# DOCUMENT INFORMATION


| Report No. | Date | Pages | ISBN | ISSN |
|---|---|---|---|---|
| CR 87 | October 1989 | 107 | 0642 51291 4 | 0810-7704 |


| | |
|---|---|
| **Title and Subtitle** | How some recent advances in econometrics might be used in road crash analysis. |
| **Author(s)** | P.O. Barnard |
| **Performing Organisation (Name and Address)** | Australian Road Research Board<br>500 Burwood Highway<br>Vermont Victoria |
| **Sponsor (Name and Address)** | Federal Office of Road Safety<br>GPO Box 595<br>CANBERRA ACT 2601 |
| **Available from (Name and Address)** | Federal Office of Road Safety<br>GPO Box 595<br>CANBERRA ACT 2601 |


**Abstract** The purpose of this report is to review the potential for applying some recent advances in econometrics, particularly in the analysis of categorical data, to road accident research.

This report should interest

(a) those involved in the analysis of road accident data
(b) those involved in road accident research
(c) statisticians and econometricians with an interest in the analysis of categorical date.

The major conclusions are that the set of statistical techniques termed quantal response models, and extensions to these techniques, appear especially suited to the analysis of road accident data in that they possess an intuitively appealing theoretical framework, they can be estimated using individual data, thus providing certain statistical advantages and survey economies, and they offer considerable flexibility.



**Keywords** Acccident / crash / severity (accident injury) / safety / accident exposure / mathematical model / probability / sample (stat) / distribution (stat) / log-linear* / latent variable* / logit* / quantral response* / econometrics* /

* Non IRRD keywords

**NOTES:**

(1) FORS Research reports are disseminated in the interests of information exchange.

(2) The views expressed are those of the author(s) and do not necessarily represent those of the Commonwealth Government.

(3) The Federal Office of Road Safety publishes two series of research report.

(a) reports generated as a result of research done within the FORS are published in the OR series;

(b) reports of research conducted by other organisations on behalf of the FORS are published in the CR series.

**Acknowledgements**

The printing and publication of this report were brought to completion with the help of David Hensher and Nerida Smith at the Transport Research Group at Macquarie University and Marcus Wigan at ARRB.

# HOW SOME ADVANCES IN ECONOMETRICS MIGHT BE USED IN ROAD CRASH ANALYSIS

**By**
**P.O. Barnard**

Australian Road Research Board
500 Burwood Highway
Vermont South
Victoria

This study was funded by the Federal Office of Road Safety

CONTENTS

## EXECUTIVE SUMMARY.

The question addressed in this report is: 'How can categorical road crash data be better analysed?'. Categorical data items are those items measured using the nominal or ordinal scales. Under the nominal scale, numbers are simply used as a classificatory device. An example is when the number 0 is used to signify road based trips which do not involve a crash and the number 1 is used to denote those trips on which a road crash occurs. With the ordinal scale, rank is established, but the distance between any two numbers in the scale is of unknown size. An example of ordinal scale measurement is the injury classification scheme devised by the National Safety Council in the United States. This scale is: 1 = no injury, 2 = a non-visible injury - a complaint of pain without visible signs of injury or momentary unconsciousness, 3 = minor visible injury - an abrasion, bruise, swelling, etc., 4 = serious injury - any condition that requires the victim to be carried from the scene of the crash, and 5 = a fatal injury. Continuous variables, on the other hand, are measured on the interval or ratio scales.

When the data item to be analysed is continuous, appropriate statistical techniques are regression analysis and analysis of variance. Both these techniques have been used extensively in road crash research. Until relatively recently, however, there have not been an analogous set of techniques to analyse categorical data, at least that have been incorporated in widely available software. This situation is now changing, with increasing attention being focused on a set of techniques known as quantal response (QR) models. Two members of this model family, particularly suited for the analysis of categorical road crash data are log-linear models and latent variable models.

Log-linear models are useful in establishing a pattern of association between a number of variables. More precisely, log-linear models provide information on the form of correlation between variables; for example, the form of correlation between road crash occurrence (0,1), driver age (say, less than 25 years = 0, and 25 years or more = 1), and sex of driver (say, male = 0, and female = 1).

Latent variable models are helpful when a more causative structure needs to be established between sets of variables. These models take the perspective that underlying the observed categorical variable measuring the phenomena under study, is an unobserved (or latent) continuous variable. Further, from this perspective, the categorical variable alters states (e.g. from no crash occurrence = 0, to a crash occurance = 1) as the underlying continuous variable crosses a threshold. The assumption of an underlying continuous variable allows analysis to proceed with latent variable models in an analogous fashion to regression methods.

To provide a concrete example, underlying (that is, latent in) most road crashes is risky driver behaviour. All driving, however, involves risks. A crash only occurs (that is, the categorical variable measuring crash occurance only switches from 0 to 1) when risks exceed a given threshold level. Moreover, the threshold level of risk that must be exceeded for a crash to result will vary according to conditions applying at the time. In a statistical analysis driver risk might be measured by travel speeds, alcohol consumption, and previous driving convictions. Similarly, the critical threshold level may be considered as dependent upon weather and road conditions, vehicle characteristics and driver experience. The latent variable model framework allows quantitative information to be obtained on how much each of these factors contributes to road crashes.

Latent variable models are also useful in analysing data from non-random samples. They allow analyses based on a segment of the population to be validly applied to the population as a whole. This means that economies in data collection can be realised, opening up the possibility of profitable utilisation of existing data from partisan sources, such as that contained in car insurance records.

Many software packages incorporating the techniques discussed above are now available. The latest version of the SPSS package, SPSS-X, contains routines to estimate log-linear models. A number of specialised econometric and psychometric software packages exist for analysing categorical data using latent variable models.

In summary, advances in multivariate statistical techniques for analysing categorical data offer a useful addition to the road crash researcher's 'toolbox'. Judiciously applied, these techniques can provide economies in data collection as well as increasing the relevance and accuracy of information supplied to policymakers.

## 1. REPORT OUTLINE.

This report serves as an introduction to the use of quantal response (QR) models in road crash analysis. The first task of the report is to demonstrate the place of QR models within an overall approach to statistical data analysis (Section 2). Following this, in Section 3, is a brief discussion of log-linear models (LLM), a particular type of QR model which has received some use in road crash analysis. In Section 4 latent variable models are reviewed. These models assume that underlying the observed categorical variable under scrutiny is an unobserved, latent, continuous variable. Changes in the value of the categorical variable, from the perspective offered by this model family, arise as the unobserved continuous variable crosses threholds. A member of the latent variable model family possessing attractive properties is the linear logit model. Variants of the linear logit model form the major focus of this report. Also in Section 4, a relationship is established between linear logit models and log-linear models. In Section 5 some extensions to the models considered in Section 4 are outlined. Section 6 contains three empirical demonstrations of the use of QR models in road crash research. Available software is reviewed in Section 7.

Throughout the first 5 Sections of this report, in the main, models are presented in a non-specific data context. This is achieved by applying the models to data arranged in the form of contingency tables. An advantage of this approach is that contingency tables are quite familiar to the road crash researcher. Furthermore, most data related to road crashes can be arranged in the form of contingency tables. This holds both for the detailed psychometric or ergonomic data on driver behaviour collected under experimental or real road conditions and for the sketchy but global data found in mass crash data records. Examples are used extensively to convey an understanding of the models. Because the

analysis context is as described above, however, these examples can easily be generalised.

The coverage of models reviewed in this report is synoptic. Entire books have been written on log-linear models (e.g. Bishop et al. 1975, Haberman 1978 and 1979) and on QR models which assume an underlying continuous variable (e.g. Hensher and Johnson 1981, Maddala 1983, Train 1985 and Wrigley 1985). The reader is referred to these works for further details on the models.

## 2. A CLASSIFICATION OF MULTIVARIATE STATISTICAL TECHNIQUES.

A useful framework for classifying statistical problems has been developed by Wrigley (1979, 1981, 1985). An enhanced version of Wrigley's classification scheme is shown in Table 1. The basic division in Table 1 is between response (or dependent) variables and explanatory (or independent) variables. Response variables are the variables of primary interest and can be considered as being generated by the explanatory variables which are selected on the basis of theory, previous empirical results or a priori reasoning. Often the statistical analysis consists of hypothesis testing that certain explanatory variables are influential in determining the state of the response variable.

