LOGLINEAR AND LOGIT MODELS IN ACCIDENT RESEARCH

Ekkehard Brühning and Gabriele Ernst,
German Federal Highway Research Institute

1. INTRODUCTION

Data used in empirical research and, in particular, in traffic and accident research, very frequently do not have the (quantitative) interval scale format which is required by most statistical methods. Instead, they are merely ordinal or nominal scaled: measurements and observations are qualitative (e.g. age in age groups; sex; type of vehicle; accident severity - killed, seriously injured, slightly injured). The categories of such a nominal scaled variable only stand in the relation "not equal to". This means that we cannot make any statements about rank orderings or distances and that the range of mathematical operations allowed with such data is different from that available for quantitative data. Thus there is an obvious need for analysis methods which are suited to such data. A further problem constantly confronted in accident research is "complexity": i.e. problems are not simple enough to permit an explanation of some variable y in terms of one other variable x.

Analysis procedures for the relations between nominal scaled variables usually rely on contingency tables. But conclusions which were valid for cross tabulations of two variables often have to be considerably modified as soon as a third (control) variable is introduced. We then have to face the question: if we want control by a third variable, why not introduce a fourth for even more control, etc.? Why not make a multivariate analysis using all the variables that are considered to be theoretically relevant.

The subject of this contribution are statistical methods on the basis of loglinear and logit models which offer the possibility of multivariate analyses while depending only on realistic assumptions about the scale types of the variables.

2. THE PRINCIPLE OF LOGLINEAR MODELS

First, we will illustrate the principle of loglinear models with the example of a simple 2 x 2 contingency table.

Figure 1: Schema of a 2 x 2 contingency table

		B	
		1	2
A	1	y_{11}	y_{12}
	2	y_{21}	y_{22}

The loglinear model for the example of Figure 1 is:

$$\eta_{ij} = \ln\mu_{ij} = \beta_0 + \beta^{A}_{i} + \beta^{B}_{j} + \beta^{AB}_{ij} \qquad i,j = 1,2$$

The logarithm of the expected cell frequency μ_{ij} of each of the cells y_{ij} in the 2 x 2 table is made up additively of a general effect β_0 (often called the "grand mean"), the main effects β^{A}_{i} and β^{B}_{j}, and an interaction effect β^{AB}_{ij}.

The cell frequencies of the contingency table are thus taken to be dependent on the underlying effects of the variables. The statistical model is based on a multiplicative connection of these effects, but it can be transformed into an additive, so-called "loglinear" model by taking the logarithms. The assumption of a multiplicative connection of the effects of the variables is not only statistically justified but can also be substantiated by empirical considerations: the frequency of accidents, for example, will change depending on the age of traffic participants by a certain proportion and not by a certain fixed amount. This is equivalent to multiplication by a corresponding factor.

The effects $\beta_0,\ldots,\beta^{AB}_{ij}$ can be explained very clearly. If the cell frequencies y_{ij} of the 2 x 2 table are the same (equal distribution), this will be expressed in the general effect β_0. The effects β^{A}_{i} and β^{B}_{j} occur when there is no equal distribution. Differences in the rows of the table result from the effect of the variable A, differences in the columns can be traced back to that of the variable B. The interaction effect β^{AB}_{ij} occurs in addition to the main effects β^{A}_{i} and β^{B}_{j} if A and B are stochastically dependent, it expresses the joint influence of A and B. If the two variables are stochastically independent, β^{AB}_{ij} will be zero.

3. THE PRINCIPLE OF LOGIT MODELS

In Section 2 loglinear models were introduced. They are characterized by treating the absolute frequencies of a contingency table as independent variables.
Logit models, instead, are characterized by using the proportion of the categories of a dichotomous variable. An example may be the number of killed pedestrians in relation to the number of all injured pedestrians.

Figure 2: Schema of a 3 x 2 x 2 contingency table

		C1	C2
A1	B1	y_{111}	y_{112}
	B2	y_{121}	y_{122}
A2	B1	y_{211}	y_{212}
	B2	y_{221}	y_{222}
A3	B1	y_{311}	y_{312}
	B2	y_{321}	y_{322}

In the analysis of the three-dimensional table of Figure 2 the value of $C=C_1/C_2$ is to be treated in dependence on the variables A and B.
This is achieved by calculating, for each A_iB_j, the value

$$\eta_{ij} = \ln \frac{\mu_{ij}}{1 - \mu_{ij}} ,$$

where $$\mu_{ij} = \frac{\mu_{ij1}}{\mu_{ij1} + \mu_{ij2}} = \frac{\mu_{ij1}}{N_{ij}}$$

is the expected value of $$y_{ij} = \frac{y_{ij1}}{N_{ij}} .$$

The linear predictor η_{ij} is called logit, it is easily interpretable as the logarithm of the quotient of the probabilities of $C = C_1$ and $C = C_2$.

