DEPARTMENT OF TRANSPORT
FEDERAL OFFICE OF ROAD SAFETY
DOCUMENT RETRIEVAL INFORMATION

Report No.	Date	Pages	ISBN	ISSN
CR 147	October 1994	34	0 644 35462 3	0810-770X

Title and subtitle

A REVIEW OF STATISTICAL METHODS FOR ROAD TRAFFIC ACCIDENT MASS DATABASES

Authors

O'Neill, T.J. Ginpil, S.

Performing organisations

Statistical Science Program, CMA
Australian National University

Federal Office of Road Safety

Sponsor

Federal Office of Road Safety
GPO Box 594
CANBERRA 2601

Available from	Price	Format
Federal Office of Road Safety GPO Box 594 CANBERRA 2601	No charge	Hard copy

Abstract

This Review considers statistical techniques that can be applied to data from mass road traffic databases, concentrating on approaches which do not require ancillary information to the database. Two general problem types are considered: estimation of the degree to which factors affect crash propensity and the evaluation of the effectiveness of factors in reducing injury outcome once the crash has occurred. In both problem types, a number of techniques are identified which promise to provide significant advantages over methods currently in use.

Keywords

Statistical analysis, databases, injury outcome, crash propensity

Notes

(1) FORS research reports are disseminated in the interests of information exchange.
(2) The views expressed are those of the authors and do not necessarily represent those of the Commonwealth Government.

A Review of Statistical Methods for Road Traffic Accident Mass Databases*

T.J. O'Neill
Statistical Science Program, CMA
Australian National University
&
S. Ginpil
Federal Office of Road Safety

Abstract

This review considers statistical techniques that can be applied to data from mass road traffic databases, concentrating on approaches which do not require ancillary information to the database. Two general problem types are considered: estimation of the degree to which factors affect crash propensity and the evaluation of the effectiveness of factors in reducing injury outcome once the crash has occurred. In both problem types, a number of techniques are identified which promise to provide significant advantages over methods currently in use.

*This research was supported by a Road Safety Seeding Research Grant from the Department of Transport

Federal Office of Road Safety

A Review of Statistical Methods for Road Traffic Accident Mass Databases

Authors

T.J. O'Neill
Statistical Science Program, CMA
Australian National University

S. Ginpil
Federal Office of Road Safety

Australian Government Publishing Service

ISBN 0 644 35462 3

ISSN 0810-770X

Produced by the Australian Government Publishing Service

Contents

0 Executive Summary

While many of the quantitative reports in road safety seek merely to describe the pattern of events involved in crashes, more analytical work has focussed on estimating the degree to which certain factors affect either crash propensity (eg Blood Alcohol Content) or the severity of injuries received after the crash has occurred (eg seat belt wearing).

Since the police in every jurisdiction keep records on all crashes which occur in excess of some given threshold level of severity, analyses which can be performed using only police data may be considerably more cost-effective that those resorting to independent data collections.

This report looks at the range of statistical techniques available for use in these two major types of analytical problems and offers evaluative comments on each.

In general, the report finds that estimates of crash propensity based on "induced exposure" methods, which require no data beyond the police database, may be appropriately used in some cases to substitute for estimates based on expensive surveys of travel behaviour.

While police databases have long been used to estimate factors affecting injury outcome, this report discusses a number of techniques which not only promise to provide more accurate estimates of factors affecting specific individuals in the crash, but also have the potential to estimate more general factors (eg vehicle mass, travelling speed).

These new techniques not only promise to provide answers to questions that have not previously been addressed with crash data, but their greater calculation efficiency means that it may be possible to estimate some issues with far less data than previously required. This could lead to a more timely evaluation of the importance of new factors, such as the airbag, and also to greater opportunities for the analysis of data from countries such as Australia, which by international standards, have relatively few crash events occurring each year.

1 Introduction

This review considers the relative advantages of various techniques available to deal with problems of interest to road safety researchers. Such problems can be broadly classified as the estimation of factors that affect the propensity to be involved in a road traffic accident and those which affect the level of injury resulting from a crash of a given severity.

In the first category are issues such as determining the relationship between a driver's blood alcohol level and the probability that the driver will be involved in a crash. The second class of problems can be illustrated by attempts to determine the degree to which wearing a seat belt can reduce resultant injury.

Many very large databases exist in the road traffic accident area which have been collected by police who have a legal requirement to attend accidents above a certain level of severity. Thus any techniques which can address these issues using only data already available from police sources is likely to be much more efficient than methods that rely on special purpose collections.

There are however particular difficulties in using police databases. In the case of estimating crash propensity, the primary problem is the lack of information about the amount of travel undertaken by various road user groups. It is necessary to show not only that road users with certain characteristics account for a high proportion of crashes, but also that this figure is disproportionately high relative to the amount and type of travel undertaken by this group.

In the case of estimating factors affecting injury outcome, the major difficulty is that the data are either completely or partially "truncated" in terms of the severity of crashes represented. In general, while almost all fatal crashes will be reported to police, there is considerable under-reporting of less severe crashes. This is one of the reasons why several countries have developed databases limited entirely to crashes resulting in fatalities.

An example within Australia is the Federal Office of Road Safety's "Fatality Files". Similar data (the "Fatal Accident Reporting System") are collected in the USA by the National Highway Traffic Safety Administration, an agency of the USA Department of Transportation.

Thus when using such databases to estimate the effectiveness of a particular factor it is not generally appropriate to simply look at the relative proportion of people in the database with and without the factor in question who survive the crash since no information is provided about those crashes in which all involved survived.

This review discusses the various problems in using police databases to answer policy relevant questions and identifies a number of techniques which are likely to have advantages over other approaches currently in use.

2 Estimating factors affecting injury outcome

2.1 Introduction

Almost all of the databases of road traffic accident data only include data on accidents where an injury level of a certain severity was attained. Accidents in which there was no injury are not included in the database. Such databases are said to be subject to truncation.

The most notable example is the *Fatal Accident Reporting System* (FARS) which is a data file compiled by the National Highway Traffic Safety Administration, an agency of the US Department of Transportation. The database began on 1 January 1975 and includes data on all fatal crashes which are crashes in which anyone dies within 30 days of the crash as a result of the crash. Information is collected on all individuals involved in the crash, not only those who died. The data is hierarchical since there is information about the overall crash, about the individual vehicles involved in the crash and about the individuals in each of the vehicles.

A similar database is collected on a biennial basis by the Federal Office of Road Safety in the Federal Department of Transport. The existing files are called the 1988, 1990 and 1992 Fatality Files.

It has long been recognised by many that if the truncation is ignored and standard techniques are applied then serious biases can result. However there continue to be a sprinkling of papers in the literature that ignore the truncation.

In this section several methods which allow for the truncation are discussed. These include the single and double pair methods of Evans (1991), conditional logistic regression and various truncation regression techniques.

In subsection 2.2 cautionary examples are considered to show that ignoring truncation can lead to serious biases even to the extent of changing the sign of effects.

2.2 Examples of the bias caused by truncation when standard methods are used

It is somewhat unusual in the road traffic literature to have access to mass data-bases that are not subject to truncation. Truncation means that we only see the data from accidents where at least one death occurred. One example where truncation was not a problem was considered by Waller, Stewart, Hansen, Stutts, Popkin, & Rodgman (1986) who used a state database records of all reported accidents from North Carolina. Databases of this type can be subject to under-reporting since there is a tendency not to report accidents where damage and injury is minimal. The effect of the rate of under-reporting being dependent on the level of the outcome can be to bias the estimates of factors.