Table 1 further classifies response and explanatory variables as being continuous, categorical or mixed. Categorical variables identify individuals, households, etc. as belonging to particular categories. They represent a classificatory mechanism. 'Mixed', in the context of explanatory variables, refers to situations where both categorical and continuous variables lie within the explanatory variable set. For response variables 'mixed' refers to situations where the response variable has both a categorical and continuous component. An example of the latter is the cost of reported crashes. The categorical component is whether or not a crash is reported. The continuous component is the cost of crashes *given* that they have been reported.

TABLE 1

CLASSES OF STATISTICAL PROBLEMS

| Response Variables | Explanatory Variables | | | |
|---|---|---|---|---|
| | Continuous | Mixed | Categorical | |
| Continuous | (a) | (b) | (c) | |
| Categorical | (d) | (e) | (f) | (g)* |
| Mixed | (h) | (i) | (j) | |

*Note: For cell (g) all variables are categorical and no distinction is made between response and explanatory variables.

Categorical variables are measured at a low level of precision using nominal or ordinal scales. Measurement at its weakest level exists when numbers or other symbols are used merely as a classification aid. When numbers or other symbols are used to identify the groups to which various objects belong, the numbers or symbols constitute a nominal or classificatory scale (Segal 1956). If objects are classified into just two groups the categorical variable is referred to as being dichotomous (e.g. a crash has or has not occurred). If more than two classificatory groups are used and no relation exists between the groups, then the categorical variable is referred to as being unordered polytomous. An example of the latter type of nominal scale measurement is the assigning of different numbers to denote urban crashes at intersections (say 0), between intersections along arterials (say 1) and between intersections along local streets and collectors (say 2).

At a slightly higher level of measurement precision is the ordinal scale. This scale is useful when categorised objects are not only different to each other, but stand in some kind of relationship. Typical relations amongst classes are (e.g. Segal

1956): higher, more preferred, more difficult, more disturbed, more mature, better driving performance, etc. The abbreviated injury scale (AIS) devised by the National Safety Council in the United States is an example of ordinal scale measurement. With ordinal scale measurement the categorical variable is referred to as being 'ordered polytomous'.

The interval and ratio scales allow a further enhancement of measurement precision. The interval scale retains all of the characteristics of the ordinal scale and has the additional feature that the distances between any two numbers of the scale are of known size. The ratio scale retains all the characteristics of the interval scale and in addition has a true zero point. The $ cost of a crash is an example of ratio scale measurement. Variables measured using interval or ratio scales are continuous and those measured using nominal or ordinal scales are categorical.

Conventional regression analysis is an appropriate technique for situation (a) where the response variable and all explanatory variables are continuous. Regression models which include some dummy exogenous variables are also appropriate for situation (b). Regression models of the type that might be applied to cell (c) where all explanatory variables are categorical can be shown to be equivalent to traditional analysis of variance models. Situations of type (d) - (g) are amenable to analysis by QR models. The remaining data situations can be analysed using integrated regression/QR model systems.

The change in emphasis from cells (d) and (e) to (f) may seem slight, but has a fundamental implication for the type of model that can be used. In particular, when all explanatory variables are categorical the raw data can be aggregated without loss of information and the response variable can be transformed into a probability estimate. In contrast, when explanatory variables are all continuous or mixed, the raw data cannot be grouped (without a loss of information) and a probability statement must be inferred directly from the categorical response variable.

To illustrate this point consider data collected from a sample of Q car drivers with a dichotomous nominal response variable defined as 0 if the individual was involved as a driver in a road crash over the course of the survey period and 1 otherwise. Suppose also two categorical explanatory variables were defined, one denoting the sex of the individual (-1 = male, +1 = female) and the other denoting the individual's age (-1 = less than 25 years old, +1 = aged 25 years or more). With this data each individual can be assigned to one of 4 groups based on the values of the explanatory variables:

g = 1 if male and $<$ 25 years,
g = 2 if male and $\geq$ 25 years,
g = 3 if female and $<$ 25 years,
g = 4 if female and $\geq$ 25 years.

A table might then be constructed, such as shown in Table 2 where the numbers within the table represent the number of sampled individuals falling into each group. The elements of this table can be transformed directly into probability terms such as those shown in Table 3. The numbers shown in Table 3 represent probability estimates of observing an individual with the characteristics of group g *and* that individual being involved (not involved) in a road crash during the survey period. These probability estimates are calculated by dividing the number of cell observations by the total number of observations. Throughout the text these probabilities are denoted in a number of ways, most commonly by $P_{rg}$ where the r can refer generally to the rth level of one variable (say R) and g to the gth level of another variable (say G), but tend to be used specifically to refer to the rth response and the gth group[1]. Longer methods of denoting these

[1]To be strictly correct, since the probability terms in Table 3 are estimates they should be denoted by $\hat{P}_{rg}$. The symbol ^ is conventionally used to refer to an estimated value, as opposed to the true population value.

TABLE 2

CONTINGENCY TABLE OF CRASH INVOLVEMENT BY AGE AND SEX

| | Male | | Female | | Total |
|---|---|---|---|---|---|
| | < 25 years | ≥ 25 years | < 25 years | ≥ 25 years | |
| crash involvement | 60 | 120 | 52 | 126 | 358 |
| no crash involvement | 820 | 3142 | 1010 | 4670 | 9642 |
| total | 880 | 3262 | 1062 | 4796 | 10000 |

TABLE 3

BASIC PROBABILITY ESTIMATES, $\hat{P}_{rg}$, FROM THE CONTINGENCY TABLE OF CRASH INVOLVEMENT BY AGE AND SEX

| | Male | | Female | | Total |
|---|---|---|---|---|---|
| | < 25 years | ≥ 25 years | < 25 years | ≥ 25 years | |
| crash involvement | 0.006 | 0.012 | 0.005 | 0.013 | 0.036 |
| no crash involvement | 0.082 | 0.314 | 0.101 | 0.467 | 0.964 |
| total | 0.088 | 0.326 | 0.106 | 0.480 | 1.000 |

probabilities are Prob(A,B) which refers to the probability of observing particular levels for variables A and B and Prob(A = i, B = j) which refers to the probability of observing the ith level for variable A and the jth level of variable B.

Another way of viewing the data in Table 2 is, *given* that an individual has the characteristics of group g, what is the probability of that individual being involved in a road crash? These probability estimates can be calculated from Table 2 by dividing the number of cell observations by the *column* total number of observations. Throughout the text these probabilities are denoted by $P^*_{rg}$. For the data of Table 2 the probability estimates, $\hat{P}^*_{rg}$, are shown in Table 4. These probabilities can also be derived from the $P_{rg}$ as $P^*_{rg} = P_{rg} / \sum_r P_{rg}$. For example, *given* that an individual is female and aged less than 25 years, the probability of that individual being involved in a crash, from Table 3 (and using 4 decimal places to avoid rounding error), is $P^*_{0,1} = 0.0052 / (0.0052 + 0.1010) = 0.0490$, which is the same as the number appearing in Table 4, calculated by dividing the number of cell observations by the column total (i.e. by dividing 52 by 1062). A shorthand method of writing $\sum_r P_{rg}$ is $P_{+g}$. Because the probability estimates in Tables 3 and 4 are continuous (but bounded by 0 and 1), they can be directly used in estimating the impact of age and sex on crash occurrence.

A simpler version of Tables 2 - 4 involving a categorisation of crash involvement by only age is shown in Tables 5 - 7. Tables 2 - 7 are used extensively as examples in this report.