If η_{ij} is the predictor for the cell frequencies, we obtain a linear model for the differences between the logarithms of the probabilities

$$\ln \mu_{ij} - \ln (1 - \mu_{ij}) = \beta_0 + \beta^{A}_{i} + \beta^{B}_{j} + \beta^{AB}_{ij} \qquad \begin{array}{l} i = 1,\ldots,3 \\ j = 1,2 \end{array}$$

A model is thus defined by:

1. An observed dependent variable (here: C), in the case of loglinear models the absolute cell frequencies.
2. A linear model constructed from the explanatory variables, predicting the vector η.
3. The probability distribution of the variable C.
4. The link function (e.g. logit) which connects the linear predictor η with the expected value μ.

The type of the model (e.g. loglinear or logit) is determined by the choice of a particular link function.

4. ASSESSMENT OF THE GOODNESS OF FIT

For evaluating any model, we have to determine to which degree the variables contained in the model (or in the contingency table) actually explain the variation in the empirical data.

In principle two relevant measures can be distinguished (cf. Arminger & Küsters, 1986):
First, we can determine that proportion of the deviance in the aggregated data contained in the contingency table which is explained by a model. (PEDAD = Proportion of Explained Deviance on Aggregate Data). This is the conventional method to assess the goodness of fit between a model and the empirical findings.
But what is frequently overlooked is that aggregation leads to a loss of information. In addition to PEDAD we therefore have to determine which proportion of the variation of the original data (later aggregated in the table) is explained by the independent variables and their interactions (PED = Proportion of Explained Deviance). Only this measure is actually comparable with the R^2 known from regression and correlation analyses.

5. A BINOMIAL LOGIT MODEL (EXAMPLE)

The following example should serve as an illustration of how data can be analysed on the basis of loglinear or logit models.
The data used in this example are taken from a set of accident data of the years 1985/86 which were supplemented by vehicle-related data, altogether containing 140 variables. These data exist for the years since 1980 and were made available in the framework of a joint project including the German Federal Highway Research Institute (BASt), the Federal Office for Motor Traffic (Flensburg), the State Authority for Data Processing and Statistics of North-Rhine Westphalia (Bock et.al., 1986).

The dependent variable considered in the following analysis is the proportion of the number of killed and seriously injured drivers to the number of all killed and injured drivers. This is a measure for the average accident severity which is frequently used in accident research. In this example, only accidents between two passenger cars were considered.

Independent variables included:

A ≙	Age of driver
A1	under 25 years
A2	25 to under 60 years
A3	60 years and more

L ≙	Location
L1	inside urban areas
L2	outside urban areas without motorways

W ≙	Unladen weight of the car
W1	under 800 kg
W2	800 kg to under 1200 kg
W3	1200 kg and over

The first category of each of the independent variables A, L, W is used as the basic category in the calculation. The remaining categories of the variables are put into proportion to the basic category (cf. Ernst & Brühning, 1986).

5.1 INTERPRETATION OF THE MODEL

Table 1: Introduction of the main effects into the logit model

Model	Introduced effect	Deviances[1]	Degrees of freedom (DF)	reduced deviances to basic model (Mod.1)	explained deviances [2]
1	GM	957,3	17	-	-
2	GM + A	807,7	15	149,6	15,6%
3	GM + L	150,2	16	807,1	84,3%
4	GM + W	950,5	15	6,8	0,7%
5	GM+A+L+W	8,3	12	949,0	99,1%

[1] The figures specify the deviance of the expected values from the actually observed frequencies. If the frequencies predicted by the model only deviate randomly from the observed values, the deviances will be in asymptotic χ^2-distribution with the degrees of freedom given in the table.

[2] If the deviance calculated with Models 2 to 5 is put into relation to the deviance given by Model 1 (deviance of the basic model), we get the proportion of deviance explained by the parameters of each model.

Model 1 which only uses the GM (= grand mean, regression constant) for the estimation of the cell frequencies represents the hypothesis that the variables A, L, W do not have any effect on the proportion of killed and seriously injured drivers. As the deviances of the Models 2 to 4 show, the explanatory contributions of the individual variables are of highly different value. A reduction of deviance by 84.3% is effected just by the variable "location" (L). The effects of the variables "age of driver" (A) and, in particular, "unladen weight" (W) are clearly smaller with 15.6% and 0.7%, respectively. All main effects taken together in one model (Model 5, the "main effects model") reduce the deviance by 99.1%.