In the following section we illustrate the type of biases which can result if techniques such as logistic regression are applied to truncated databases.

2.2.1 The bias of logistic regression for single vehicle accidents with two occupants

The dangers of ignoring truncation can be illustrated with the following examples involving a database referring to single vehicle crashes.

Consider the simplified case in which each vehicle contains two occupants who have independent probabilities of dying in the crash and that these probabilities are determined by a single factor; whether the occupant is male or female.

In the first example, assume all crash involved vehicles contain one male and one female occupant and that the probability of being killed in a crash is .1 for males and .4 for females. Thus if we look at the total number of males and females who die and the number who survive the crash, the relative risk for females relative to males would be four and the odds ratio would be six.

However databases which record only fatal crashes will not include any information about vehicles in which both occupants survive the crash. As a result, the odds ratio (as calculated from the database) will be 24, and thus the use of a technique such as logistic regression without any correction for truncation would give a significantly exaggerated estimation of the relative disadvantage of females in terms of survival.

It is interesting to note that in this particular case there would be no change to the calculated relative risk.

In the second example, it is assumed that the probabilities of crash survival for males and females are still .1 and .4 respectively, but now half of all crash involved vehicles contain only two male occupants while the other half of crash involved vehicles contain only two female occupants.

In this case, the relative risk and the odds ratio change to 1.19 and 1.5 respectively. In other words, the disadvantage for females has in this instance, been underestimated.

These illustrations make clear that the ignoring of the truncation effect could lead to very serious errors in the estimation of crash survival, and that the nature of this error varies with the distribution of the factor in the crash involved population.

2.2.2 The bias of logistic regression for an outcome not directly subject to truncation

Other binary outcomes which are not directly truncated that are often measured in truncated accidents. Although the problem is more subtle, substantial biases can still result if the truncation is ignored.

For simplicity and concreteness, we will consider cars with only a driver and that we only see the data if the driver died. We wish to examine the effect of airbags on the odds of ejection. Having conditioned on all relevant explanatory variables x and letting A denote the event of an airbag deployment, E denote an ejection and D denote a death, the odds ratio for ejection among dead drivers is

$$\begin{aligned}\tilde{\psi}(E \mid A) &= \frac{P(E \mid D, A)}{P(\bar{E} \mid D, A)} \Big/ \frac{P(E \mid D, \bar{A})}{P(\bar{E} \mid D, \bar{A})} \\ &= \frac{P(D \mid E, A)P(E \mid A)}{P(D \mid \bar{E}, A)P(\bar{E} \mid A)} \Big/ \frac{P(D \mid E, \bar{A})P(E \mid \bar{A})}{P(D \mid \bar{E}, \bar{A})P(\bar{E} \mid \bar{A})} \\ &= \psi(E \mid A) \times \frac{P(D \mid E, A)}{P(D \mid E, \bar{A})} \times \frac{P(D \mid \bar{E}, A)}{P(D \mid \bar{E}, \bar{A})},\end{aligned}$$

where $\psi(A_1 \mid A_2)$ denotes the odds ratio of event A_1 given event A_2. So the odds ratio will be affected by the truncation unless

$$\frac{P(D \mid E, A)}{P(D \mid E, \bar{A})} \times \frac{P(D \mid \bar{E}, A)}{P(D \mid \bar{E}, \bar{A})} = 1.$$

Now it is plausible that

$$\frac{P(D \mid E, A)}{P(D \mid E, \bar{A})} = 1$$

since the probability of death will be predominately affected by what happened after ejection. However, it is not plausible that

$$\frac{P(D \mid \bar{E}, A)}{P(D \mid \bar{E}, \bar{A})} = 1$$

since this would say that for those drivers who are not ejected, the probability of death is unaffected by whether they have an airbag deployment.

So any analysis of the effect of airbags on the probability of ejection must allow for the fact that the data is subject to truncation. To further illustrate this point let us consider an artificial example. We suppose that the following probabilities hold:

- $P(D \mid E, A) = P(D \mid E, \bar{A}) = .9$
- $P(D \mid \bar{E}, A) = .1, P(D \mid \bar{E}, \bar{A}) = .7$
- $P(A) = .2$
- $P(E \mid A) = .4, P(E \mid \bar{A}) = .6.$

	Ejected	*Not Ejected*
Airbag Deployed	.0972	.0162
Airbag Not Deployed	.584	.303

Table 1: Observed Frequencies of ejection and airbag deployment among dead drivers

Then among dead drivers the proportion of those with an airbag deployment who were ejected is .86, the proportion of those without an airbag deployment who were ejected is .66. So although the true odds ratio for ejection given airbag is $\psi(E \mid A) = .44$, the observed odds ratio among the dead drivers is $\tilde{\psi}(E \mid A) = 3.11$. The observed frequencies in each category are given in table 1.

Thus any analysis on the dead drivers which ignores the truncation can have seriously biased estimates with estimated effects going in the wrong direction. In the following sections we consider a variety of methods which can adjust for the truncation.

2.3 Evans' Pair and Double Pair Comparison Method

One of the most popular techniques in the road traffic community for assessing the effects of risk factors on the probability of death in a road traffic accident has been the *Pair* and *Double Pair Comparison Method* (DPC) of Evans (1985). These are discussed at length in the book of Evans (1991).

Evans' approach is not model based and he is not explicit about the assumptions concerning relative risk. However the fact that he commonly uses expressions like "Fatality risk from similar physical insults for females relative to males of the same age versus age" (Evans, 1991, p. 25) suggests that he is using multiplicative relative risks.

In the single pair comparison method, in order to examine the effect of an independent variable E on the probability of death, he matches on all other relevant variables x. For example to compare the driver and passenger seating positions he would consider only accidents where the driver and the passenger were similar in all respects such as age and sex for example. Evans then implicitly assumes that

$$\frac{P(D \mid E, x)}{P(D \mid \bar{E}, x)} = \lambda,$$

that is the relative risk is constant. If $n_{1,x}$ is the number of deaths for E, x and $n_{0,x}$

is the number of deaths for $\bar{E}, x$, then Evans estimates λ by

$$\frac{\sum_x n_{1,x}}{\sum_x n_{0,x}}.$$

This will converge to

$$\frac{\int P(D \mid E, x) f(x) dx}{\int P(D \mid \bar{E}, x) f(x) dx}. \tag{1}$$

In the double pair comparison method, Evans uses a ratio of two such quantities. An example would be if he wanted to compare male to female drivers, he would compare them both to a common type of front seat passenger. If x and y denote the covariates for the cars with the male and female drivers respectively, then Evans' estimate of the relative risk of male to female drivers would be

$$\hat{R} = \frac{\sum_x n_{1,x}}{\sum_x n_{0,x}} \Big/ \frac{\sum_y n_{1,y}}{\sum_y n_{0,y}} = \frac{N_1}{N_2} \Big/ \frac{N_3}{N_4},$$

where $n_{1,x}, n_{0,x}, n_{1,y}, n_{0,y}$ are the number of male driver deaths, passenger deaths in cars driven by a male, female driver deaths and passenger deaths in cars driven by a female respectively.