Suppose, however, that there was also strong evidence that a continuous variable 'distance travelled as driver during the previous year' exerted an influence on the likelihood that an individual would be involved in a crash. One method of accommodating this variable into an analysis of crash involvement would be to segment distance travelled into a number of

TABLE 4

CONDITIONAL PROBABILITY ESTIMATES, $\hat{P}^{*}_{rg}$, FROM THE CONTINGENCY TABLE OF CRASH INVOLVEMENT BY AGE AND SEX

| | Male | | Female | |
|---|---|---|---|---|
| | < 25 years | ≥ 25 years | < 25 years | ≥ 25 years |
| crash involvement | 0.068 | 0.037 | 0.049 | 0.026 |
| no crash involvement | 0.932 | 0.963 | 0.951 | 0.974 |
| total | 1.000 | 1.000 | 1.000 | 1.000 |

TABLE 5

CONTINGENCY TABLE OF CRASH INVOLVEMENT BY AGE

| | Age < 25 years | Age ≥ 25 years | Total |
|---|---|---|---|
| crash involvement | 180 | 178 | 358 |
| no crash involvement | 3962 | 5680 | 9642 |
| total | 4142 | 5858 | 10000 |

TABLE 6

BASIC PROBABILITY ESTIMATES, $\hat{P}_{rg}$, FROM THE CONTINGENCY TABLE OF CRASH INVOLVEMENT BY AGE

| | Age < 25 years | Age ≥ 25 years | Total |
|---|---|---|---|
| crash involvement | 0.018 | 0.018 | 0.036 |
| no crash involvement | 0.396 | 0.568 | 0.964 |
| total | 0.414 | 0.586 | 1.000 |

TABLE 7

CONDITIONAL PROBABILITY ESTIMATES, $\hat{P}^{*}_{rg}$, FROM THE CONTINGENCY TABLE OF CRASH INVOLVEMENT BY AGE

| | Age < 25 years | Age ≥ 25 years |
|---|---|---|
| crash involvement | 0.044 | 0.030 |
| no crash involvement | 0.956 | 0.970 |
| total | 1.000 | 1.000 |

distinct categories and then proceed along the lines outlined above. However, the cut-off points could only be arbitrarily determined. Further, we might expect distance travelled (in contrast to age, perhaps) to exert a continuously increasing impact on the probability of crash involvement. Any categorization would then involve a loss of information. Ideally the individual observations should be retained with the tabular form being of the kind shown in Table 8.

In Table 8 the explanatory analysis variables (age, sex and now distance travelled) are not shown. This should not, however, present any impediment to an understanding of the analysis environment. Rather we know that these explanatory variables are associated with the individuals we have surveyed. We can therefore summarily describe the data simply by referring to these individuals. The total number of individuals surveyed is indexed by Q. (For Tables 2 - 7, Q = 10000.) In the remainder of this report we tend to use the letter 'q' to refer to any particular individual. It is likely that each individual surveyed will possess a unique set of values for the explanatory variables. One individual, for example, might be male, less than 25 years old and have travelled 15,132 kilometers in the past year; another individual while falling into the same sex and age group may have travelled slightly further, say 15,725 kilometers. There are only two possible outcomes for the response variable, however - either the individual will have been involved in a crash or will not have been involved. When an individual has been involved in a crash a 1 is recorded in the 'crash involvement' row of Table 8 and a 0 recorded otherwise. A similar recording system applies to the 'no crash involvement' row. With this recording system, Table 8 can easily be generalized to polytomous response variables as has been done in Table 9. Here the total number of possible response outcomes is indexed by R. It is clear that to analyse the data of Tables 8 and 9 an estimation procedure is required that utilizes the discrete individual observations.

In this report a family of models is introduced that enable sophisticated analysis of data of the type displayed in Tables 2 -

TABLE 8

TABULAR FORM FOR A DICHOTOMOUS RESPONSE VARIABLE AND CONTINUOUS OR MIXED EXPLANATORY VARIABLES

| | Individual Observations | | | |
|---|---|---|---|---|
| | 1 | 2 | | ...Q |
| crash involvement | 1 or 0 | 1 or 0 | | 1 or 0 |
| no crash involvement | 1 or 0 | 1 or 0 | | 1 or 0 |
| total | 1 | 1 | . . | 1 |

TABLE 9

TABULAR FORM FOR A DICHOTOMOUS RESPONSE VARIABLE AND CONTINUOUS OR MIXED EXPLANATORY VARIABLES

| Response Categories | Individual Observations | | | |
|---|---|---|---|---|
| | 1 | 2 | | ...Q |
| 1 | 1 or 0 | 1 or 0 | | 1 or 0 |
| 2 | 1 or 0 | 1 or 0 | | 1 or 0 |
| 3 | 1 or 0 | 1 or 0 | | 1 or 0 |
| R | 1 or 0 | 1 or 0 | | 1 or 0 |
| total | 1 | 1 | . . | 1 |

9. In the next section one member of this family, log-linear analysis, is described. Classical log-linear analysis may be thought of as applying to type (g) problems in Table 1. This is because classical log-linear analysis makes no distinction between response and explanatory variables, in effect treating all variables as response variables. A requirement of classical log-linear analysis is that all variables are categorical so that the data can be reduced to the form shown in Tables 2 - 7. When the discrete individual observations must be directly analysed, as in data situations described by cells (d) and (e) in Table 1 (and by Tables 8 and 9), an appropriate set of statistical techniques are the so called 'latent variable' models (LVMs). These models originated in the biometrics literature, but were further developed by a number of econometricians, particularly McFadden (e.g. McFadden 1974, 1978, Manski and McFadden 1981). It is also shown how LVMs may be applied to cell (g) type problems. Cell (h) - (j) problems can be analysed using integrated latent variable / regression model systems. An overview of this emerging area with potential road crash research applications is provided in Section 5.

Before concluding this Section on classes of statistical problems and accompanying analysis techniques, it is worth noting that often a research problem can be approached in a number of ways and data selected or manipulated to suit the approach. Essentially the approach selected should depend on the required accuracy and detail needed to satisfactorily address the problem and the costs of data collection and analysis associated with each possible approach.

To provide a concrete example, suppose a need arose to ascertain the effect of seat belt usage on road crashes. A broadbrush, aggregate data level approach could be conducted along the following lines. Firstly, data on individual crashes contained on mass crash data tapes might be aggregated over time, geographical regions or both. The effect of this process is to render what is a discrete variable at a disaggregate level (i.e.

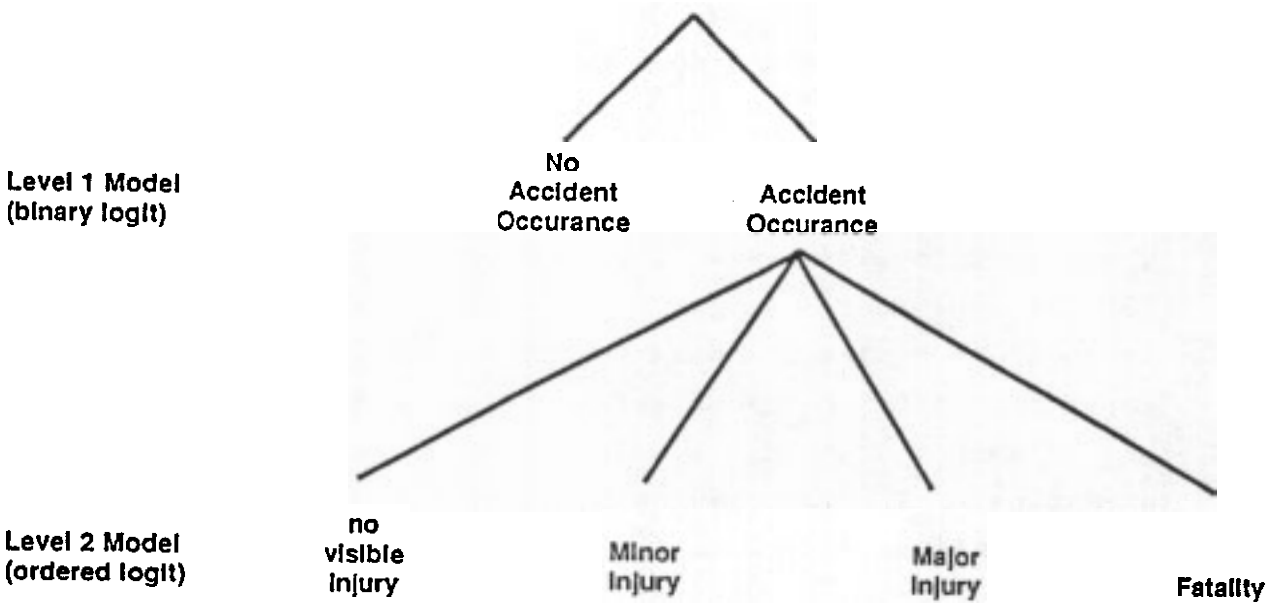
crash / non-crash defined at some level of severity) into a continuous variable at an aggregate data level (i.e. number of crashes per year, per State, etc.). Data on aggregated explanatory variables such as vehicle kilometers travelled, proportion of the population falling into various age categories and proportionate use of seat belts are available in survey information provided by Government bodies and private agencies. A regression analysis might then be applied to the data described above. Under this approach the effect of seat belt usage on road crash injuries and deaths would be measured by observing the size, sign and statistical significance of the parameter attached to the seat belt variable in the regression equation.

