ELSEVIER

Accident Analysis and Prevention 37 (2005) 591–600

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes

F.D. Bijleveld*

SWOV Institute for road safety research, PO Box 1090, 2260 BB Leidschendam, The Netherlands

Received in revised form 20 January 2005; accepted 22 January 2005

Abstract

In this study some statistical issues involved in the simultaneous analysis of accident related outcomes of the road traffic process are investigated. Since accident related outcomes like the number of victims, fatalities or accidents show interdependencies, their simultaneous analysis requires that these interdependencies are taken into account. One particular interdependency is the number of fatal accidents that is always smaller than the number of fatalities as at least one fatality results from a fatal accident. More generally, when the number of accidents increases, the number of people injured as a result of these accidents will also increase. Since dependencies between accident related outcomes are reflected in the variance-covariance structure of the outcomes, the main focus of the present study is on establishing this structure.

As this study shows it is possible to derive relatively simple expressions for estimates of the variances and covariances of (logarithms of) accidents and victim counts.

One example reveals a substantial effect of the inclusion of covariance terms in the estimation of a confidence region of a mortality rate.

The accuracy of the estimated variance-covariance structure of the accident related outcomes is evaluated using samples of real life accident data from The Netherlands. Additionally, the effect of small expected counts on the variance estimate of the logarithm of the counts is investigated.

Keywords: Multivariate model; Covariance; Likelihood; Accident counts; Logarithm; Overdispersion

1. Introduction

1.1. The need for multivariate modelling of influences on road safety

The development of (road) traffic safety is very often analysed studying the development of only one single accident related outcome. However, in practice there are always several accident related outcomes: the number of accidents themselves, the number of fatalities, the number of people injured, the cost of the material damage, and so on. Road safety measures usually do not have the same (quantitative) effect on each of these accident outcomes. For example, it is likely that the compulsory use of seat belts mainly has an effect on the consequences of an accident, whereas measures aiming to reduce the occurrence of drink-driving mainly have an effect on the number of accidents (and therefore on the number of victims). Speed reducing measures are supposed to have an effect on both the number of accidents and the consequences of an accident. Particular theories, such as, for example, the risk homeostasis theory (Wilde, 1994) and the zero risk theory (Summala and Näätäänen, 1988) state that theoretically likely developments may be counteracted because of behavioural adaptation. An example of this is the use of seat belts which may, according to some theories, result in higher speeds and other more dangerous behaviour, so that the expected reduction in the number of injuries is (at least partly) undone by the fact that the number of accidents increases.

In order to get a better understanding of the (quantitative) effect of road safety measures, their potentially differentiated effect on each of the accident related outcomes should be investigated. This becomes more important when the effects

* Tel.: +33 70 317 3333; fax: +33 70 320 1261.
E-mail address: Frits.Bijleveld@Swov.nl.

0001-4575/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2005.01.004

of different road safety measures introduced in a brief period of time are to be decomposed. In principle this can often be done by careful definition of dependent variables using separate univariate models, but in many cases a multivariate framework, where the dependence between the outcome variables is acknowledged is likely to be preferable.

The multivariate approach allows for the simultaneous estimation of unknown quantities based on all relevant data, rather than one estimate per dependent variable. This feature will become more important as new statistical techniques become practical that allow for unobserved-latent-components. One example is a multivariate extension of Harvey and Durbin (1986), other applications include methods like Schafer (1987), that implements an errors-in-variables approach by means of the EM-algorithm (Dempster et al., 1977) to generalized linear models by "casting the true covariates as 'missing data' ". For instance alternative methods exist that estimate a sufficient statistic for the explanatory variables. One example of an application in traffic safety is Johansson (1996) in which exposure is modelled by means of a latent variable that is estimated by means of one dependent variable, not all dependent variables simultaneously. This subject is discussed in more detail in Section 4.2.

1.2. The issue of dependence among outcomes

One important issue is that multivariate road safety outcomes may not be independent. A notable example is the fact that no more fatal accidents can occur in a period of time than the total number of fatalities in that period of time, as at least one fatality occurs in a fatal accident. A more general example is that when more road accidents occur in a year than usual, it is likely that more people get injured in road traffic as well. The former restriction is not imposed in this study, rather an approximation is made based on the latter aspect by developing an expression for the covariance (matrix) of the counts of accident related outcomes (and logarithms thereof). In some cases it will be possible to redefine the problem to a problem with independent road safety outcomes. This will however only simplify matters as far as the off-diagonal elements of the covariance matrix are concerned. The diagonal elements still have to be estimated in which the covariance matrix is implicitly used. In some cases the use of multivariate models will be inevitable.

Ignoring the covariance may have serious consequences in inference. Suppose the differentiated effect of a safety measure is to be evaluated on two different outcome variables: the annual number of accidents and the annual number of injured people. Also suppose that two models (A and B, say) are fitted on the data. For a certain observation (year) it is found that model A overestimates the observed number of accidents as well as the observed number of injured by say 10. On the other hand, for the same year model B overestimates the observed number of accidents also by an amount of 10, but underestimates the number of injured by an amount of ten. Expressed in terms of 'fit' (based on the assumption that the errors follow a symmetrical but not necessarily normal distribution) both models are equally likely when the covariance between the two outcomes is ignored. However, once the positive covariance between the two outcome variables is taken into account, model A yields a better fit than model B, as it should. Even worse, if model B overestimates the observed number of accidents by an amount of ten, but underestimates the number of injured by an amount of five, then it would yield a better fit than model A if the covariance between the two outcome variables is ignored. This could possibly result in false conclusions concerning the differentiated effectiveness of safety measures.