In both methods Evans (1985) bases variance approximations on an assumption of independent Poisson processes (Evans, 1985, p. 222). Evans suggests that an appropriate variance estimate for $\log R$ is $\sigma_u^2 + 1/N_1 + 1/N_2 + 1/N_3 + 1/N_4$ where σ_u^2 is a term meant to express unavoidable error which persists even in very large samples. This term is arbitrary and is not based on statistical arguments.

Although the implicit assumptions for Evans' methods may be sustainable for low incidence events such as rare diseases in epidemiology there is a potential problem in the road traffic application. Evans commonly finds quite high λ, for example 3 or 5. For example a driver who is older than 60 is found to have a relative risk of 3 or greater compared to a twenty year-old driver. However there are instances in which the probability of death is very high. For example Joksch (1993) reported that frontal crashes with a ΔV of 60 mph (97kph) have an average probability of death of .55. In this case a relative risk of 3 would imply a probability of death greater than 1. Crashes with such impact speeds may represent relatively uncommon events (due to vehicle braking prior to impact). However when investigating factors which have an important large effect on crash survival, the relative risk model could nevertheless imply probabilities of greater then one in the case of less severe, but more frequently observed crash events.

Even in low speed impacts where the probability of death is sufficiently low to avoid probabilities greater than one however, it is still possible that the use of the multiplicative model will result in distortions if the multiplicative effect does not hold for the range of crashes encountered.

A further problem is the assumption that the different seating positions yield independent Poisson processes. The basic Poisson process is the occurrence of the accidents. The outcomes, for example for the different seating positions where all other relevant variables have been matched, cannot be regarded as independent Poisson processes. This comment applies to both the single and double pair comparison methods. Consequently the expressions for the variance approximations and the resulting inference are questionable.

Because of the range of probabilities that are extant in road traffic accident data, the logistic model for probabilities of death is indicated. It is useful to consider the behaviour of the Evans' method when it is applied to some data for which the logistic model holds.

2.3.1 Logistic Example

Consider a hypothetical example with two factors which influence survival. The first is the scaled crash severity x where we suppose that $x \sim N(0,1)$. The second factor is the sex of the individual where E denotes a female. We further suppose that

$$P(D \mid E, x) = \frac{e^{x+1}}{1+e^{x+1}}$$

and

$$P(D \mid \bar{E}, x) = \frac{e^{x}}{1+e^{x}}.$$

So the log-odds ratio of death for females is always one greater than that for males for any crash severity. Then Evans' estimate will converge to

$$\int \frac{e^{x+1}}{1+e^{x+1}}\varphi(x)dx \Big/ \int \frac{e^{x}}{1+e^{x}}\varphi(x)dx = 1.393,$$

where $\varphi(x)$ is the standard normal density function. In Figure 1 we compare the Evans' relative risk with the true relative risk at scaled age x which is $e(1+e^{x})/(1+e^{x+1})$. It can be clearly seen that the Evans' estimate will be a flawed description of the relative risk of the population. More satisfactory descriptions of the effects of independent variables on the probability of death are obtained using the logistic regression techniques which are described in subsequent sections.

2.3.2 Evans' Effectiveness Interpretation

Evans, in recognizing that it may be impossible to estimate the effect of a factor in particular cases, nevertheless claims that his statistics provides an estimate of the average effect of the factor. This is open to two objections. First, as discussed in

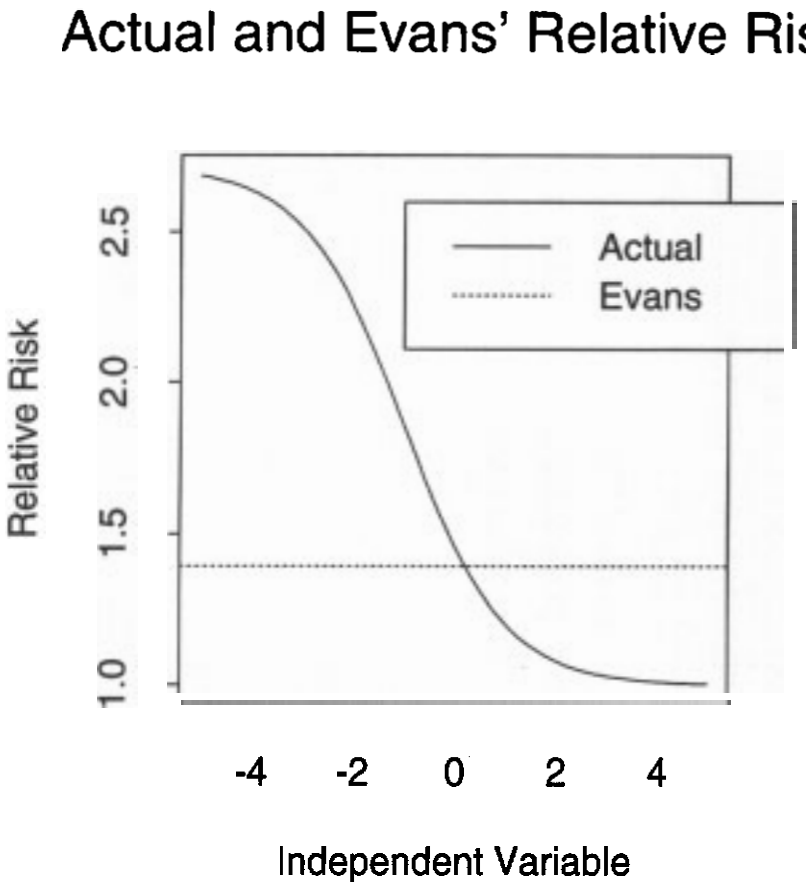

Figure 1: The Evans' and True Relative Risks

section 2.3, distortions could occur in cases not strictly limited by a ceiling effect and this average may be similarly distorted.

On the other hand, even if the effectiveness measure was an unbiased estimate on the average effect, it would only be appropriate for estimating population wide effects. Thus an effectiveness estimate of $X\%$ for seat belt wearing could be used to predict the total reduction in death that might be expected within the non-wearing population, if all those who do not wear a belt were to do so, but would not imply that an individual would reduce his or her risk by $X\%$ by wearing the seat belt (this figure would be expected to be in excess of $X\%$).

Similarly the effectiveness measure makes little sense when applied to non-behavioral factors. For example, in estimating the effect of being elderly or female on crash survival, we are not interested in predicting the number of fatalities that could be reduced if the population of vehicle occupants could be made male or younger. Rather we require estimates of the effect of these factors in individuals.

Even when the effectiveness measure makes sense it may have been calculated using an inappropriate population. In the single pair comparison method, the distribution of the other covariates is taken to be that determined by the matching. In other words, the reduction in deaths is estimated for a population which has a distribution of covariates the same as the matched sample used in the estimation. For example if we used Evans' technique to estimate the effect of seating position then we would be estimating the effect on deaths of making the passenger seat equivalent in safety to the driver's seat. The Evans' technique would use accidents where the driver and passenger are matched very closely in all variables that are thought to influence survival. The resulting numerator and denominator in equation 1 will be integrated with respect to the distribution of the covariates in accidents where the two occupants match in all covariates. This distribution from the subset of the accident population where the driver and the passenger happen to be closely matched can be substantially different from that the desired marginal distribution of covariates for each seat from the full accident population. By requiring a particular type of passenger, an atypical population of accidents may be obtained. For example, when the driver and the passenger were matched on age and sex in the 1992 FARS data, the percentage of age 16–24 increased from 28% to 66%.