A more detailed disaggregate data level approach is described in Figure 1. Firstly, a model is developed of the probability of an individual being involved in a road crash at any level of severity. Secondly, given that an individual has been involved in a road crash, a model is developed of the probability of the individual sustaining no injury, a minor injury, a major injury or being killed. Both models would contain a binary variable indicating whether the individual was wearing a seat belt. In the Level 1 model this variable in effect measures the impact of seat belt usage on road crash involvement and thus drives at the heart of the risk homoeostasis debate. In the Level 2 model this variable measures the role of seat belts in reducing injury severity once a crash has occurred. That is, the one integrated modelling framework permits estimates to be made of the effect of seat belts on the overall number of crashes, their effect on the number of minor injuries, their effect on the number of major injuries and their effect on the number of deaths. Major and minor injuries could be broken down by type if desired.

From the example provided it may be observed that the comparative advantage held by the aggregate data level approach lies in its minimal cost. The comparative advantage held by the disaggregate data level approach lies in the extra information provided to the policymaker and circumvention of a number of statistical problems associated with the use of aggregated data.

FIGURE 1

SCHEMA FOR A PROBABILISTIC MODEL OF ACCIDENT OCCURRENCE AND INJURY SEVERITY



These problems may be of such a magnitude to destroy the validity of results obtained from aggregated data.[2]

## 3. A DESCRIPTION OF LOG-LINEAR MODELS.

The starting point for a detailed consideration of QR models is log-linear models applied to cell (g) type problems. As a specific example we use Table 5 and assume that no distinction is made between response and explanatory variables. The initial reaction of many researchers faced with Table 5 and an assignment to test for a relationship between the variables would be to apply a chi-square test. It is shown in this Section that log-linear models embrace the chi-square test of simple independence but may also cover more complex tests of interdependence between variables.

Consider the situation where two variables A and B (e.g. crash involvement and age in Table 4) are independent. Then the joint probability of an observation belonging to the ith category of variable A and the jth category of variable B, $P_{ij}$, can be expressed as the product of two marginal probabilities; that the observation falls into the ith category of variable A and that the observation falls into the jth category of variable B:

$$\bar{P}_{ij} = \text{Prob}(A = i) \times \text{Prob}(B = j) \qquad (1)$$

Essentially the chi-square test involves computing from equation (1) the frequencies to be expected in each cell of the contingency table under a hypothesis of independence and comparing these values with the observed cell frequencies. The chi-square statistic is of the form:

[2]These statistical issues are covered in most texts. For example, see Kmenta 1971, pp. 322 - 336 and Maddala 1977, pp. 268 - 274. De Donnea 1971 contains a striking example, in the context of travel demand, of the pitfalls concomitant with the use of aggregated data.

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(QP_{ij} - Q\bar{P}_{ij})^2}{Q\bar{P}_{ij}} \qquad (2)$$

where $\bar{P}_{ij}$ are the computed 'independence' probabilities from equation (1), $P_{ij}$ are the observed cell probabilities and Q is the (randomly drawn) sample size.

The log-linear model provides an alternative method of testing the independence hypothesis. From equation (1) and taking the natural logarithm of both sides we have:

$$\log (\bar{P}_{ij}) = \log \{ \text{Prob}(A = i)\} + \log \{ \text{Prob}(B = j)\} \qquad (3)$$

When A and B are dichotomous, taking values +1 and -1,[3] equation (3) can be re-expressed as (see e.g. Bishop et al. 1975):

$$\log (\bar{P}_{ij}) = u + u_1 A + u_2 B \qquad (4)$$

where:

[3]From the discussion in Section 2 it is clear that any two numbers can be assigned to dichotomous nominally measured variables. Elsewhere in this text (Table 8 and Section 6, in particular) either dichotomous response variables or both dichotomous response and explanatory variables are assigned the numbers 0 and 1. Here it is convenient to use the numbers +1 and -1. That is, for crash involvement (variable A) -1 is used to denote crash involvement and +1 is used to denote no crash involvement. Similarly for age (variable B) -1 is used to denote $<$ 25 years and +1 for $\geq$ 25 years.

$$u = \frac{1}{IJ} \sum_{i,j} \log (P_{ij})$$

$$u_1 = \frac{1}{J} \sum_{j} \log (P_{ij}) - u$$

$$u_2 = \frac{1}{I} \sum_{i} \log (P_{ij}) - u$$

which is conventionally referred to as the log-linear model of independence for two variables. The model is reminiscent of an analysis of variance model with u fulfilling a similar role to the 'grand mean', $u_1$ to the 'main effects of variable A' and $u_2$ to the 'main effects of variable B'. The log-linear model of equation (4) for the 2 x 2 contingency table can be generalised by adding an interaction term between variables A and B so that:

$$\log (P_{ij}) = u + u_1 A + u_2 B + u_{12} AB \qquad (5)$$

where $u_{12} = \log P_{ij} - \left[ \frac{1}{J} \sum_{j} \log (P_{ij}) + \frac{1}{I} \sum_{i} \log (P_{ij}) - u \right]$.

This is the most general form of model for the two dichotomous variable case and is known as the saturated log-linear model. In this model there are as many parameters as cells in the contingency table. As a result the model will perfectly predict the observed cell probabilities. The difference in predictions between the models of equations (4) and (5) is indicative of the traditional chi-square 'independence' hypothesis. With equations (4) and (5) being estimated by maximum likelihood using an algorithm such as Newton-Raphson, statistically the acceptability of this hypothesis may be determined by applying a log likelihood ratio test.[4] Note,

---

[4]Maximum likelihood estimation of log-linear models is discussed by Bishop et al. 1975, Chapters 3 and 5 and by Haberman 1978 and 1979. The log-likelihood ratio test is detailed in virtually all texts which cover maximum likelihood estimation. For example, see Bishop

however, that the traditional chi-square test of independence is just one of a number of hypotheses that might be considered when applying log-linear models to a 2 x 2 contingency table. Additional restrictions to $u_{12} = 0$ that might be applied are $u_1 = 0$, $u_2 = 0$, and $u_1 = u_2 = 0$. The last set of restrictions (i.e. $u_{12} = u_1 = u_2 = 0$) is equivalent to an hypothesis of equal probabilities in all cells of the contingency table, while the former two correspond, respectively, to categories of the A variable being equally probable and categories of the B variable being equally probable. Parsimony of modelling effort suggests that a hierarchical approach be adopted in which the saturated model is first estimated, the restriction $u_{12} = 0$ then tested and only if this is accepted should further restrictions (in order $u_1 = 0$, $u_2 = 0$ and $u_1 = u_2 = 0$) be applied.

The advantages offered by the log-linear approach become ever more pronounced as the number of variables considered increases. For example Table 10 details the hierarchical set of log-linear models relevant for a three dimensional (2 x 2 x 2) contingency table - such as that given in Table 2.

## 4. A DESCRIPTION OF LATENT VARIABLE MODELS.

In this section a family of models termed in the statistics literature 'latent variable models' is briefly reviewed. The focus of attention is on two members of this family; the binary and multinomial logit models.

### 4.1. THE LINEAR LOGIT MODEL APPLIED TO GROUPED RESPONSE DATA.

To introduce the linear logit model we return to Table 2, but in contrast to Section 3 now explicitly view crash / non-crash as the response variable and age and sex as explanatory

---

et al. 1975, pp 125-130, Haberman 1979, p86, and Maddala 1977, p177.