Thus, in the multivariate analysis (in the sense that multiple outcome variables are analysed) of road safety the dependencies between the dependent variables should be taken into account.

1.3. An approximating solution

In the following sections, an analytical procedure is proposed for the estimation of the variance-covariance matrix from accident data that can be used in models based on these moments, such as normal approximations, which includes the vast majority of available multivariate models in which multiple outcome variables are analysed.

Some multivariate models for multiple count variables do exist, see Cameron and Trivedi (1998, Chapter 8). Cameron and Trivedi (1998, p. 252) however state that "applications of multivariate count models are relatively uncommon. Practical experience has been restricted to some special computationally tractable cases."

For this reason and the fact that accident related outcomes are not restricted to count data, it is attempted in this paper to define a generally applicable, albeit approximate, method for analysing multivariate accident related data that may work in less tractable cases. To that end, estimates of the mean and covariance of those data are developed. It is intended that (possibly derivatives of) these estimates are used in weighted models based on the normal distribution. The methods will therefore not be suitable for observations based on a small number of accidents. With respect to the estimates derived in this paper the results in this paper are more extensive than developed in Evans (2003, paragraph 3 and appendices). This paper adds expressions for covariance between outcomes as well as a framework for developing estimates of higher moments, when needed.

It should be noted that this method differs from methods, for instance in time series analysis, in which (auto)covariances (of errors) are estimated from within a model, using aggregated data. The proposed method is different in that it estimates covariances from within the accident data, using individual accident information on the relevant outcomes, for instance the number of victims and fatalities. Note that in road safety it is rare to have information on non-injured persons, except for vehicle drivers. In principle it is possible to decide when the driver of a

vehicle is not among the victims and the vehicle was not parked that the driver probably is a non-injured person. Due to the possible complications, this possibility is ignored in this study.

Furthermore, only covariances within an observation are estimated in contrast to for instance time series analysis in which covariances between observations are estimated.

Recently Hutchings et al. (2003) published a method based on generalized estimating equations that allows for the estimation of covariances. The estimation of the covariances is in this case from within a model, but uses individual accident records. This approach used information on non-injured car occupants that is not available in many cases including the current study, where only victims and drivers are registered.

All theory in this study is based on the assumption that accident counts follow a Poisson distribution. At first this seems to be in conflict with modern theories on generalized Poisson modelling. However, this is not the case. The Poisson assumption is supported by limit theorems such as Feller (1968, p. 282), of which a less general version can be found in McCullagh and Nelder (1989, p. 105), concerning the asymptotic distribution of the sum of n (n large) independent Bernoulli trials[1] with variable (but quite small) probabilities of success. If each Bernoulli trial is equivalent to an encounter in road traffic with a unique but small probability of an accident, the distribution of the total number of accidents will tend to the Poisson distribution. This result however neglects the impact the small number of accidents can have on the large number of encounters.

In practice it is impossible to perform true replications, so the assumption that accident counts follow a Poisson distribution cannot be verified by means of an experiment. Observations that are *assumed* to be replications in practice produce larger variation than can be explained by the Poisson distribution. This is by no means a disproof of the assumption. More on causes of this overdispersion phenomenon can be found in Hauer (2001). In these cases however, it is assumed that overdispersion is mostly a *modelling* issue rather that a data issue. As described in Section 4.2 this approach of overdispersion can also be taken using this study.

In contrast to the strict assumptions on the accident counts, the assumptions on the distribution of the outcomes are rather relaxed. It is assumed that the moments of the variables that are a consequence of the accident (e.g. the number of victims, cost of damage) is finite. This assumption will in practice always be met as in practice damage is limited. Additionally it is assumed that the outcomes are independently and identically distributed for all accidents. It is likely that the results can be extended to not identically distributed outcomes.

The fact that the proposed method uses information on individual accidents will prohibit its use in cases where only aggregated accident information is available.

[1] A Bernoulli trial with parameter p is an experiment with probability p of 'success' (unfortunately an *accident* in this case) and probability $1 - p$ of 'failure' (no accident).

1.4. Overview of the paper

Section 2 describes and discusses the results of an analytical derivation of the covariance between the total number of accidents (not necessarily injury accidents) and the total number of victims in a time period. The latter could also have been the total cost of damages or any other measurable quantity resulting as a consequence of an accident. The analytical results can be used to derive higher order moments than covariances as well.

The analytical results are compared with results based on simulation in Section 3. This is done using samples from real accident data from The Netherlands in the period 1980–1999. The purpose of this is to assess the accuracy of the variance-covariance estimates based on the analytical derivations. Simulation studies have been performed on four sets of accidents: the first set consists of fatal car-only accidents, in the second set injury only accidents are included as well, the third set consists of fatal accidents (not just car-only accidents) and in the fourth injury accidents are included.

In Section 4, some examples are discussed of (possible) applications of the methods.

2. The covariance structure of road safety related outcomes

2.1. Introduction

This section describes how an estimate of the covariance matrix of the number of (injury) accidents and victims can be computed. In the following 'number of victims' may be read as 'the cost of damage' or any other consequence of an accident, as long as this consequence is equally and independently distributed with finite moments for all accidents. Details on the derivations can be found in Appendix A.

2.2. Results

Using the results of the derivations reported in Appendix A, variance-covariance estimates are formulated between either the count variables or the logarithms of those count variables. As stated above, these results are applicable to any variable with finite moments that is the consequence of an accident.

Two cases are developed in this study: the total number of accidents, victims and fatalities and the logarithm of the total number of accidents, victims and fatalities.

Table 2 provides an overview of all results while Table 1 contains an explanation of abbreviations used in Table 2.