As a result the population over which the Evans' estimate is averaging can be very different from the correct populations. The Evans' estimate may not estimate a meaningful quantity for the population of interest. This effect may still persist when estimates are averaged over various types of control passengers.

Thus, while the restriction of the sample to vehicles with multiple occupants is a feature of all available methods, the control of potentially confounding variables through case deletion in the single pair comparison technique introduces particular difficulties. For example the estimation of seating position effects could be seriously

biased if there was an interaction between seat effects and person characteristics (or crash characteristics such as vehicle type and impact speed which vary systematically with person characteristics).

A similar objection applies to the double pair comparison method. Consider the example of an investigation to investigate the difference between passive and lap-sash belts for the driver. Evans would advocate a double pair comparison where we match to a specific type of passenger. There are three relevant distributions of the covariates x:

- The distribution of x in the accidents where the driver is wearing a passive belt and has a closely matched passenger, f_{PA} say.
- The distribution of x in the accidents where the driver is wearing a lap-sash belt and has a closely matched passenger, f_{LS} say.
- The distribution of x in the accidents where the driver is wearing a passive belt and the passenger is ignored, f_A say.

Then, using Evans' terminology and letting D_d and D_p denote driver and passenger death respectively and LS_d and PA_d denote that the driver is wearing either a lap sash or passive restraint respectively, the Evans' estimate of the effectiveness in moving from a passive belted population to a lap-sash belted population will be a consistent estimate of

$$\left(\frac{\int P(D_d \mid LS_d, x) f_{LS}(x)dx}{\int P(D_p \mid LS_d, x) f_{LS}(x)dx}\right) \Big/ \left(\frac{\int P(D_d \mid PA_d, x) f_{PA}(x)dx}{\int P(D_p \mid PA_d, x) f_{PA}(x)dx}\right). \tag{2}$$

This can be quite different from the desired quantity

$$\frac{\int P(D_d \mid LS_d, x) f_A(x)dx}{\int P(D_d \mid PA_d, x) f_A(x)dx}.$$

In his discussion of possible biases in his estimate, Evans (1985) assumes that the densities in equation 2 are the marginal distributions that exist in the population rather than the conditional distributions given that the occupants have been matched. As a result his conclusions on the magnitude of the biases may not not valid.

Because of the above comments, the large sample size requirements and the difficulty of handling complex interactions with Evans' method, it may be preferable to use a logistic regression based procedure instead of Evans' approach.

2.4 Greenland's Method

Another method which is based on the assumption of multiplicative relative risks was recently proposed by Greenland (1994). Unlike Evans (1985) it explicitly makes the model assumptions. It is also a true regression technique rather than at best a stratified approach.

Greenland (1994) argues that the method is desirable since it enables the estimation of the relative risk reduction or case load reduction when a risk factor is changed. However, since the method assumes multiplicative relative risks, the same reservations that were expressed in section 2.3 about the unsuitability of this assumption for the range of probabilities encountered in the road traffic area also hold for this new method.

The approach considers a matched pair model. If y is a random variable which is 1 if an individual dies and 0 otherwise, and the outcomes for the driver and passenger in the jth car are denoted by y_{1j} and y_{2j} respectively, then the relative risk assumption is that

$$P(y_{ij} = 1 \mid z_{ij}, j) = \exp(\alpha_j + z_{ij}^T \beta)$$

where α_j is the effect of the covariates common to the two occupants and z_{ij} is the vector of covariates specific to the individual occupant. So the relative risk for the driver versus the passenger is

$$\phi_j = P(y_{1j} = 1 \mid z_{1j}) / P(y_{2j} = 1 \mid z_{2j}) = \exp(d_j^T \beta),$$

where $d_j = z_{1j} - z_{2j}$. The problem of estimating the ratio of expectations is often called the *Fieller-Creasy* problem in the statistical literature and an estimating equation approach based solution was considered by McCullagh & Nelder (1989). In the case of road traffic data, it is not possible to choose the optimal weights for the estimating equation because of the truncation inherent in the FARS data. Greenland (1994) advocates a particular selection of weights which gives the estimate of β as the solution of

$$S(\beta) = \sum_j s_j(\beta) = \sum_j (y_{1j} - \phi_j y_{2j}) d_j / (\phi_j + 1) = 0.$$

The estimated covariance matrix of $\hat{\beta}$ is $D(\hat{\beta})^{-1} V_s(\hat{\beta}) D(\hat{\beta})^{-1}$ where $V_s(\beta) = \sum_j s_j(\beta) s_j(\beta)^T$ and $D(\beta) = \sum_j (y_{1j} + y_{2j}) \phi_j / (\phi_j + 1)^2 d_j d_j^T$. The inference proceeds as usual in estimating equations. Greenland's method has much to recommend it over the methods of Evans (1985). It is

- based on sound contemporary statistics,
- the model assumptions are explicit,

- it is a true regression technique,
- theory for statistical inference is readily available.

Greenland's technique is the analogue for multiplicative relative risks of the conditional logistic approach discussed in section 2.5 which is applicable when a logistic model is assumed.

2.5 Conditional Logistic Regression

Conditional logistic regression is one of the most under-utilized statistical techniques in the road traffic literature. It finesses the problem of truncation and enables a regression analysis of the FARS database. It is a very accessible technique which has been available in the statistical literature for many years. Early readable references on the application of the technique to matched pair designs are Breslow, Day, Halvorsen, Prentice, & Sabai (1978), Breslow & Day (1980) and Holford, White, & Kelsey (1978). The technique is discussed in a number of statistical textbooks such as Hosmer & Lemeshow (1989) who give a good introduction to the technique and consider diagnostics for the model. The technique is available in most modern statistical packages such as SAS, S-Plus and EGRET.

The principal application in the road traffic literature of conditional logistic regression has been by Lui, McGee, Rhodes, & Pollack (1988) who looked at the effects of variables such as seat belt usage and principal point of impact on the probability of driver death. The data from FARS which was used was two car collisions which involved a driver death.

For clarity of exposition we will restrict the discussion to accidents which only involved two individuals. The extension to more than two individuals is relatively straightforward and is discussed in the references cited above. Conditional logistic regression begins by assuming that a logistic model holds for the probability of death in an accident. If y is a random variable which is 1 if an individual dies and 0 otherwise, and the outcomes for the two individuals in the jth accident are denoted by y_{1j} and y_{2j} respectively, then the logistic assumption is that

$$P(y_{ij} = 1 \mid z_{ij}, j) = \frac{\exp(\alpha_j + z_{ij}^T \beta)}{1 + \exp(\alpha_j + z_{ij}^T \beta)}$$

where α_j is the effect of the covariates common to both individuals in the accident and z_{ij} is the vector of covariates specific to the individual occupant. The technique only looks at accidents where exactly one death occurred which are accidents where $y_{1j} + y_{2j} = 1$. For such accidents it follows that

$$P(y_{1j} = 1 \mid y_{1j} + y_{2j} = 1) \quad = \quad p(\beta, d_j)$$

$$= \frac{\exp(z_{1j}^T\beta)}{\exp(z_{1j}^T\beta)+\exp(z_{2j}^T\beta)}$$
$$= \frac{\exp(d_j^T\beta)}{1+\exp(d_j^T\beta)}, \quad (3)$$

where $d_j = z_{1j} - z_{2j}$. From equation 3 we can see that the conditional logistic model can be fitted by an ordinary logistic regression of y_{1j} on d_j in accidents with exactly one death. The estimate $\hat{\beta}$ is found by maximizing $\prod_{\text{accidents } j} p(\beta, d_j)$ or equivalently solving

$$S(\beta) = \sum_j u_j(\beta) = \sum_j \{y_{1j} - p(\beta, d_j)\}d_j = 0 \quad (4)$$

and the estimated covariance matrix of $\hat{\beta}$ is $V(\hat{\beta})$ where

$$V^{-1}(\beta) = \sum_{\text{accidents } j} p(\beta, d_j)\{1 - p(\beta, d_j)\}d_j d_j^T.$$