TABLE 10

THE HIERARCHICAL SET OF LOG-LINEAR MODELS FOR A 2 x 2 x 2 CONTINGENCY TABLE

| Parameters Constrained to equal 0 | Number of models of this type* | Model specification [log $(P_{ij})$ = ...] |
|---|---|---|
| 0 | 1 | $u_0 + u_1A + u_2B + u_3C + u_{12}AB + u_{13}AC + u_{23}BC + u_{123}ABC$ |
| 1 | 1 | $u_0 + u_1A + u_2B + u_3C + u_{12}AB + u_{13}AC + u_{23}BC$ |
| 2 | 3 | $u_0 + u_1A + u_2B + u_3C + u_{12}AB + u_{13}AC$ |
| 3 | 3 | $u_0 + u_1A + u_2B + u_3C + u_{12}AB$ |
| 4 | 1 | $u_0 + u_1A + u_2B + u_3C$ |
| 4 | 3 | $u_0 + u_1A + u_2B + u_{12}AB$ |
| 5 | 3 | $u_0 + u_1A + u_2B$ |
| 6 | 3 | $u_0 + u_1A$ |
| 7 | 1 | $u_0$ |

*Note: In some cases the model specification shown is but one of a number of models of the general type implied by the restrictions. For example, the restrictions imposed on the third model in this table are $u_{123} = u_{23} = 0$. There are two other restrictions in this general group, viz. $u_{123} = u_{12} = 0$ and $u_{123} = u_{13} = 0$. Thus, three models are contained within this group.

variables (i.e. variables determining the probability of being involved in a road crash). With this perspective the analysis task can be refurbished as: *given* that an individual has the characteristics of group g, what is the probability of he or she being involved in a crash? That is, we are now interested in the probability of crash involvement (non-involvement) *conditional* upon knowledge of the individual's characteristics. The appropriate probability terms are the $P^{*}_{rg}$ introduced in Section 2. By predicting these probabilities we can answer questions such as: 'If circumstances *a* were to occur, what would be the implic-ations for *b*?'

In addition a different parameterisation is used for the probability terms. In particular the binary logit probability function is used which has the general form:

$$L(\mu) = \frac{\exp(\mu)}{1 + \exp(\mu)} \qquad (6)$$

where $\mu$ can be treated as the mean value of a single variable or be made a function of a group of independent variables and unknown parameters. In the case of the linear logit model this function is linear in the parameters, such that:

$$\mu = Z\delta \qquad (7)$$

where $\delta$ is a vector of parameters ($\delta' = \delta_1, \delta_2, \ldots\ldots, \delta_p$) and Z is a vector of independent variables ($Z' = z_1, z_2, \ldots\ldots, z_p$) which can include simple transformations of the 'raw' variables.[5] It can be seen from equation (6) that $L(\mu)$ is constrained to lie between between 0 and 1 (which, naturally, is a desirable constraint for a

---

[5] For those not familiar with matrix algebra $Z\delta = \delta_1 z_1 + \delta_2 z_2 + \ldots + \delta_p z_p = \sum_{\ell}^{p} \delta_\ell z_\ell$.

probability function) and takes the shape shown in Figure 2.

Applying equations (6) and (7) to Table 2 the probability, $P^*_{1_q}$, of individual q being involved in a crash may be expressed as:

$$P^*_{1_q} = L(\mu_q) = \frac{\exp(\mu_q)}{1 + \exp(\mu_q)} \qquad (8)$$

where $\mu_q = \delta_1 + \delta_2\ AGE_q + \delta_3\ SEX_q$. $AGE_q$ is a dummy variable with value -1 if the individual's age is less than 25 years and +1 otherwise and $SEX_q$ is a dummy variable with value -1 if male and +1 if female. With these definitions the information in Table 4 can be rearranged as in Table 11.

In estimating the model, equation (8) can be rewritten to produce a linear expression:

$$\log\left(\frac{P^*_{1_q}}{P^*_{2_q}}\right) = \delta_1 + \delta_2\ AGE_q + \delta_3 SEX_q \qquad (9)$$

where $P^*_{2_q}$ is the probability of not being involved in a road crash ( $= 1 - P^*_{1_q}$). Equation (9) can be estimated using ordinary regression techniques (OLS) or, more appropriately, weighted least squares (WLS) to allow for the heteroskedasticity of the error terms.[6]

For illustrative purposes, the results from applying the model of equation (9) to the data of Table 2 are shown in Table 12. Such

---

[6]Heteroskedastic error terms arise because the variance between the estimated probabilities of Table 4 and the true probabilities will not be constant but rather will depend on the probability values and sample size in each cell.

FIGURE 2

FORM OF THE LOGISTIC DISTRIBUTION

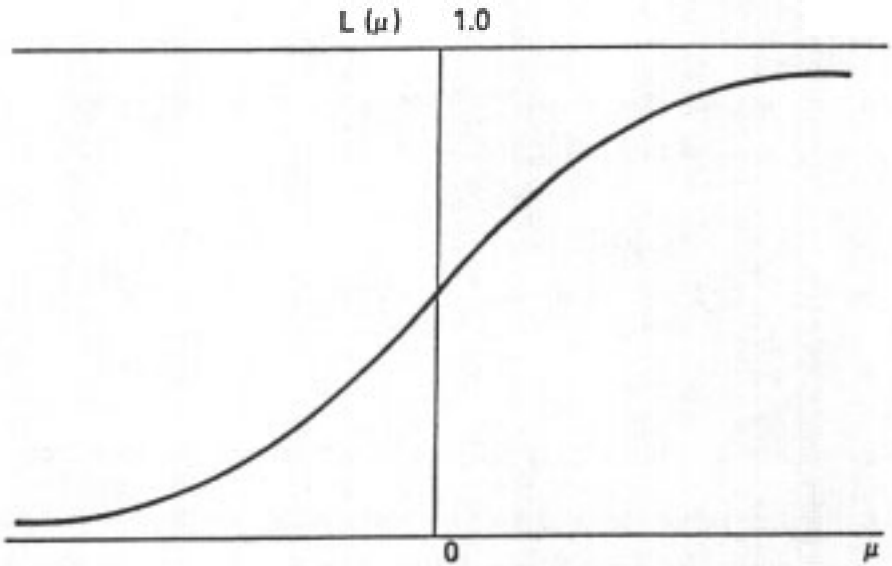


TABLE 11

REARRANGEMENT OF THE DATA OF TABLE 3 FOR LOGIT MODEL ESTIMATION

| Constant | Age | Sex | $\hat{P}_i^*$ |
|---|---|---|---|
| 1 | -1 | -1 | 0.068 |
| 1 | +1 | -1 | 0.037 |
| 1 | -1 | +1 | 0.049 |
| 1 | +1 | +1 | 0.026 |

TABLE 12

RESULTS FROM WEIGHTED LEAST SQUARES ESTIMATION OF A LOGIT MODEL USING THE DATA OF TABLES 2 AND 11

| Variable Name | Parameter Estimate | Standard Error |
|---|---|---|
| Constant | -3.1166 | 0.0041 |
| Age | -0.3246 | 0.0041 |
| Sex | -0.1796 | 0.0038 |

Note: Predicted conditional probabilities from the model of a crash occurrence, $\hat{P}_i^{\#}$, are:

age < 25 years and male, $\hat{P}_i^{\#} = 0.068$

age ≥ 25 years and male, $\hat{P}_i^{\#} = 0.037$

age < 25 years and female, $\hat{P}_i^{\#} = 0.049$

age < 25 years and female, $\hat{P}_i^{\#} = 0.026$

which replicate those of Table 11. $\hat{P}_{ig}^*$ is used to represent the conditional probability of crash involvement for an individual belonging to group g *as estimated from the basic data* (e.g. as in Table 11). $\hat{P}_{ig}^{\#}$ is used to represent the conditional probability of crash involvement for an individual belonging to group g as *estimated from the model*.

a model, for instance, may be used to obtain information of the effect of an aging population on the incidence of road crashes.