Table 1 shows that not all information is available in standard publications on accidents. This is indicated by a "*". Detailed sources on individual accidents are needed to get more precise estimates. In that case, probably the individual fatality counts f_i and victim counts v_i per accident will be available. In that case, for instance, the variance of the

Table 1
Abbreviations used in the derived equations for variances and covariances and estimates

Number	Realisation	Abbreviation
Accidents (acc)	n	n
Victims in accident i	v_i^*	
Fatalities in accident i	f_i^*	
Sum over all accidents of number	Estimate	Abbreviation
Victims (vic)	$\sum_{i=1}^{n} v_i$	Σv
Fatalities (fat)	$\sum_{i=1}^{n} f_i$	Σf
Sum over all accidents of the square of the number	Estimate	Abbreviation
Victims	$\sum_{i=1}^{n} v_i^{2\,*}$	Σv^2
Fatalities	$\sum_{i=1}^{n} f_i^{2\,*}$	Σf^2
Sum over all accidents of the cross product of the numbers	Estimate	Abbreviation
Victims and fatalities	$\sum_{i=1}^{n} v_i f_i\,^*$	Σfv

The quantities marked * are usually not available in aggregated accident data.

total number of victims can be computed as the sum of the squared victims counts as indicated in Table 2. It can be seen that the variance of such victim counts is generally larger than the variance of similar accident counts. The amount of 'extra' variance depends on the distribution of the number of victims per accident. When more victims tend to occur in certain types of accident the variance of the number of victims tends to be higher.

Table 2
Derived equations for variances and covariances and estimates

Results based on counts variance	Estimate	Equation
The total number of accidents	n	
The total number of victims[a]	Σv^2	(A.7)
The total number of fatalities	Σf^2	(A.7)
Covariance	Estimate	Equation
The total number of accidents and victims	Σv	(A.13)
The total number of accidents and fatalities	Σf	(A.13)
The total number of victims and fatalities	Σfv	(A.14)
Results based on logarithms of counts variance	Estimate	Equation
The total number of accidents	$1/n$	(A.16)
The total number of victims	$\Sigma v^2/(\Sigma v)^2$	(A.18)
The total number of fatalities	$\Sigma f^2/(\Sigma f)^2$	(A.18)
Covariance	Estimate	Equation
The total number of accidents and victims	$1/n$	(A.19)
The total number of accidents and fatalities	$1/n$	(A.19)
The total number of victims and fatalities	$\Sigma fv/(\Sigma v \times \Sigma f)$	(A.20)

[a] The variance of the total number of victims is up to about 50% higher than the total number of victims based on data used in the simulation study (Section 3).

3. Simulation studies

To assess the accuracy of the variance-covariance estimates in Section 2 a number of simulation studies were performed. From injury accidents that occurred in The Netherlands in the years 1980 through 1999 the number of victims and the number of fatalities were recorded for each individual accident as well as the month and year in which the accident occurred. Additional simulations were performed using accidents that only involved cars, using accidents that involved only fatal accidents and using accidents that involved exclusively fatal car-only accidents. All simulation studies were performed by selecting a random number of accident records with replacement from a specific month. The number of accidents to be selected was a random number sampled from a Poisson distribution with expected value equal to the number of accidents that actually occurred that particular month.

This scheme should produce a selection of accidents that could have occurred almost as likely as the selection of accidents that actually occurred. For each thus created sample the total number of accidents, victims and fatalities were computed, as well as the logarithms thereof. Covariances were computed using a large number of such samples.

Table 3 compares results of estimates based on simulations with the estimates in Table 2. The estimates based on simulations were computed as the sample variance-covariance matrix. Each sample consisted of the 50,000 simulated months. One sample of 50,000 was drawn for each of the 240 months in the range starting january 1980 through december 1999. For each month, sample estimates (e_{sample}) and computed estimates ($\hat{e}$) were compared by means of the difference measure $d = (e_{\text{sample}} - \hat{e})/e_{\text{sample}}$.

For each statistic in the first column of Table 3 this resulted in 240 difference measures. The mean values and standard deviations of those 240 difference measures are listed horizontally in Table 3 for each of the four accident selections. The mean values and standard deviations reflect the amount of similarity between the sample estimates and the computed estimates. A large departure from zero of a mean value indicates a systematic difference (bias) between both estimates whereas a relatively large standard deviation is an indication of inaccuracy of the estimate.

It should be noted that the sampling scheme implies a Poisson distribution of the number of accidents and therefore simulation checks on the estimation of the variance of the number of accidents cannot be used to check the variance estimate. As discussed in the introduction, without true replicates it is impossible to validate the Poisson assumption. True replicates would mean for instance months of traffic sites with *exactly* the same accident distribution. The entry 'var(acc)' in Table 3 is thus for reference only.

Table 3 shows that the two types of estimates are quite similar except in the case of log-fatalities, indicating no important difference between the sample statistics and the estimates proposed in this study in the other cases. Particularly the logarithmic case with a smaller number of acci-

Table 3
Means and standard deviations of the relative differences $(e_{\text{sample}} - \hat{e})/e_{\text{sample}}$ between simulation sample estimates of the measures 'e_{sample}' and computed estimates '$\hat{e}$', 50,000 simulations, based on accident data from 1980–1999 in The Netherlands