It can be seen from equation 3 that the covariates that are common to all individuals in an accident do not appear in the conditional probabilities. This can be regarded as both a virtue and a defect of conditional logistic regression. It is a drawback since for example the effects of variables such vehicle mass and speed cannot be estimated from single vehicle accidents. However since these variables are sometimes difficult to measure or are unavailable it can be useful to have an estimate which does not require knowledge of these variables. These variables are not omitted from the inference without cost. In the subsequent sections we consider estimation procedures which utilize all the available covariates and so can be expected to yield more precise estimates of all the coefficients.

2.6 Truncated Logistic Regression

A technique which is related to conditional logistic regression was recently proposed by O'Neill & Barry (1994c) and O'Neill & Barry (1994b). Suppose that the binary variable y is 0 if an individual survives and 1 if the individual dies. Also suppose that x is a vector of covariates thought to influence survival. Then the logistic model is that

$$Pr(y = 1) = \frac{\exp \beta^T x}{1 + \exp \beta^T x} = p(\beta, x) = 1 - q(\beta, x), \quad (5)$$

where β is a vector of unknown covariates. The conventional logistic regression estimate of β is the maximizer of

$$\prod_{\text{sample}} p(\beta, x_i)^{y_i} q(\beta, x_i)^{1-y_i}. \quad (6)$$

We have seen in section 2.2 that this method will result in biased estimators of regression parameters if it is applied to truncated data. The *Truncated Logistic Regression* (TLR) approach conditions on the probability that an accident is observed which is the probability that it results in at least one fatality. This has the effect of introducing a divisor to logistic regression likelihood equation 6. The truncated logistic regression estimator of β is the maximizer of

$$\prod_{j \in \text{accidents}} \frac{\prod_{i \in \text{accident}_j} p(\beta, x_{ij})^{w_{ij}} q(\beta, x_{ij})^{1-w_{ij}}}{P(\beta, j)} \tag{7}$$

where $P(\beta, j) = 1 - Q(\beta, j) = 1 - \prod_{i \in \text{accident}_j} q(\beta, x_{ij})$. This modification of the logistic regression likelihood equation 6 gives a well behaved estimator which has all the usual desirable properties of maximum likelihood estimators. $\hat{\beta}$ can also be written as the solution of

$$\sum_{j \in \text{accidents}} \sum_{i \in \text{accident}_j} x_{ij} y_{ij} - \mu(\beta, j) = 0$$

where

$$\mu(\beta, j) = P(\beta, j)^{-1} \sum_{i \in \text{accident}_j} p(\beta, x_{ij}) x_{ij}.$$

The estimated variance matrix of $\hat{\beta}$ is $V(\hat{\beta})$ where

$$V^{-1}(\beta) = \sum_{j \in \text{accidents}} \left\{ \sum_{i \in \text{accident}_j} p(\beta, x_{ij}) q(\beta, x_{ij}) / P(\beta, j) x_{ij} x_{ij}^T \right\} - Q(\beta, j) \mu(\beta, j) \mu(\beta, j)^T.$$

Truncated Logistic Regression extends naturally to ordinal data. Conditional Logistic Regression method cannot be extended to the ordinal case. A full discussion of *Truncated Ordinal Regression* (TOR) is given in O'Neill & Barry (1994b). A special case of TOR was considered by Weiss (1993) who looked at a probit model for the bivariate case of neck and body injuries in motorcycle accidents.

An ordinal response variable is assumed to have $k + 1$ levels and the case $k = 1$ corresponds to binary data. An example where $k = 3$ would be a four point scale for injury:

1. No injury
2. Injury
3. Died after hospitalization

4. Died at scene

The data is said to be group truncated if the responses for a group are only known if at least one of the group attained a specified level, j say. In the above example the cutoff might be $j = 3$ in which case the injury levels are only recorded if at least one person in the accident dies. The *Truncated Ordinal Regression* (TOR) likelihood is the natural generalization of equation 7. The method allows for different relationships between the covariates to the logistic link given in equation 5.

The following general properties hold.

Relative Advantages of Truncated Logistic Regression

- Since the TLR uses the full information from the sample it can be expected to lead to more accurate inference than CLR or DPC.
- More effects can be fitted using TLR. The conditional logistic regression likelihood equation 4 only includes terms which vary within a given accident. For example, for single vehicle accidents, since the speed of the car is constant for all the occupants, its effect on the survival prospects cannot be estimated using CLR. TLR on the other hand can be used to estimate its effect.
- TOR can be used to estimate the relative seriousness of crashes for occupants. The TLR method allows us to estimate the probability that a given type of crash will kill a given type of occupant. The TOR method enables the estimation of the probabilities of the various categories of injury.
- Only TLR can be used to estimate the total number of potentially fatal crashes. The TLR method allows us to estimate the probability that a particular configuration of factors results in a fatality. By dividing the observed number of crashes of this type by this probability we obtain an estimate of the total number of potentially fatal crashes of this type. The estimates can then be summed over the categories of crashes to obtain an estimate of the total number of potentially fatal crashes.
- TOR can be generalized to different link functions. Various researchers have found that the logistic link given in equation 5 does not work well when dealing with very rare events. The TOR method allows us to choose the link function which best fits a given data set.

2.7 Special rate comparison techniques

A collection of special techniques uses the observation that if a factor does not affect injury outcome, then the frequency of the factor should be the same in the general

accident population as in the truncated database. Any differences in the rates is attributed to the effectiveness of the factor in reducing injury.

For example an analysis of the effectiveness of airbags was performed by Zador & Ciccone (1991) who assumed that airbags are only effective in frontal impacts. They expected therefore that the frequency of frontal impacts among dead drivers would be lower in airbag equipped vehicles than those without. The technique concludes that the reduction in odds of frontal impact is the deaths prevented by airbags. This inference is questionable since it requires that the device is only effective in one direction and has no effect on other impacts. This is a strong assumption which would normally require verification before the method could be applied. Since the method only uses the information from the dead drivers it will be inefficient compared to conditional or truncated logistic regression. It will also be necessary to stratify the data to control for car type and driver characteristics.

2.8 Summary and comparison of statistical techniques applicable to truncated data

We have seen that the techniques for the analysis of accident level truncated binary data on deaths can be divided into those that are based on an assumption of multiplicative relative risks and those based on a logistic assumption. It is felt that there are some difficulties with the use of a multiplicative relative risk assumption for road traffic data. If the assumption is made, then the method of Greenland (1994) is a regression technique that is quite attractive. The pair and double pair comparison methods of Evans (1985) are not regression or model based but can be effective on large databases such as the FARS database.