## 4.2. THE LINEAR LOGIT MODEL APPLIED TO INDIVIDUAL OBSERVATION DATA.

It has been shown in the previous Subsection how the logit model can be applied to grouped data. In the majority of real world investigations, however, the state of the response variable will be affected by the values of both categorical and continuous variables. As pointed out in Section 2 where some of the explanatory variables are continuous, if there is to be no loss of efficiency in analysis, the data must be considered at the individual response level.

There may be occasions when consideration of the data at the individual observation level is warranted even when all the explanatory variables are categorical. The illustrative data in Tables 2 and 5 reflect a situation where, with a randomly selected sample, one of the responses (i.e. crash involvement) is infrequently observed. This sort of situation is not uncommon in road crash research. A consequence is that some cells may have zero entries, particularly as the dimensions of the contingency table increase. Since the log of zero is undefined, the models of equations (4), (5) and (9), based on the proportions observed in a contingency table, become difficult or impossible to estimate. More generally, the WLS method of estimating logit models and methods normally used to estimate log-linear models tend to fail when there exist many cells with small observed frequencies.

In this Section it is shown how the logit model can be applied to individual observation data. The basic tenet of the model is that an unobserved continuous variable underlies the observed discrete responses. Essentially the method involves specifying a regression for the unobserved variable and then relating this variable to the observed responses. The net effect is that, whereas to this point the probabilities have been treated as given, an attempt is made in developing the logit model with individual

observation response data to provide an interpretation for the mechanisms giving rise to these probabilities.

Interestingly, this distinction in the treatment of discrete response data can be traced to the very foundations of modern statistical analysis. Pearson (1900) insisted that it always made sense to assume an underlying continuous variable for a dichotomy or polytomy. Yule (1900), on the other hand, chose to analyse the cross-classified data as they are and can thus be considered as the founder of the LLM school (Feinburg 1975, Maddala 1983).

### 4.2.1. The Linear Logit Model Applied to Dichotomous Individual Observation Response Data.

A convenient starting point to convey an understanding of the logit model applied to discrete response data is the data of Table 8. The response variable contained in this table is dichotomous with 0 indicating no crash involvement and 1 indicating that the individual was involved in a road crash over the course of the survey period. The observed dichotomous response is denoted below by $y_q$. The logit model assumes that underlying $y_q$ is an unobserved continuous variable, $y_q^*$. It is left to the researcher to provide an interpretation for the underlying continuous variable; however, in the context of crash involvement a reasonable interpretation for $y_q^*$ is 'risky driving behaviour'.

If $y_q^*$ were to be observed, a regression relationship could be specified as:

$$y_q^* = Z_q\delta + \epsilon_q \qquad (10)$$

where $\epsilon_q$ is an error term, and other terms are as previously defined. $Z_q$ should contain the determinants of risky driver behaviour or proxy variables for these such as socio-economic characteristics, activity time constraints, previous driving convictions, etc. However, $y_q^*$ is not observed and consequently

equation (10) cannot be estimated directly using standard regression techniques. Nevertheless, it is possible to view the observed response (crash involvement) as the outcome of the unobserved variable (risky driving behaviour) crossing an unobserved threshold value. Since $y_q^*$ is measured on an interval scale, without loss of generality this threshold value can be normalised to zero so:

$$\begin{aligned} y_q &= 1 \qquad \text{if } y_q^* > 0 \\ y_q &= 0 \qquad \text{otherwise} \end{aligned} \tag{11}$$

The probability of crash involvement can then be expressed as:

$$\text{Prob}(y_q = 1) = \text{Prob}(\epsilon_q > - Z_q\delta) \tag{12}$$

Where the $\epsilon_q$ are independently and identically distributed (iid) extreme value type 1 (EV1)[7] then the RHS of equation (12) through integration is equal to (see, for example, Johnson and Kotz 1970, Domencich and McFadden 1975) $1 - L( - Z_q\delta)$ so:

$$\text{Prob}(y_q = 1) = 1 - L( - Z_q\delta) \tag{13}$$

Also,

$$\text{Prob}(y_q = 0) = L( - Z_q\delta) \tag{14}$$

The likelihood function formed from equations (13) and (14) is:

---

[7] If the $\epsilon_q$ were assumed to be normally distributed then the resulting QR model would be a probit model rather than a logit model. This report focuses on the logit model because of its computer tractability when the response is polytomous and because the binary logit and probit models yield very similar results.

$$\lambda = \prod_{y_q=0} L(-Z_q\delta) \prod_{y_q=1} [1 - L(-Z_q\delta)] \qquad (15)$$

which can be maximized using the Newton-Raphson, Fletcher-Powell-Davidson or similar algorithms to obtain estimates for the parameter vector, $\delta$. Standard errors for these estimates can be extracted from the information matrix in the usual manner and an overall 'goodness-of-fit' statistic for the model, $\rho^2$, which is not dissimilar to the $R^2$ regression statistic, has been suggested by McFadden (1974).

The model presented above can be further refined by endogenising the critical threshold value. Suppose, as before, that a regression of risky driver behaviour can be defined by:

$$y_q^* = Z_q\delta + \epsilon_q \qquad (16)$$

However, in contrast to the previously developed model, where the critical threshold was specified as 0, now assume that the threshold is in itself determined by a number of factors of interest to the analyst, so:

$$t_q^* = X_q\beta + \bar{\epsilon}_q \qquad (17)$$

where $t_q^*$ is the critical unobserved threshold value, $\beta$ is a vector of unknown parameters ($\beta' = \beta_1, \beta_2, \ldots\ldots, \beta_K$). $X_q$ is a row vector of variables determining the critical threshold value ($X_q = x_1, x_2, \ldots\ldots, x_K$) and $\bar{\epsilon}_q$ is an error term. In analysing road crash occurrence factors likely to determine the critical threshold value that, once crossed, results in risky driver behaviour being converted into an observed crash, include vehicle character-istics, such as braking performance and acceleration profiles, driver characteristics, especially reaction time (or 'proxies' for this), road characteristics and weather conditions.

From equations (16) and (17), a crash is observed if:

$$y_q = 0 \quad \text{if} \quad Z_q\delta - X_q\beta + \epsilon_q - \bar{\epsilon}_q > 0 \tag{18}$$

Similarly a crash is not observed if the risk element of driving does not exceed the critical threshold value. That is:

$$y_q = 0 \quad \text{if} \quad Z_q\delta - X_q\beta + \epsilon_q - \bar{\epsilon}_q \leq 0 \tag{19}$$

Once again if the $\epsilon_q - \bar{\epsilon}_q$ are distributed iid EV1 then the probability expressions can be derived from the cumulative logistic function:

$$\text{Prob}(y_q = 1) = 1 - L( - Z_q\delta + X_q\beta)$$

$$\text{Prob}(y_q = 0) = L( - Z_q\delta + X_q\beta)$$

and maximum likelihood used to estimate the parameter vectors.

The process described in the previous two paragraphs would seem to closely mirror the chain of events that actually gives rise to road crashes. More generally, the concept of discrete responses arising from an unobserved continuous variable crossing a threshold value has achieved wide acceptance in a number of research fields. The model described by equations (11) - (14) was first developed in biometrics. A single explanatory variable (i.e. Z variable) was used, namely, level of dosage of insecticide, with the response variable being whether the insect died or not. Other areas of application have included union membership (e.g. Lee 1978), transport choices (e.g. Domencich and McFadden 1975), purchase of consumer durables (e.g. Cragg and Uhler 1970), birth (e.g. Heckman and Willis 1975), voting (e.g. Deacon and Shapiro 1975) and horse racing (e.g. Figlewski 1979). More complete lists are provided by Amemiya (1981), Maddala (1983) and Wrigley (1985).

4.2.2. The Linear logit Model Applied to Unordered Polytomous Individual Observation Response Data.

The model presented above can be extended to cases where the response variable is polytomous and unordered. An example, empirically developed later in this report, involves an analysis of perceived bicycle safety on commuter mode choice. In this work the response variable was mode choice and for individual q is coded as:

$y_q = 1$ if car driver,
$y_q = 2$ if car passenger,
$y_q = 3$ if bus,
$y_q = 4$ if bicycle,
$y_q = 5$ if walk.

This is an example of measurement using the nominal or classificatory scale.