Measure	All accidents		All car accidents		Fatal accidents		Fatal car accidents	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
var(acc)	−0.0000	0.0064	−0.0004	0.0065	0.0004	0.0065	0.0003	0.0063
var(vic)	−0.0006	0.0062	−0.0002	0.0062	−0.0002	0.0065	0.0004	0.0063
var(fat)	−0.0001	0.0064	0.0004	0.0066	0.0003	0.0063	0.0004	0.0065
cov(acc, vic)	−0.0005	0.0067	−0.0005	0.0069	0.0000	0.0069	0.0004	0.0067
cov(acc, fat)	0.0006	0.0259	−0.0004	0.0352	0.0003	0.0066	0.0004	0.0066
cov(vic, fat)	0.0004	0.0201	0.0006	0.0208	0.0000	0.0068	0.0004	0.0068
var(log(acc))	0.0004	0.0064	0.0021	0.0066	0.0126	0.0078	0.0513	0.0292
var(log(vic))	−0.0003	0.0063	0.0023	0.0063	0.0116	0.0099	0.1516	0.1194
var(log(fat))	0.0046	0.0068	0.0687	0.0411	0.0118	0.0076	0.0684	0.0400
cov(log(acc), log(vic))	0.0000	0.0068	0.0026	0.0070	0.0165	0.0087	0.1068	0.0724
cov(log(acc), log(fat))	0.0018	0.0262	0.0164	0.0382	0.0134	0.0079	0.0669	0.0367
cov(log(vic), log(fat))	0.0005	0.0203	0.0123	0.0234	0.0149	0.0083	0.1165	0.0766

Abbreviations: var(*), variance of *; cov(*,#), covariance of * and #; acc, number of accidents; vic, number of victims; fat, number of fatalities; log(*), logarithm of *.

dents (near the bottom of Table 3) appears to be biased. This bias may be the result of the approximation of the logarithm used to obtain an estimate of the variance in these cases, as it does not occur for instance in the case of var(fat) in combination with relatively little fatalities. Apparently estimates of the variance may not be that accurate and care must be taken in case of logarithms in combination with small counts that these inaccuracies do not influence inferences too much. The relative error of the variance estimate of the logarithm of a Poisson distributed random variable is the subject of Section 4.3, which is concerned with this issue.

4. Examples

4.1. The mortality ratio

As a first application one can use derived statistics as the mortality ratio. The mortality ratio is the number of fatalities divided by the number of accidents and as such is an application of multivariate accident related outcomes. In this example, the number of hospitalized victims as well as the fatalities are used. The ratios are obtained by dividing the number of victims by the number of accidents resulting in hospitalized victims or worse. Obviously, each ratio is a nonlinear function of the respective number of victims and the number of accidents, which themselves are not independent. The textbook approximation method to this ratio (delta method, see for instance Rice (1995, Chapter 4.6)) is used to obtain the expected value and the variance of the ratio $Z = Y/X$ in Rice (1995, p. 153):

$$E(Z) \approx \frac{\mu_Y}{\mu_X} + \sigma_X^2 \frac{\mu_Y}{\mu_X^3} - \frac{\sigma_{XY}}{\mu_X^2} = \frac{\sum v_i}{n} + n\frac{\sum v_i}{n^3} - \frac{\sum v_i}{n^2} = \frac{\sum v_i}{n}.$$

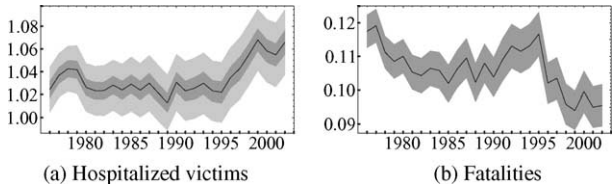
As a nice consequence but to no surprise it can be concluded that although the number of victims and the number of accidents is correlated, correction for it cancels out. The variance is approximated as in Rice (1995, p. 153)):

$$\begin{aligned}\mathrm{Var}(Z) &\approx \sigma_X^2 \frac{\mu_Y^2}{\mu_X^4} + \frac{\sigma_Y^2}{\mu_X^2} - 2\sigma_{XY}\frac{\mu_Y}{\mu_X^3} \\ &= n\frac{(\sum v_i)^2}{n^4} + \frac{\sum v_i^2}{n^2} - 2\left(\sum v_i\right)\frac{\sum v_i}{n^3} \\ &= \frac{\sum v_i^2}{n^2} - \frac{(\sum v_i)^2}{n^3}.\end{aligned}$$

In Fig. 1, the annual ratios for The Netherlands are plotted for the years 1976 through 2002. The dark gray areas denote the approximate 95% confidence regions based on the estimates from this study. The light gray areas (only visible in Fig. 1(a)) are 95% confidence regions ignoring covariance and using the mean estimate like the number of accidents as the variance estimate. The differences between the confidence regions are evident. The panels roughly resemble El-Sadig et al. (2002, Figs. 6 and 7) except that accidents with more serious outcomes as well as more seriously injured are counted in this case. El-Sadig et al. (2002, p. 472 and Discussion) and others also notice an increase in severity from traffic accidents. Although out of the scope of this study, it appears that the rates have suddenly changed in the late 1990's.

4.2. Multivariate state space modelling and the Kalman filter

One possible field of application of the proposed method is in applications of multivariate time series, for instance using state space models (Durbin and Koopman, 2001). As stated in the introduction, Johansson (1996, p. 75) acknowledged that the level of exposure is an unobservable variable and in that study it is modelled as a latent variable. This approach is similar to taken in Bijleveld et al. (in press) in which the



(a) Hospitalized victims (b) Fatalities

Fig. 1. The ratio between the number of hospitalized victims (a) and the number of fatalities (b) and the number of accidents with hospitalized victims or fatalities. Dark gray areas denote the approximated 95% confidence regions. Light gray areas (only visible in (a)) are 95% confidence regions ignoring covariance and estimating the variance of the number of victims like the number of accidents using the number of victims.

