_i, not about the underlying distribution.

Assuming that the components of $\mathbf{Y}$ are independent and that the mean and variance of the components are given by

$$\mathbb{E}(Y_i) = \mu_i, \quad \mathbb{V}\text{ar}(Y_i) = \sigma^2 V_i(\mu_i),$$

where σ^2 is allowed to be unknown and V is a known function. The factor σ^2 is hence playing the role of the overdispersion parameter φ. The stochastic variables

$$U_i = \frac{Y_i - \mu_i}{\sigma^2 V_i(\mu_i)}$$

have the following properties:

$$\mathbb{E}(U_i) = 0, \quad \mathbb{V}\text{ar}(U_i) = \frac{1}{\sigma^2 V_i(\mu_i)}, \quad -\mathbb{E}\left(\frac{\partial U_i}{\partial \mu_i}\right) = \frac{1}{\sigma^2 V_i(\mu_i)}.$$

Because the asymptotical theory of likelihood functions is based on these properties, it is no surprise that the integral

$$Q_i(\mu_i; y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma^2 V_i(t)} dt$$

behaves like a log likelihood function for μ_i (when it exists). The integral Q_i is called the quasi-likelihood function for μ_i based on y_i. The quasi-likelihood for the complete vector $\mathbf{Y}$ based on the vector of observations $\mathbf{y}$ is

$$Q(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{n} Q_i(\mu_i; y_i).$$

For many variance functions V the sum of integrals Q coincides with the log likelihood function of a known distribution. The parameters $\beta_0, \ldots, \beta_p$ can hence be estimated by maximizing Q. Two possible estimates for the unknown σ^2 are the deviance and Pearson's χ^2, both divided by the number of degrees of freedom.

In case of the number of road crashes it is assumed that $\mathbb{V}\text{ar}(Y_i) = \sigma^2 \mu_i$, i.e., that $V_i(\mu_i) = \mu_i$ for $i = 1, \ldots, n$. The quasi-likelihood for μ_i based on y_i is then

$$Q_i(\mu_i; y_i) = y_i \cdot \log(\mu_i) - \mu_i.$$

This is exactly the log likelihood function for the Poisson distribution. So maximising Q gives exactly the same values of $\beta_0, \ldots, \beta_p$ as in the standard Poisson model. The expected variances and scaled deviance are different: the variances are now a factor σ^2 larger and the scaled deviance is a factor σ smaller.