Underlying each value of the y variable, following McFadden (1974), assume that there exists a continuous variable, $y^*_{qr}$, and interpret this to be the utility or level of satisfaction to be obtained from using that mode. Further, endogenise in a regression-like context the five unobserved continuous variables, as:

$$y^*_{qr} = Z_{qr}\delta + \epsilon_{qr} \qquad \text{for } r = 1,2,\ldots,5 \qquad (20)$$

Presuming economic utility maximising behaviour, for mode 1 to be chosen by individual q the utility provided by this mode must be greater than the utility provided by any of the other modes. That is, for car driver to be chosen by individual q then, $y^*_{q1} > y^*_{q2}$ *and* $y^*_{q1} > y^*_{q3}$ *and* $y^*_{q1} > y^*_{q4}$ *and* $y^*_{q1} > y^*_{q5}$. To generalise this condition, mode j will be chosen if:

$$y^*_{qj} > y^*_{qr} \qquad \text{for } r = 1, 2, \ldots,5,\ r \neq j \qquad (21)$$

Here the threshold condition has been cleverly[8] generalised even further than that specified in the two logit models previously considered.

Because a portion of the $y^*_{qr}$ are unobservable by the analyst (i.e. the $\epsilon_{qi}$s), a probability must be assigned to the use of each mode:

$$\text{Prob}(y_q = j) = \text{Prob}(Z_{qj}\delta - Z_{qr}\delta > \epsilon_{qr} - \epsilon_{qj}, \text{ for all } r \neq j) \qquad (22)$$

Where the $\epsilon_{qr} - \epsilon_{qj}$ are distributed iid EV1, equation (22) defines McFadden's conditional multinomial logit (MNL) model:

$$\text{Prob}(y_q = j) = \frac{\exp(Z_{qj}\delta)}{\sum_r \exp(Z_{qr}\delta)} \qquad (23)$$

which again can be estimated by maximum likelihood to obtain values for the parameter vector.

In the application developed later in this report, the utility yielded by each mode is made a function of the time taken to travel to work by that mode, the travel cost and certain other attributes specific to the various modes. One of the $Z_{q4}$ variables (i.e. one of the variables contributing to the utility provided by the use of bicycle for commuting) is specified as perceived bicycle safety. The statistical significance and parameter value associated with this variable indicates the influence of perceived safety on commuting bicycle use.

[8]The word 'cleverly' is used in relation to the pioneering work of Daniel McFadden in this area and not to the particular application considered in the author's study.

4.2.3. The Linear Logit Model Applied to Ordered Polytomous Individual Observation Data.

Yet another development of the basic model presented in Section 4.2.1 concerns application of the principles enunciated to ordered polytomous response data. An example of such data in crash research is the injury classification scheme refered to as the abreviated injury scale (AIS). An example of this scale is:

0 = no injury,
1 = minor injury,
2 = moderate injury,
3 = severe, but not life threatening, injury,
4 = severe, life threatening injury, but recovery is probable,
5 = critical injury, involving non-immediate death or a permanent impairment of bodily function such as paralysis, and
6 = death within 24 hours

From Section 2 this is an example of ordinal scale measurement. The categories bear a ranked relationship to one another (e.g. a moderate injury is 'worse' than a minor injury which in turn is 'worse' than no injury), but the numbers assigned do not indicate distance between categories, as would be the case for interval or ratio scale measurement. To emphasize this last point, the distance of 3 scale points between no injury and a serious injury (e.g. a broken thema) should not be taken as equivalent to the distance of 3 scale points between a serious injury and death within 24 hours. Clearly, when measuring the 'actual' severity of injuries, the distance between the latter two points in the scale is greater than the former; but this is not indicated from the assigned numerical values.

In reality injury severity does not occur in discrete intervals but rather varies continuously; it is only for ease of measurement that a limited number of injury classifications are created. Consequently it is natural to envisage a continuous

variable underlying the discrete measured values of the AIS scale. The discrete points of the AIS scale can then be viewed as arising from this unseen variable crossing, not one, but a number of threshold values.

Suppose the severity of a crash, expressed in terms of the forces acting on the vehicle occupants, is represented by $y_q^*$. Crash severity will depend on observable factors such as the change in vehicle velocity on impact, vehicle size, occupant seating position and restraint usage. Relevant variables may be placed in a vector $Z_q$ so that:

$$y_q^* = Z_q\delta + \epsilon_q \tag{24}$$

The process giving rise to the AIS scores may now be viewed in terms of $y_q^*$ crossing 6 threshold values. The first value (say $t_0^*$) will be the crossing point between no observed injury and a minor injury, the second ($t_1^*$) between a minor injury and a moderate injury, and so on. It is clear that $t_{0q}^* < t_{1q}^* < t_{2q}^* < t_{3q}^* < t_{4q}^* < t_{5q}^*$.

These threshold values may be endogenised by setting,

$$t_q^* = X_q\delta + u_q, \tag{25}$$

and relating the individual threshold values to $t_q^*$ by,

$$\begin{aligned} t_{0q}^* &= t_q^* \\ t_{1q}^* &= t_q^* + k_1 \\ t_{2q}^* &= t_q^* + k_2 \\ &\vdots \\ t_{5q}^* &= t_q^* + k_5 \end{aligned} \tag{26}$$

with $k_1 < k_2 < \ldots < k_5$.

Note that the threshold values are allowed to vary across individuals, reflecting different capacities to tolerate automobile collision forces without sustaining an (observed) injury. The $X_q$ vector would include variables such as age and sex. Concerning the former, the frailty of the elderly is well known. Given, therefore, a fixed level of crash severity, an older person would be more likely injured than a younger person (Viano et al. 1978). It is also possible that sex may play a part in the body's ability to withstand injury.

Combining (24), (25) and (26) we have:

$$y_q = 0 \quad \text{if} \quad -\infty \leq Z_q\delta + \epsilon_q < t^*_{0q}$$
$$y_q = 1 \quad \text{if} \quad t^*_{0q} \leq Z_q\delta + \epsilon_q < t^*_{1q}$$

$$y_q = 6 \quad \text{if} \quad t^*_{5q} \leq Z_q\delta + \epsilon_q < +\infty,$$

or,

$$y_q = 0 \quad \text{if} \quad -\infty \leq Z_q\delta - X_q\delta + \epsilon_q - u_q < 0$$
$$y_q = 1 \quad \text{if} \quad 0 \leq Z_q\delta - X_q\delta + \epsilon_q - u_q < k_1$$

$$y_q = 6 \quad \text{if} \quad k_5 \leq Z_q\delta - X_q\delta + \epsilon_q - u_q < +\infty.$$

The probability of observing an injury of severity level j is:

$$\text{Prob}(y_q = j) = \text{Prob}(k_j - \bar{Z}_q\bar{\delta}) - \text{Prob}(k_{j-1} - \bar{Z}_q\bar{\delta}), \qquad (27)$$

where $\bar{Z}_q$ is a row vector containing all variables in vectors $Z_q$ and $X_q$, $\bar{\delta}$ is to be similarly interpreted in parameter space, and $k_0 = 0$, $k_{-1} = -\infty$ and $k_J = +\infty$. With the error terms $\epsilon_q - u_q$ independently and identically logistically distributed the probabilities are defined by an ordered logit model:

$$\text{Prob}(y_q = j) = \frac{1}{1 + \exp(\bar{Z}_q\bar{\delta} - k_j)} - \frac{1}{1 + \exp(\bar{Z}_q\bar{\delta} - k_{j-1})}.$$

The model presented above is more detailed and realistic than other models of injury severity that have appeared in the general road crash literature. It recognizes the ordinality of injury scale data, but retains many of the advantages of regression analysis. The model provides probability estimates of injury occurring at the various scale levels in individual crashes. Also an index of the continuous underlying variable $y_q^* - t_q^*$ can be recovered once estimates have been obtained for $\beta$ and $\delta$.

The model presented here is considered in more detail later in this report. It is, however, just one of a set of probabilistic models that may be applied to ordinal data. An excellent overview of some members of this set of models is contained in McCullagh (1980) and the comments by discussants of that paper.

## 4.3. RELATING LOG-LINEAR MODELS TO LOGIT MODELS.

The log-linear model presented in Section 3 makes no distinction between response and explanatory variables, in effect treating all variables as response variables. In this Section consideration is given to how the parameters of log-linear models may be estimated using a series of logit models, and the concept of model systems is introduced.