It has been argued that preferred methods for the regression analysis of accident level truncated binary data on deaths should be based on the logistic assumption. We have seen that viable regression estimation can be performed using either Truncated Logistic Regression (TLR) or Conditional Logistic Regression (CLR). The methods are applicable to any situations in which binary variates and associated covariates are observed if and only if at least one member of the group has a positive binary response. The efficiency loss due to truncation can be substantial but not catastrophic.

The choice between the use of truncated and conditional logistic regression in an analysis of group truncated binary data is governed by several factors. For random or unstructured accident level effects, conditional logistic should be used since the method eliminates all accident level effects from the likelihood. If the individual probabilities can be adequately modelled in terms of known accident level and individual level covariates, then either TLR or CLR can be used. TLR will normally be the preferred method since it has higher efficiency and can estimate accident level

effects. TLR can also be extended to allow the estimation of the total number of accidents, not only those that were observed. This can be of considerable interest to road traffic researchers. A disadvantage of TLR is that information about all covariates which affect the probability of death in the crash must be available. It will generally be the case that major determinants of survival such as impact speed will not be available and can only be approximated by measures of travel speed.

Alternatively, the two methods may be regarded as complementary and in some situations both can be fitted and the results compared. This may provide insight into the appropriateness of the models and may indicate the true pattern of the data.

The CLR estimates may be found using a variety of software packages such as SAS, S-Plus or EGRET. A library of S-Plus functions and C routines for TLR has been developed by Simon Barry and the author and is available by request at no cost.

In summary, the main issue that will determine the application of TLR is the availability of accident level covariates. When TLR can be applied, it alone offers the very desirable possibility of estimating accident level effects and the total number of accidents and has higher efficiency than CLR. If accurate accident level covariates are unavailable, then the efficacy of available surrogates is an open issue.

3 Estimating factors affecting crash propensity

3.1 Direct calculation of rates

The factors that affect the propensity of individuals to have accidents can be measured by direct observational studies on the population. For example, it may be possible to survey the extent of travel by different modes and calculate rates such as fatalities per distance unit as a measure of crash propensity. These studies are often prohibitively expensive. Consequently it is attractive to use methods which make use of existing databases.

3.2 Case-Control Studies

Case-control studies can be used to study the effect of factors on crash propensity.The use of case-control studies has been fairly limited in the road traffic literature. They are not internal to the existing databases and normally require the collection of supplementary information. They have been used primarily for the study the effect of blood alcohol content. There are several reasons for the limited use of case-control studies. They are relatively expensive to conduct. Access to appropriate data is often constrained by civil liberties issues. It is often very difficult to match adequately.

Case-control studies could be performed by selecting cases from the mass databases and selecting controls from a secondary source such as a registry of licensed drivers. The study could either be matched or unmatched. Such a study would not properly fit into the class of methods that only require data from the mass database. However there may be considerable scope for using case-control studies when a secondary source of controls is available. One example of an unmatched case-control study is given in Wong, Lee, Phoon, Yiu, Fung, & McLean (1990) who looked at the effect of driving experience on the probability of having a motorcycle accident.

3.3 Exposure

3.3.1 Introduction

A prime aim of road traffic research is to explain the effect of driver characteristics on the probability of having an accident. A recurring obstacle to the use of fairly standard epidemiological techniques for the estimation of relative risk or the odds ratio is that it is often very difficult to estimate exposure from the existing road traffic mass databases. The numerator in a relative risk calculation, namely the number of cases, is usually well known but the denominator is not and it would normally be very difficult or impossible to collect such information. For example, to determine an accident rate per distance unit for each age group would require accurate distances

travelled by each age group. This information is typically not available. A variety of techniques have been suggested to finesse this difficulty. The most popular is the induced exposure method which was originally introduced by Thorpe (1964).

3.3.2 Induced exposure

The fundamental assumption proposed by Thorpe (1964) and subsequently by others such Haight (1973) in the Induced Exposure Method is that drivers involved in accidents can be split into two groups, responsible and innocent. There are two different techniques for identifying drivers who are responsible for the accident. Thorpe (1964) uses the set of single vehicle accidents where every driver is judged to be responsible and then infers the distribution of innocents from the two car collisions. Carr (1969) assumes that each two car collision has an identified responsible and innocent driver and proceeds directly. Carr (1969) argues that this is a reasonable assumption.

Further, the distribution of the accident innocents within a specific combination of the independent variables is assumed to be the same as the overall distribution of individuals over that combination of the independent variables. The validity of the method will depend on having either the complete set of accidents or an unbiased sample of them for each specific combination of the independent variables.

In order to fix the concepts involved in the method, suppose that we wish to estimate the distribution of some exposure variable E, such as distance travelled at night, given a population characteristic A, such as the age of the driver. We will use the approach suggested by Cerrelli (1973). The quantity that we wish to estimate is $P(E \mid A)$. The method uses the fact that

$$P(E \mid A) = P(A \mid E)\frac{P(E)}{P(A)}. \tag{8}$$

The marginal distribution $P(E)$ of exposure variables can be readily obtained from surveys of road usage and does not require information on vehicle occupants. The marginal distribution $P(A)$ of the population variable would normally be obtained from a central registry such as the registry of licensed drivers. The final quantity necessary to complete the estimation of the right hand side of equation 8 and hence $P(E \mid A)$ is $P(A \mid E)$. This is not directly observable. The assumption is made that the proportion of drivers of age A given they were not responsible for a crash that occurred at exposure level E is the same as the proportion of age A drivers at exposure level E,

$$P(A \mid \text{Not responsible for accident}, E) = P(A \mid E). \tag{9}$$

So the method assumes that the number of not responsible accidents in a given set of conditions for a particular type of driver is proportional to the relative exposure

of that particular type of driver to that set of environmental conditions. Provided that equation 9 holds any type of accident can be used. For example accidents from a truncated database could be used.

Some related quantities are often estimated (Cerrelli, 1973). If R denotes responsibility for the accident, then define the relative exposure index (REI), the liability index (LI) and the hazard index (HI) as follows:

$$\text{REI}_i = \frac{\%\text{ innocent in class i}}{\%\text{ licensed drivers in class i}} = \frac{P(A=i \mid \bar{R}, E)}{P(A=i)},$$

$$\text{LI}_i = \frac{\%\text{ responsible in class i}}{\%\text{ licensed drivers in class i}} = \frac{P(A=i \mid R, E)}{P(A=i)},$$

$$\text{HI}_i = \frac{\text{LI}_i}{\text{REI}_i} = \frac{\%\text{ liability in class i}}{\%\text{ exposure of drivers in class i}}.$$

Note that since the denominators are the same, this can be calculated directly from the two vehicle accident data. Hence since $P(\bar{R} \mid E) = P(R \mid E)$, this is

$$\text{HI}_i = \frac{\%\text{ liable in class i}}{\%\text{ innocent in class i}} = \frac{P(R \mid A=i, E)}{P(\bar{R} \mid A=i, E)}.$$

The auxiliary information about the frequency of class i in the general population is not required for the calculation of the hazard index. The hazard index is simply the odds ratio for responsibility given the age and exposure variable.

Subsequently many authors have considered the induced exposure approach. A variation on the fully responsible-innocent dichotomy was considered by Wasielewski & Evans (1985) who assumed that there is a proportion of two vehicle accidents ρ in which both drivers are responsible. Under the original model, the observed distribution of age for example is $D_i = (\text{LI}_i + \text{REI}_i)/2$. With the modified model, it becomes $D_i = \{(1+\rho)\text{LI}_i + (1-\rho)\text{REI}_i\}/2$. They use an estimate of REI from a secondary source, LI from single vehicle accidents and D from two vehicle accidents to obtain an estimate of ρ. This extension of the basic model does not seem to have been taken up in the literature.