(extended) Kalman filter (Harvey and Durbin, 1986, p. 160) is used to model the development of the number of fatal accidents inside and outside urban area's in the Netherlands together with vehicle kilometres inside and outside urban area's. In that approach it appeared possible to reconstruct the vehicle kilometres inside and outside urban area's in years in which only the total (the sum of inside and outside urban area's) number of vehicle kilometres is available. The multivariate approach allows for estimation of the latent exposure in Johansson (1996, p. 75) for all dependent variables, not just on a per dependent variable basis. That is likely to yield different exposure developments per dependent variable which may not be optimal.

If counts are modelled linearly in state space models, an additive approach to overdispersion will be taken. This is due to the way variances are handled in the Kalman filtering approach. This can be seen from (Durbin and Koopman, 2001, p 66, (4.7)) where the *prediction error covariance* is decomposed into a part based on the *system* error (the model part) and an *observation* part, the latter composed from the (co)variances of the observation errors, which this study serves. This approach is very much along the lines of (Cox, 1983, Generalization (b), pp. 272–273).

4.3. The relative error of the variance estimate of the logarithm of a Poisson distributed random variable

In Section 3, it was found that the approximation of the variance of the logarithm of counts performs poorly when counts are small. In order to obtain a better insight into this matter this section is devoted to estimating the error in approximation. To this end, the theoretical variance of the logarithm of a Poisson distributed number as a function of its expected value λ is computed numerically and compared to the estimate $1/\lambda$. Table 4 lists the numerical approximation (s^2) of the theoretical variance of the logarithm of a Poisson distributed number as well as the estimate $1/\lambda$ for that number over a range of small values of the expected value λ of the Poisson distributed number. In Fig. 2, the relative difference (approximation-estimate)/approximation is graphed as a function of the expected value λ. Obviously, both results for the logarithmic case are approximations, but the numerical approximation is much more precise, so all differences are attributed to the estimate. As Fig. 2 shows, the relative error for the variance of the number of accidents can be substantial if λ is less than about 20–30. Similar results will hold for victims and fatalities, which do not obey a Poisson law but are dependent on one.

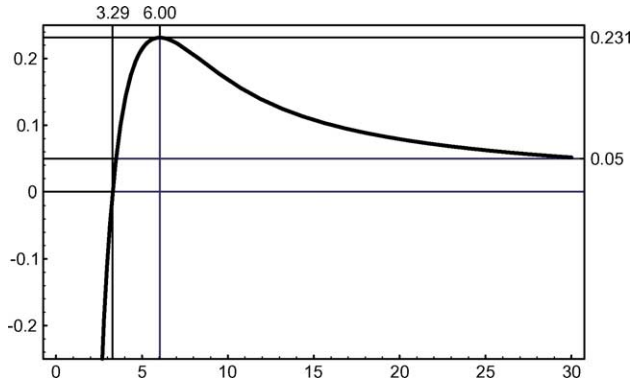


Fig. 2. The relative error of the variance estimate $1/\lambda$ of the logarithm of a Poisson distributed random variable with respect to a numerical approximation as a function of the expected value λ of the Poisson distributed number (horizontal axis). The relative error is computed as (approximation-estimate)/approximation.

5. Conclusions

In this study some statistical issues involved in the simultaneous analysis of accident related outcomes (such as the number of victims, fatalities or accidents and costs) of the

Table 4
Estimates and approximations of variances of the logarithm of the number of accidents when the expected number of accidents is small

λ	$1/\lambda$	$s^2(N_\lambda)$	λ	$1/\lambda$	$s^2(N_\lambda)$	λ	$1/\lambda$	$s^2(N_\lambda)$	λ	$1/\lambda$	$s^2(N_\lambda)$
1	1.0000	0.1343	11	0.0909	0.1072	21	0.0476	0.0515	31	0.0323	0.0340
2	0.5000	0.2631	12	0.0833	0.0967	22	0.0455	0.0490	32	0.0313	0.0328
3	0.3333	0.3037	13	0.0769	0.0881	23	0.0435	0.0467	33	0.0303	0.0318
4	0.2500	0.2898	14	0.0714	0.0808	24	0.0417	0.0446	34	0.0294	0.0308
5	0.2000	0.2546	15	0.0667	0.0747	25	0.0400	0.0427	35	0.0286	0.0299
6	0.1667	0.2168	16	0.0625	0.0695	26	0.0385	0.0409	36	0.0278	0.0290
7	0.1429	0.1840	17	0.0588	0.0649	27	0.0370	0.0393	37	0.0270	0.0282
8	0.1250	0.1574	18	0.0556	0.0610	28	0.0357	0.0378	38	0.0263	0.0274
9	0.1111	0.1366	19	0.0526	0.0574	29	0.0345	0.0364	39	0.0256	0.0267
10	0.1000	0.1202	20	0.0500	0.0543	30	0.0333	0.0351	40	0.0250	0.0260

See also Fig. 2.

road traffic process were investigated. The main focus of this study was on the estimation of the variance-covariance structure of such outcomes. Correction for covariance is needed in order to enhance the statistical reliability of techniques applied to the simultaneous analysis of accident related outcomes. It turns out to be possible to derive relatively simple expressions for the variances and covariances of (logarithms of) accidents and victim counts.

It is argued that when multiple accident outcomes are modelled, their covariance should be taken into account . One example reveals a substantial effect of the inclusion of covariance terms in the estimation of a confidence region of a mortality rate.

The variances and covariances were compared with estimates obtained in a simulation study. Not surprisingly, it was found in the logarithmic case that bias increases as the number of accidents decreases. In general, estimates will deteriorate when the number of accidents decreases.