An expansion of the main effects model by the interactions of the 1st level (interactions between two variables) and of the 2nd level (interactions between three variables) does not provide more explanatory power because all the interactions of the 1st and 2nd level are not significantly different from zero.

The main effects model contains the variable "unladen weight" with 3 categories. The category W2 (800 kg to under 1200 kg) however does not result in an estimate that is significantly different from that of the basic category W1 (less than 800 kg). W2 will therefore not be treated as a separate main effect in the optimal model which will include only those main effects which are different from zero with an error probability of $\alpha = 0.05$.

Table 2: Estimation of the parameters and the standard deviation for the optimal model: A + L + W3

Deviance	DF
9.366	13

	ESTIMATE	S.E.	Parameter
1	-1.639	0.03211	GM
2	-0.3593	0.03717	A2
3	0.1668	0.06643	A3
4	0.9844	0.03485	L2
5	-0.1058	0.04876	W3

The interpretation of the model parameters (effects)

The model parameters and their estimates given in Table 2 define the optimal model. Since the main effect "age of the driver" (A) has more than two categories (including the basic category), the optimal model contains 5 parameters.

The model parameters included in the optimal model are easily interpretable:

Negative values of "ESTIMATE" mean that the proportion of killed and injured drivers will decrease when the parameters in question are present, positive values in turn mean that the proportion will increase. The largest positive main effect is caused by the parameter L2 (outside urban areas): there is a relatively strong increase of the proportion of killed and injured drivers outside urban areas.

The signs of the parameters A2 (middle age group) and A3 (old drivers) show directly that the proportion of killed and seriously injured drivers of the middle age group is smaller than that of young drivers while that of old drivers is larger (A1 is the basic category, i.e. the parameter of this category has the value zero and is between the values of A2 and A3).

The effect of the parameter W3 (unladen weight over 1200 kg) is relatively small (negative), i.e. the unladen weight of the car has a little effect on the proportion of killed and seriously injured drivers (W1 and W2 form the basic category).

The simple model shown in Table 2 illustrates how the logit model can describe strong and weak dependencies; non-significant parameters can be easily detected and excluded from the further development of an optimal model.
This model explains 99.0% of the deviance of the basic model (PEDAD); but the proportion of explained variation in the individual data (PED) is only 4.4%.

5.2 EXPANSION OF THE LOGIT MODEL WITH THE INDEPENDENT VARIABLE "ACCIDENT CAUSATION"

In empirical research we always have to ask the question whether the relevant influences have been taken into account in an analysis or whether some important intervening variables have been overlooked. The analysis described above using loglinear or logit models allows an easy way of expanding the model by further variables.

Let us check whether the average accident severity is statistically correlated with the fact whether the driver in question is responsible for the accident or not (according to the opinion of the police). For this, we expanded the multivariate analysis presented in Section 5.1 by the variable "accident causation":

C	≙	Accident causation
C1		driver responsible for the accident
C2		driver not responsible for the accident

Additional use of this dichotomous variable doubles the number of cells in the data matrix of the logit model. This will find an expression in the number of degrees of freedom as well as in the values for deviance. The number of degrees of freedom in the basic model is now DF = 15, the deviance in the basic model is increased to 1519.

The development of the optimal model is performed analogously to Section 6.1; there will be no detailed description of this procedure here because of reasons of time and space.

Table 3: Estimation of the parameters and their standard deviations for the optimal model: C + L + A2 + C2.L2

Deviance	DF
25.64	31

	ESTIMATE	S.E.	Parameter
1	-1.175	0.03695	GM
2	-0.9366	0.04910	C2
3	0.8545	0.04941	L2
4	-0.2669	0.03566	A2
5	-0.2243	0.07057	C2.L2

The model parameters and their estimates given in Table 3 again define the optimal model.

The optimal model includes 5 parameters:
The variable "unladen weight" (W) did not yield estimates significantly different from zero, neither for the categories of the main effect nor for the interaction effects; it is therefore no more explicitly included in the optimal model.
The category A3 of the variable "age of the driver" also did not have an estimate significantly different from zero, it will be collapsed with the category A1 (young drivers) in the optimal model.

In addition to the three main effects (C2, L2, A2), the optimal model includes one interaction of the 1st level (C2.L2).

The outstanding influence in this model is not any more exerted by the location parameter L2; the strongest main effect is to be found with the parameter C2 ('driver not responsible for the accident'). Accordingly high is the influence of this variable on the proportion of killed and seriously injured drivers. The average accident severity is therefore considerably lower for drivers not accused by the police.

For the case that C2 and L2 are both present, the model formula also contains the (significant) interaction effect C2.L2 with positive value (0.2243). This constellation (C2.L2) leads to an average accident severity which is higher by the value of C2.L2 than that which would be calculated if just C2 and L2 were included in the model formula. This fact can be interpreted as follows: the inclination of innocent drivers to indicate slight injuries to the police is smaller in the case of accidents outside urban areas than it is in accidents inside urban areas.