Another variation was recently discussed by Cuthbert (1994) who uses the ratio of the single vehicle and multiple vehicle accident data to derive an estimate of the two exposures. The method is essentially a variation on the Thorpe model which assumes that the at faults are at higher representation in the single vehicle accidents. Cuthbert (1994) postulates that the exposure for driver type i to factor level j is

$$\sigma_{ij} = p_j(s_i + \gamma_j). \tag{10}$$

The "driver type exposure" s_i can clearly be identified with the at fault category in Thorpe's terminology. The "at random component" γ_j which is determined by the level of the factor can clearly be identified with the innocent drivers. An approximation was introduced which generated a multiplicative model for the differences of the logarithms of the ratios of the single vehicle to two vehicle involvement. This model has two possible criticisms:

- The practical suitability of the model given in equation 10 is not clear.
- The accuracy of the approximation and hence the multiplicative model has not been established.

Recently Janke (1991) again argued the case for an induced exposure approach rather than an accidents per mile approach. Lyles & Stamatiadis (1991) reconsider only the hazard index and rename it the Involvement Ratio (IR).

The sampling distribution of the various estimates have only been briefly considered in the literature (Wasielewski & Evans, 1985) and there remains work to be done to put inference concerning induced exposure on a sound statistical basis.

3.4 Survival Analysis

3.4.1 Introduction

The technique of survival analysis can be used to study the crash free periods for individuals and the effect of factors on that time. Survival analysis has been a topic of great research activity in statistics for the last twenty years. The activity was largely caused by the two papers by Kaplan & Meier (1958) and Cox (1972). The former deals with estimating the distribution of the time to death when censoring is present and in the absence of covariates that might influence survival. The latter extended the method to allow for the influence of explanatory variables on the hazard of death. We will not describe here the mathematical details of the Proportional Hazards Model since it is very accessible both in the literature and in statistical computing packages such as SAS, SPSS and S-Plus. The key assumption is that when we look at the set $\mathcal{R}(t)$ of individuals that are at risk of death at a time t, the probability that it is individual (i) who dies is

$$\frac{\exp(\beta^T x_{(i)})}{\sum_{j \in \mathcal{R}(t_{(i)})} \exp(\beta^T x_j)}. \tag{11}$$

where x is the vector of covariates influencing survival and β is the vector of unknown parameters. MacKensie (1986) considered the suitability of the Proportional Hazards Model for road traffic data. However he gave no consideration of the practical issue of the availability of suitable data to fit the model. Mannering (1993) also looked

at the application of survival methods to road traffic data. He looked at a Weibull log-linear model fitted to a data set on time between accidents that he collected in a survey of students. He did not consider how the models might be applied to the existing mass database information.

The application of survival analysis techniques to the road traffic area is limited by the availability of suitable data. To apply classical survival analysis techniques in the road traffic area it is necessary to identify a set of individuals with the following characteristics:

- The time when the individual first became at risk of death must be known. For example in a particular study it may be the age when the individual gained a license.
- The time when the individual ceased to be at risk, either because of death by road accident or by other causes or ceased to be a road user must be known.

This information would usually only be available from secondary sources of data such as registries of licensed drivers. Registries can be used to establish the set of people at risk. The truncated road traffic database can be used to identify the road deaths. Using two sources of data however introduces the problems of incompatibilities in the data-bases. The truncated database will include some unlicensed drivers and some drivers having licenses who are not captured by the registry, for example foreign drivers. The license registry will include individuals who have stopped driving or who have left the catchment area for the truncated database. These incompatibilities may introduce biases into a survival analysis. There may also be privacy restrictions which limit the matching of the data-bases. However the technique of using ancillary data-bases to truncated database on traffic deaths offers considerable promise for the application of survival techniques to data from truncated databases.

3.4.2 Survival Analyses based on FARS

There have been very few survival analyses based only on FARS. Lui & Marchbanks (1990) looked at the length of time from a road traffic infringement to subsequent death in a car accident. They conditioned on actually being included in FARS and fitted a Weibull distribution in each age category. This approach did not allow the effect of other variables to be considered except by extending the stratification.

Lui & Pollack (1991) used techniques that were derived for the analysis of AIDS data by Lagakos, Barraj, & De Gruttola (1988) to again look at the length of time from a road traffic infringement to subsequent death in a car accident. The technique only requires data internal to the FARS database where the time of the last traffic infringement had been recorded. In contrast the methods of the previous section

required access to another database to obtain a suitable cohort to look for in the FORS database. The technique only uses individuals who had a traffic infringement and died in a traffic accident in a fixed time interval. All such individuals would be included in the FARS files. In the particular example that is considered they look at the 36 month period from 1986–1988. If X denotes the time in months to the traffic infringement and T denotes the time from the traffic infringement to death in a traffic accident, then $S = 36 - T$ is a left truncated variable in that S must be at least X for it to be included. Also it is easy to show that if X and T are independent then the hazard function of S at s is the hazard function of T at $36 - s$. So modelling the hazard function of S is equivalent to modelling the hazard function of T until 36 months.

The analysis proceeds very similarly to an ordinary Cox Proportional Hazards Regression. The only modification is in the risk set at time s which consists of all those known to be at risk at that time which is the set of individuals such that $X_i \leq s$ and $S_i \geq s$. Subject to this modification of the risk set the usual Cox Proportional Hazards Regression methods apply.

The primary objection to this technique is the assumption of independence of the time to the traffic infringement and the recurrence time to the death.

4 Further statistical techniques relevant to road traffic data

4.1 Meta-Analysis

The technique of meta-analysis seeks to combine the results of a set of randomly selected independent trials to obtain an overall test of significance of an effect. It is important that the set of trials can be regarded as a random sample from the set of all trials on the effect.

The medical area has been the leader in the meta-analysis field. There is an extensive literature in the medical application of meta-analysis. An excellent entry to the field is the August 93 issue of Statistics in Medicine (Everitt, 1993) which was dedicated to meta-analysis and contained several review articles. A major complication in the medical area is that meta-analyses have typically been performed on published results. The literature is culled for all papers testing a particular factor. There is a tendency that only papers with significance levels close to .05 to be published. Consequently the published literature represents a biased sample of the studies that were conducted to assess the effect of the given factor. This is called the "File Drawer Problem" - trials with non significant results have been consigned to the file drawer. This effect has tended to make the results of meta-analyses somewhat controversial.

Hauer (1983) has advocated the combination of results from independent studies in the road traffic context. If the studies are gleaned from the published literature, then the file drawer problem will apply. However, the application of meta-analysis to the road traffic area is more straightforward. Databases in the area are often more accessible than the medical databases where access is often determined by the original investigators. Researchers would not normally be limited to published material and could readily have access to the original data. For example the FARS data is available for purchase. As a result the file drawer problem can be avoided in the road traffic area. Several analyses can be envisaged. For example the results of airbag analyses from disjoint databases could be combined using meta-analysis techniques.

Meta-analysis offers considerable promise for the analysis of road traffic data.