As a special case it is recommended not to use Normal approximations to the Poisson distribution (Feller, 1968, Chapter VII) of the variance of, for instance, the number of victims in a year by estimating its value using the observed number of victims. The actual variance may be substantially larger. The amount of 'extra' variance depends on the distribution of the number of victims per accident. When more victims tend to occur in certain types of accident the variance of the number of victims tends to be higher. As a result this study confirms that is better to approximate this variance by the sum of the square of the number of victims per accident rather than by the sum of the number of victims per accident.

In order to compute the statistics, in some cases information at the level of individual accidents is needed. For instance the estimate of the variance of the number of fatalities is computed by summing the squares of the number of fatalities for each individual accident. This information may not always be available.

Appendix A. The covariance structure of accident related outcomes

Here it is shown how to derive an expression for the variance-covariance matrix of the number of (injury) accidents and victims. It is assumed that a basic simplification can be used: the number of victims per accident is equally distributed with finite variance for all accidents, although it may be possible to relax this assumption. No further assumptions on the shape of the distribution of the accident outcomes are made. The derivation extends the result in (Feller, 1968, p. 286).

A.1. The expected value and variance of the number of victims

First define N as the number of accidents in a certain period of time. N is assumed to be Poisson distributed with parameter λ.

Let the stochastic variables V_i $(i = 1, \ldots, N)$ denote the number of victims in accident i. The V_i are assumed to be independently identically distributed. The distribution of the V_i has characteristic function $\phi(t)$ and expected value μ and all moments are finite. The symbol v_i is used to denote a realisation of the number of victims in accident i.

Let the total number of victims be V, defined as $V = \sum_{i=1}^{N} V_i$, thus V is a sum over a Poisson distributed random number (N) of accidents. Defining $\Phi(t)$ as the characteristic function of the distribution of V, we have that

$$\Phi(t) = \mathrm{E}(\mathrm{e}^{itV}) = \mathrm{E}(\mathrm{E}(\mathrm{e}^{itV}|N)), \tag{A.1}$$

where i is the imaginary number ($i^2 = -1$). Since

$$\mathrm{E}(\mathrm{e}^{itV}|N = n) = \mathrm{E}(\mathrm{e}^{it\sum_{i=1}^{n} V_i}|N = n)$$

$$= \mathrm{E}\left(\prod_{i=1}^{n} \mathrm{e}^{itV_i}\right) = \prod_{i=1}^{n} \phi(t) = \phi^n(t) \tag{A.2}$$

then substituting (A.2) in (A.1) and because N follows a Poisson distribution, we get

$$\Phi(t) = \mathrm{E}(\phi^N(t)) = \mathrm{e}^{-\lambda} \sum_{n=0}^{\infty} \frac{\phi^n(t)\lambda^n}{n!} = \mathrm{e}^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda\phi(t))^n}{n!}.$$

Since $\mathrm{e}^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ (see Feller, 1968, p. 286) we obtain

$$\Phi(t) = \mathrm{e}^{-\lambda+(\lambda\phi(t))} = \mathrm{e}^{\lambda(\phi(t)-1)}. \tag{A.3}$$

As $\mathrm{E}(|V|^3)$ exists and is finite, $\mathrm{E}(V) = i^{-1}\Phi'(0)$ and $\mathrm{E}(V^2) = -\Phi''(0)$. Because $\phi(0) = 1$, $\Phi(0) = 1$, $\phi(0)' = i\,\mathrm{E}(V_k) = i\mu$ and $\phi(0)'' = -\mathrm{E}(V_k^2)$, we get the following expected value for the total number of victims V:

$$\mathrm{E}(V) = i^{-1}[\lambda\phi'(t)\Phi(t)]_{t=0} = i^{-1}\lambda\phi'(0) = \lambda\mu. \tag{A.4}$$

This quantity can be estimated using:

$$\mathrm{m}(V) = \sum_{i=1}^{N} v_i. \tag{A.5}$$

This estimator is unbiased since

$$\mathrm{E}(\mathrm{m}(V)) = \mathrm{E}\left(\mathrm{E}\left(\sum_{i=1}^{N} v_i|N\right)\right) = \mathrm{E}(N\mu) = \lambda\mu.$$

The variance of the total number of victims V is $\sigma^2(V) = \mathrm{E}(V^2) - \mathrm{E}^2(V)$. Because

$$\mathrm{E}(V^2) = -[\lambda\phi''(t)\Phi(t) + (\lambda\phi'(t))^2\Phi(t)]_{t=0}$$

$$= \lambda\mu. - \lambda\phi''(0) - (\lambda\phi'(0))^2$$

we obtain

$$\sigma^2(V) = \lambda\mathrm{E}(V_k^2). \tag{A.6}$$

This can be estimated using

$$s^2(V) = \sum_{i=1}^{N} v_i^2. \tag{A.7}$$

Again, this estimator is unbiased since

$$E(s^2(V)) = E\left(E\left(\sum_{i=1}^{N} v_i^2 | N\right)\right) = E(NE(V_k^2)) = \lambda E(V_k^2).$$

The expected value and the variance of the number of fatalities can be derived in the same way.