The optimal model allows to give estimates of the expected values of the proportion of killed and seriously injured drivers for any cell of the data matrix. Let us note that, in addition to the main effects, for certain variable combinations the interaction effects of the optimal model will enter the calculation.

Table 4: Proportion of killed and seriously injured drivers in relation to all killed and injured drivers in accidents between two cars and certain selected variable constellations

No.	Variable constellation				Propor-tion
	C	A	L	W[1]	
1	1	1	1	1	23,6%
2	2	1	1	1	10,8%
3	2	2	1	1	8,5%
4	1	2	1	1	19,1%
5	1	1	2	1	42,1%
6	2	1	2	1	26,3%
7	2	2	2	1	21,4%

[1] W1 ≙ W2 ≙ W3

On the assumption that the basic category holds for each of the parameters contained in the model, the expected value of the proportion of killed and seriously injured drivers is 23.6% (Table 4, No.1).
If the driver is not responsible for the accident (C2) and all other variables in the model have the basic category, there is an expected value of 10.8% (No.2).
The lowest expected accident severity is calculated for "drivers not responsible for the accident" of the middle age group (A2) with 8.5% (No.3). The expected value for this age group is higher by more than 10% if the driver is responsible for the accident, viz. 19.2% (No.4).
The highest expected value estimated by the model for the proportion of killed and injured drivers is 42.1% (No.5: outside urban areas and all other variables = basic category). In that constellation, "drivers responsible for the accident" have the corresponding expected value 26.3% (No.6).

The examples show that logit models allow to calculate an expected value for all the variable combinations included in the multivariate analysis.

The expected values deviate from the percentages in a multi-dimensional contingency table because all effects contained in the optimal model as well as those excluded from the model development implicitly enter the calculation of an expected value. However only those parameters that are statistically significant enter the actual model calculation.

The parameters contained in the model explain 98.3% of the deviance of the basic model (PEDAD), but the proportion of explained variation in the individual data (PED) is only 6.9%. There is a remaining variation of 93.1% which results from influences other than those explicitly contained in the model. Therefore it is advisable to use more variables in the analysis if we want to represent reality in an adequate way with the aim of making predictions for individual cases.

6. SUMMARY

The analysis of multivariate dependencies on the basis of logit models allows complex results of a quality which cannot be achieved by the conventional analysis of multidimensional contingency tables. The reason for this is that all effects included in the optimal model as well as those originally excluded from the model development enter the calculation of expected values. The model calculation itself is only based on statistically significant parameters.

The example presented above was only meant as an explanation of the methodological facts. But nevertheless we detected a fact that is frequently overlooked: If we use the proportion of drivers with severe accident consequences related to the number of all killed and injured drivers as a measure of the average accident severity, the result will depend crucially on whether the particular driver is responsible for the accident or not. There is a simple reason for this: drivers that did not cause an accident will be more inclined to report slight injuries to the police than those that caused the accident.
The degree of this difference depends considerably on the definition of injury[1] and the methods of data collection in a country. But the general result is valid also for other countries.

Loglinear and logit models are useful in controlling this and other distorting variables.

[1] The definitions used in Germany:
Seriously injured ≙ person injured in a traffic accident who were in hospital for stationary treatment
Slightly injured ≙ all others injured in an accident who did not receive stationary treatment

Bibliography

Arminger, G., Küsters, U.; 1986.
Statistische Verfahren zur Analyse qualitativer Variabeln.
Bericht zum Forschungsprojekt 8302/3 der Bundesanstalt für Straßenwesen.
Bergisch Gladbach, 1986.
[Statistical methods for the analysis of qualitative variables]

Bock et.al.; 1987.
Aufbereitung und Auswertung von Fahrzeug- und Unfalldaten zur Hebung der Verkehrssicherheit. Gemeinsamer Bericht der Bundesanstalt für Straßenwesen und des Kraftfahrt-Bundesamtes.
[Preparation and analysis of vehicle and accident data for improving traffic safety. Joint report of the German Federal Highway Research Institute and the Federal Office for Motor Traffic (Flensburg)]

Ernst, G., Brühning, E.; 1987.
Einführung in das Arbeiten mit GLIM zur Analyse mehrdimensionaler Kontingenztafeln mittels loglinearer und Logit-Modelle.
Forschungsberichte der Bundesanstalt für Straßenwesen, Bereich Unfallforschung; Nr.148. Bergisch Gladbach, 1987.
[Introduction into the use of GLIM for the analysis of multi-dimensional contingency tables on the basis of loglinear and logit models]