4.2 Bayesian Techniques

In very recent years, there has been great research interest by statisticians in a variety of Bayesian techniques. The Bayesian paradigm assumes that the unknown parameters are themselves random variables which have a distribution called a prior. The prior is chosen by the researcher before the data is collected to represent the researcher's prior belief about the unknown parameter. The primary objection to Bayesian techniques has been that the choice of the prior is subjective. Different

researchers will have different priors and consequently arrive at different conclusions from the same set of data. This problem can be averted by choosing non-informative priors which essentially assume that the researcher has no apriori knowledge about the unknown parameter. In this way the classical frequentist inference can be mimicked. However the methods can be applied to a variety of problems that are intractable using conventional statistical inference.

As an example consider a database subject to group truncation. The techniques of section 2 allow us to estimate the effects of covariates on the probability of death. Suppose that we would also like information on the overall number of particular types of accidents, not only those that result in a fatality. Further suppose we would also like information on the overall distribution of the covariates in the general population of accidents. The Bayesian method proposes a distribution for the unknown total number of accidents and the distribution of the covariates in those accidents. The parameters of those distributions are themselves given a distribution, called a prior. Then the probable distribution of the parameters given the observed data, called the posterior, is calculated. From the posterior various quantities such as confidence intervals can be calculated. The mathematics involved in calculating the posterior is often intractable since it will often involve high dimensional integrals. That is the case when truncation is involved. Gibbs sampling is a technique which allows us to obtain estimates in such situations. Gibbs sampling and variants such as the Metropolis-Hastings algorithm are methods that generate observations from the posterior distributions and use the resulting data to perform inference. A good review of simulation based Bayesian techniques can be found in Smith & Roberts (1993). Gibbs sampling has been developed for truncated data by Simon Barry and the author and some details are given in O'Neill & Barry (1994a).

As a hypothetical example suppose we wished to estimate the total number and character of accidents in rural areas using data from a truncated database. The Gibbs sampling technique would use the data from only accidents with a fatality and would give confidence intervals for the total number of accidents and the overall pattern of accidents.

Gibbs sampling is one example of a set of modern statistical simulation techniques which are designed to explore likelihoods and posterior distributions. The methods are very computational and have only become feasible for general use in the last few years. A good reference is the book by Tanner (1993). The collection of techniques offers considerable promise for the solution of some difficult problems in the road traffic area.

References

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research. 1: The analysis of case-control studies.* No. 32. International agency for research on cancer, Lyon.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L., & Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology, 108*, 299–307.

Carr, B. R. (1969). A statistical analysis of rural Ontario traffic accidents using induced exposure data. *Accident Analysis and Prevention, 1*, 343–357.

Cerrelli, E. C. (1973). Driver exposure: the indirect approach for obtaining relative measures. *Accident Analysis and Prevention, 5*, 147–156.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological, 34*, 187–220.

Cuthbert, J. R. (1994). An extension of the induced exposure method of estimating driver risk. *Journal of the Royal Statistical Society, Series A, General, 157*, 177–190.

Evans, L. (1985). Double pair comparisons - a new method to determine how occupant characteristics affect fatality risk in traffic crashes. *Accident Analysis and Prevention, 18*, 217–227.

Evans, L. (1991). *Traffic Safety and the Driver.* Van Nostrand Reinhold, New York.

Everitt, B. S. (Ed.). (1993). *Statistical Methods in Medical Research*, Vol. 2. Edward Arnold, Sevenoaks.

Greenland, S. (1994). Modelling risk ratios from matched cohort data: an estimating equation approach. *Applied Statistics, 43*, 223–232.

Haight, F. A. (1973). Induced exposure. *Accident Analysis and Prevention, 5*, 111–126.

Hauer, E. (1983). Reflections on methods of statistical inference in research on the effect of safety countmeasures. *Accident Analysis and Prevention, 15*, 275–285.

Holford, T. R., White, C., & Kelsey, J. L. (1978). Multivariate analysis for matched case-control studies. *American Journal of Epidemiology, 107*, 245–256.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied Logistic Regression.* John Wiley & Sons, Inc., New York.

Janke, M. K. (1991). Accidents, mileage, and the exaggeration of risk. *Accident Analysis and Prevention, 23*, 183–188.

Joksch, H. C. (1993). Velocity change and fatality risk in a crash - a rule of thumb. *Accident Analysis and Prevention, 25*, 103–104.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*, 457–481.

Lagakos, S. W., Barraj, L. M., & De Gruttola, V. (1988). Nonparametric analysis of truncated survival data, with application to aids. *Biometrika, 75*, 515–523.

Lui, K. J., & Marchbanks, P. A. (1990). A study of the time between previous traffic infractions and fatal automobile crashes, 1984–1986. *Journal of Safety Research, 21*, 45–51.

Lui, K. J., McGee, D., Rhodes, P., & Pollack, D. (1988). An application of conditional logistic regression to study the effects the effects of safety belts, principal impact points and car weights on drivers fatalities. *Journal of Safety Research, 19*, 197–203.

Lui, K. J., & Pollack, D. (1991). An application of proportional hazards model to study the recurrent time between traffic accidents or infractions and subsequent fatal automobile crashes, 1986–1988. *Journal of Safety Research, 22*, 163–170.

Lyles, R. W., & Stamatiadis, P. (1991). Quasi-induced exposure revisited. *Accident Analysis and Prevention, 23*, 275–285.

MacKensie, G. (1986). A proportional hazards model for accident data. *Journal of the Royal Statistical Society, Series A, General, 149*, 366–375.

Mannering, F. L. (1993). Male/female driver characteristics and accident risk: some new evidence. *Accident Analysis and Prevention, 25*, 77–84.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (Second edition). Chapman and Hall, London.

O'Neill, T. J., & Barry, S. (1994a). Empirical priors and Gibbs sampling for truncated regression. In Preparation.

O'Neill, T. J., & Barry, S. (1994b). Group truncated ordinal regression. *Statistics and Probability Letters.* In Press.

O'Neill, T. J., & Barry, S. (1994c). Truncated logistic regression. *Biometrics.* In Press.

Smith, A. F. M., & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B, Methodological, 55,* 3–23.

Tanner, M. A. (1993). *Tools for statistical inference* (Second edition). Springer-Verlag, New York.

Thorpe, J. D. (1964). Calculating relative involvement rates in accidents without determining exposure. *Australian Road Research, 2,* 25–36. Reprinted (1967) Traffic Safety Research Review, 11, 3–8.

Waller, P. F., Stewart, J. R., Hansen, A. R., Stutts, J. C., Popkin, C. L., & Rodgman, E. A. (1986). The potentiating effects of alcohol on driver injury. *Journal of the American Medical Association, 256,* 1461–1466.

Wasielewski, B., & Evans, L. (1985). A statistical approach to estimating driver responsibility in two-car crashes. *Journal of Safety Research, 16,* 37–48.

Weiss, A. A. (1993). A bivariate ordered probit model with truncation: helmet use and motorcycle injuries. *Applied Statistics, 42,* 487–499.

Wong, T., Lee, J., Phoon, W., Yiu, P., Fung, K., & McLean, J. A. (1990). Driving experience and the risk of traffic accident among motorcyclists. *Social Science and Medicine, 30,* 639–640.

Zador, P. L., & Ciccone, M. A. (1991). Driver fatalities in frontal impacts: Comparisons between cars with air bags and manual belts. Tech. rep., Insurance Institute for Highway Safety.