A.2. The covariance between the number of accidents and the number of victims

The covariance between the number of injury accidents and the number of victims is more complicated. Its derivation is based on the same characteristic function argument as used above. The characteristic function of the random vector (N, V) is defined as

$$\Phi(s,t) = E(e^{isN+itV}) \equiv E(f(N) \times g(V)).$$

Using the same property of conditional expectations

$$E(E(f(N) \times g(V)|N)) = E(f(N)E(g(V)|N))$$

then using (A.2) we obtain

$$\Phi(s,t) = E(e^{isN}\phi^N(t)) = \quad = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda\phi(t)e^{is})^k}{k!} =$$

$$= e^{-\lambda}e^{\lambda\phi(t)e^{is}} = e^{\lambda(\phi(t)e^{is}-1)}. \tag{A.8}$$

In order to derive the covariance, we have $E(N) = \lambda$ by the Poisson law of N and $E(V) = \lambda\mu$ is already available in (A.4). In order to complete the derivation of the covariance, we need to evaluate

$$E(N\,V) = -\left[\frac{\partial^2\Phi(s,t)}{\partial s\partial t}\right]_{s=t=0}. \tag{A.9}$$

The derivative of Φ with respect to s is

$$\frac{\partial\Phi(s,t)}{\partial s} = i\lambda\phi(t)\Phi(s,t)$$

and the derivative of the latter with respect to t

$$\frac{\partial^2\Phi(s,t)}{\partial s\partial t} = i\lambda\phi(t)\lambda e^{is}\phi'(t)\Phi(s,t) + i\lambda\phi'(t)\Phi(s,t) =$$

$$= i\lambda\Phi(s,t)\phi'(t)(\phi(t)\lambda e^{is} + 1).$$

Because $\Phi(0,0) = \phi(0) = 1$, and $\phi'(0) = i\mu$ it follows that

$$\left[\frac{\partial^2\Phi(s,t)}{\partial s\partial t}\right]_{s=t=0} = i^2\lambda\mu(\lambda+1).$$

Therefore

$$E(N\,V) = \lambda\mu(\lambda+1) \tag{A.10}$$

and thus

$$\mathrm{Cov}(N,V) = \lambda^2\mu + \lambda\mu - \lambda\lambda\mu = \lambda\mu. \tag{A.11}$$

This quantity can be estimated using

$$s(N,V) = \sum_{i=1}^{N} v_i. \tag{A.12}$$

Again, this estimator is unbiased because

$$E(s(N,V)) = E\left(E\left(\sum_{i=1}^{N} v_i|N\right)\right) = E(N\mu) = \lambda\mu. \tag{A.13}$$

The covariance between the number of accidents and the number of fatalities can be derived in the same manner.

A.3. The covariance between the number of victims and the number of fatalities

Let the random variable F_i be the number of fatalities in accident i. Let $F = \sum_{i=1}^{N} F_i$. Define $\Psi(s,t)$ as the characteristic function of the random vector (V, F) and $\psi(s,t)$ as the characteristic function of the random vector (V_i, F_i). Then, for each $i \neq j$, (V_i, F_i) is independent of (V_j, F_j). However, the V_i and F_i are not independent because $V_i \geq F_i$ a.s. Now

$$\Psi(s,t) = E(e^{itV+isF}) = E(E(e^{itV+isF}|N)).$$

Following a derivation similar to (A.2) we obtain:

$$E(e^{itV+isF}|N=n) = E(e^{\sum_{i=1}^{n}(itV_i+isF_i)}|N=n)$$

$$= E\left(\prod_{i=1}^{n} e^{itV_i+isF_i}\right)$$

$$= \prod_{i=1}^{n}\psi(s,t) = \psi^n(s,t).$$

Analogous to (A.8) it is found

$$\Psi(s,t) = E(\psi^N(s,t)) = e^{-\lambda}\sum_{n=0}^{\infty}\frac{\psi^n(s,t)\lambda^n}{n!}$$

$$= e^{-\lambda}\sum_{n=0}^{\infty}\frac{(\lambda\psi(s,t))^n}{n!},$$

from which it follows that

$$\Psi(s,t) = e^{-\lambda+(\lambda\psi(s,t))} = e^{\lambda(\psi(s,t)-1)}.$$

Using the same argument as in the derivation of (A.10) we obtain

$$\mathrm{E}(V\,F) = \lambda^2 \mathrm{E}(F_i)\mathrm{E}(V_i) + \lambda \mathrm{E}(F_i V_i),$$

and therefore

$$\mathrm{Cov}(V, F) = \lambda \mathrm{E}(F_i V_i).$$

This can be estimated with

$$\mathrm{s}(V, F) = \sum_{i=1}^{N} f_i v_i. \tag{A.14}$$

Again, this estimator is unbiased.

A.4. Derivation for the logarithm of counts

A.4.1. The expected value and variance of the logarithms of number of accidents and victims

Unfortunately, it is not possible to derive an explicit characteristic function as simple as the one in Eq. (A.3) in the case of the logarithm of the number of accidents and victims. For that reason, approximations need to be made in order to get a useful expression for the covariance between the logarithm of the number of accidents and victims. This is done using the 'delta' method. The basic idea is that the logarithms of N and V are approximated by a series expansion of order k (usually order one) about their expected values. This results in $\log(N)$ being approximated by a polynomial in N of order k, that is, $\log(N) \approx^k a_0 + a_1 \times (N-\lambda) + \ldots + a_k \times (N-\lambda)^k$.

In the present case, a first order approximation about the expected value (λ) of the number of accidents is:

$$\log(N) \overset{1}{\approx} \log(\lambda) + \frac{N-\lambda}{\lambda} \tag{A.15}$$

where $\approx^1$ means variance of the first order approximation. Thus, the expected value of this first order approximation is equal to $\log(\lambda)$: $\mathrm{E}(\log(N)) \approx^1 \log(\lambda)$. Similarly, the square of the linear approximation is

$$\frac{(N-\lambda)^2}{\lambda^2} + \log^2(\lambda) + 2\left(\frac{N-\lambda}{\lambda}\right)\log(\lambda).$$

The latter part has expected value 0, so its expected value is

$$\frac{\sigma^2(N)}{\lambda^2} + \log^2(\lambda) = \frac{1}{\lambda} + \log^2(\lambda),$$

so combining we have

$$\sigma^2(\log(N)) \approx \frac{\sigma^2(N)}{\lambda^2} = \frac{1}{\lambda}, \qquad \mathrm{s}^2(\log(N)) \approx \frac{1}{N} \tag{A.16}$$

In the case of $\log(V)$, approximations are about the expected value $\lambda\mu$ of V:

$$\log(V) \overset{1}{\approx} \log(\lambda\mu) + \frac{V-\lambda\mu}{\lambda\mu} \tag{A.17}$$

For that reason $\mathrm{E}(\log(V)) \approx^1 \log(\lambda\mu)$ and using first (A.16) and then (A.5) we obtain

$$\sigma^2(\log(V)) \approx \frac{\sigma^2(V)}{(\lambda\mu)^2} = \frac{\sigma^2(V)}{(E(V))^2}, \qquad \mathrm{s}^2(\log(V)) \approx \frac{\mathrm{s}^2(V)}{\mathrm{m}(V)^2}. \tag{A.18}$$

Results for fatalities are derived in a similar way.

A.4.2. The covariance between the logarithm of the number of accidents and the logarithm of the number of victims and fatalities

Extending the first order approximations of both log-accident counts and log-victims, it can be seen that, using (A.15) and (A.17)

$$\begin{aligned}
&\mathrm{Cov}(\log(N), \log(V)) \\
&\quad \approx \mathrm{Cov}\left(\log(\lambda) + \frac{N-\lambda}{\lambda}, \log(\lambda\mu) + \frac{V-\lambda\mu}{\lambda\mu}\right) \\
&\quad = \mathrm{E}\left(\frac{N-\lambda}{\lambda}\frac{V-\lambda\mu}{\lambda\mu}\right) \quad = \frac{\mathrm{Cov}(N, V)}{\lambda^2\mu} \\
&\quad \text{using } (A.11) := \frac{1}{\lambda} \qquad \mathrm{s}(\log(N), \log(V)) = \frac{1}{n}
\end{aligned} \tag{A.19}$$

Again, results for fatalities are derived in a similar way.

A.4.3. The covariance between the logarithm of the number of victims and the logarithm of the number of fatalities

In this case a similar approach can be taken:

$$\begin{aligned}
&\mathrm{Cov}(\log(V), \log(F)) \\
&\quad \approx \mathrm{Cov}\left(\log(\lambda\mu_V) + \frac{V-\lambda\mu_V}{\lambda\mu_V}, \log(\lambda\mu_F) + \frac{V-\lambda\mu_F}{\lambda\mu_F}\right) \\
&\quad = \mathrm{E}\left(\frac{V-\lambda\mu_V}{\lambda\mu_V}\frac{F-\lambda\mu_F}{\lambda\mu_F}\right) \quad = \frac{\mathrm{Cov}(V, F)}{\lambda^2\mu_V\mu_F} \\
&\quad \text{using } (A.14) := \mathrm{s}(\log(V), \log(F)) = \frac{\sum_{i=1}^{n} v_i f_i}{\sum_{i=1}^{n} v_i \sum_{i=1}^{n} f_i}
\end{aligned} \tag{A.20}$$

References

Bijleveld, F.D., Commandeur, J.J.F., Koopman, S.J., van Montfort, K., in press. Nonlinear time series aspects of road safety research. J. Transp. Econ. Policy.

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge.

Cox, D.R., 1983. Some remarks on overdispersion. Biometrika 70 (1), 269–270.

Dempster, A.P., Liard, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B (34) 183–202.

Durbin, J., Koopman, S.J., No. 24 in Oxford statistical science series 2001. Time Series Analysis by State Space Methods. Oxford University Press, Oxford.

El-Sadig, M., Norman, J.N., Lloyd, O.L., Bener, A., 2002. Road traffic accidents in the united arab emirates: trends of mobidity and mortality during 1977–1998. Acc. Anal. Preven. 34, 465–467.

Evans, A.W., 2003. Estimating transport fatality risk from past accident data. Acc. Anal. Prevent. 35, 459–472.

Feller, W., 1968. An Introduction to Probability Theory and its Applications, vol. I, third ed. John Wiley & Sons Inc., New York.

Harvey, A.C., Durbin, J., 1986. The effects of seat belt legislation on British road casualties: a case study in structural time series modelling. J. Roy. Stat. Soc. A 149 (3), 187–227.

Hauer, E., 2001. Overdispersion in modelling accidents on road sections and in Empirical Bayes estimation. Acc. Anal. Prevent. 33, 799–808.

Hutchings, C.B., Knight, S., Reading, J.C., 2003. The use of generalized estimating equations in the analysis of motor vehicle crash data. Acc. Anal. Prevent. 35, 3–8.

Johansson, P., 1996. Speed limitation motorway casualties: a time series count data regression approach. Acc. Anal. Prevent. 28 (1), 73–87.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed.. Chapman & Hall.

Rice, J.A., 1995. Mathematical Statistics and Data Analysis. JDuxberry Press, Belmont, CA.

Schafer, D.W., 1987. Covariate measurement error in generalized linear models. Biometrika 74 (2), 385–391.

Summala, H., Näätäänen, 1988. The zero-risk theory and overtaking decisions. In: Rothengatter, J.A., de Bruin, R.A. (Eds.), Road User Behaviour; Theory and research. van Gorkum, Assen/Maastricht, pp. 82–92.

Wilde, G.J.S., 1994. Target Risk. PDE Publications, Toronto.