To demonstrate the close relationship between logit models and log-linear models we take the example data of Table 2 which contains three dichotomous variables - crash involvement (A), age (B) and sex (C). The saturated log-linear model for this table is shown as the first model of Table 10. The first task of this Section is to show how the parameters of this model could have been generated by a series of logit models.

To facilitate this demonstration it is convenient to introduce a further set of notation. The probability of observing a

particular combination of values for the variables A,B, and C is denoted Prob(A,B,C). The conditional probability of observing a particular value of variable A given values for variables B and C is written as Prob(A|B,C). Analogously we write Prob(B|A,C) and Prob(C|A,B) for the conditional probabilities of variables B and C respectively, given values for the remaining variables. Following conventional use for log-linear models the dichotomous values assigned to variables A, B and C are -1 and +1.

From Table 10 the saturated log-linear model for the dichotomous 3 variable case is:

$$\log[\text{Prob}(A,B,C)] = u_0 + u_1A + u_2B + u_3C + u_{12}AB + u_{13}AC + u_{23}BC + u_{123}ABC \quad (28)$$

Also it follows that[9]:

$$\text{Prob}(A|B,C) = \frac{\text{Prob}(A,B,C)}{\text{Prob}(-1,B,C) + \text{Prob}(+1,B,C)} \quad (29)$$

Combining equations (28) and (29) yields, after simplification:

$$\text{Prob}(A|B,C) = \frac{\exp(u_1A + u_{12}AB + u_{13}AC + u_{123}ABC)}{\exp(u_1 + u_{12}B + u_{13}C + u_{123}BC) + \exp(-u_1 - u_{12}B - u_{13}C - u_{123}BC)} \quad (30)$$

noting that the u terms must sum to 0 over all categories of the variable, which for a dichotomous variable means that they are equal in value but opposite in sign.

The logit expression corresponding to equation (30) is the log of the ratio of the two probabilities associated with variable A. By setting $k = \exp(u_1 + u_{12}B + u_{13}C + u_{123}BC) + \exp(- u_1 - u_{12}B - u_{13}C - u_{123}BC)$ this expression can be derived from equation (30) as:

[9]This equation is simply a re-expression of the $P^*_{rg}$ terms.

$$\log\left[\frac{\text{Prob}(+1\,|B,C)}{\text{Prob}(-1\,|B,C)}\right] = L_{A|BC}$$
$$= u_1 + u_{12}B + u_{13}C + u_{123}BC - \log k$$
$$- (- u_1 - u_{12}B - u_{13}C - u_{123}BC - \log k)$$
$$= 2u_1 + 2u_{12}B + 2u_{13}C + 2u_{123}BC \qquad (31)$$

By a similar process,

$$L_{B|AC} = 2u_2 + 2u_{12}A + 2u_{23}C + 2u_{123}AC \qquad (32)$$

and,

$$L_{C|AB} = 2u_3 + 2u_{13}A + 2u_{23}C + 2u_{123}AB \qquad (33)$$

Equations (31) - (33) permit the estimation of 7 parameters ($u_1$, $u_2$, $u_3$, $u_{12}$, $u_{13}$, $u_{23}$, $u_{123}$) which is the number of free parameters in the saturated log-linear model (the parameter $u_0$ is determined automatically from the other parameters to ensure the probabilities sum to 1). The parameter estimates obtained from the conditional logit models will be exactly twice the estimates of the corresponding LLM.

The equivalence in parameter estimates between a series of conditional logit models and LLMs estimated from the same contingency table does not necessarily hold for unsaturated model forms. When a restriction can be imposed across all conditional logit models so that they all imply the one LLM, the resulting estimates will possess the equivalence established above and will be efficient. An example of such a restriction is $u_{123} = 0$. Some restrictions, however, will mean that the set of conditional logit models will imply more than one LLM. For example, by restricting $u_{123} = u_{23} = 0$ the set of conditional logit models is:

$$L_{A|BC} = 2u_1 + 2u_{12}B + 2u_{13}C \qquad (34a)$$
$$L_{B|AC} = 2u_2 + 2u_{12}A \qquad (34b)$$
$$L_{C|AB} = 2u_3 + 2u_{13}A \qquad (34c)$$

Whereas the latter two of these equations imply an LLM:

$$\log P_{ij} = u_0 + u_1A + u_2B + u_3C + u_{12}AB + u_{13}AC \qquad (35)$$

the former implies an LLM:

$$\log P_{ij} = u_0 + u_1A + u_2B + u_3C + u_{12}AB + u_{13}AC + u_{23}BC \qquad (36)$$

In this case matching and efficient estimates will only be obtained through joint estimation of equation system (34).

The models presented above have been taken a step further by Nerlove and Press (1973) and Schmidt and Strauss (1975) by specifying the main effects in the conditional logit models as a function of explanatory variables. Their model can be written as:

$$\log \left[ \frac{\text{Prob}(A=+1 \mid X,B)}{\text{Prob}(A=-1 \mid X,B)} \right] = X\beta + 2u_{12}B$$

$$\log \left[ \frac{\text{Prob}(B=+1 \mid Z,B)}{\text{Prob}(B=-1 \mid Z,B)} \right] = Z\delta + 2u_{12}A \qquad (37)$$

Note again the equality of the parameters attached to A and B in the two equations. *This suggests that the model, as with the LLM framework in general, is more a correlation model than a causal model* (Maddala 1983). It is possible to construct recursive causal logit model systems, but this is beyond the scope of the current report.

In summary, the intent of this rather complicated and detailed section on the relationship between LLMs and logit models has been to establish two points:

1. Firstly, to substantiate the close relationship between LLMs and logit models. The parameters of many LLMs can be estimated using logit software, treating the explanatory variables as exogenous.

2. Secondly, to distinguish between the essentially correlative role of LLMs from the more causative function of logit models as typically formulated. Heckman (1978), in particular, has argued that LLMs are not sufficiently rich in parameters to discriminate structural association among discrete variables from purely statistical association.

It should also be recalled that the introduction of continuous variables in LLMs is difficult, in contrast to their ready embodiment in a logit framework.

Finally, we note the words of Maddala (1983, p146):

'.....*the focus of analysis in the log-linear model and that of the latent variable model are different. In the log-linear model we are interested in knowing which of the explanatory variables are significant determinants of the different main effects and interactions. In the latent variable model we are interested in the effects of different exogenous variables on the unobserved latent variables, as well as which of the observed dependent (i.e. response) variables are significant indicators of the unobserved latent variables. Very often the latter question is more interesting*'.

## 5. SOME EXTENSIONS TO THE BASIC LATENT VARIABLE MODELS.

In this Section three extensions to the basic LVMs outlined in Section 4 are examined. The extensions were chosen on the basis of their high potential for profitable application in road crash research. Two of the extensions address sampling issues. In Sections 5.3 and 5.4 it is shown how LVMs can be used to correct results obtained from non-random samples so that the results reflect the population as a whole. In turn, LVMs themselves can be estimated using non-random samples and simple correction mechanisms applied to permit population inferences to be drawn. These matters are discussed in Section 5.1. Material contained in Section 5.2 covers the introduction of endogenous variables into LVMs, in effect amplifying on the concept of a causal model system alluded

to in Section 4.3. Throughout this section we attempt to keep the mathematics to a minimum, preferring instead to convey an intuitive feel for the models.

## 5.1. ESTIMATING LATENT VARIABLE MODELS WITH NON-RANDOM SAMPLES.

The tables used as examples throughout this report exhibit two characteristics that pervade many road crash research problems. Firstly, the phenomena under study may seldom occur. In Tables 2 - 7 road crashes are studied directly and these are (fortunately) relatively rare events. This problem of infrequency becomes even more acute when types of crashes (e.g. pedestrian or front-on car / truck collisions) are being studied. It is also a problem in studying some forms of deviant driver behaviour which are of interest to a road crash researcher. Secondly, often the researcher is especially interested in a minority group - e.g. drivers with a previous drink driving conviction. By definition little information will be collected on these groups in any random sample based on the population of road users. In Tables 2 - 7 younger drivers are not well represented, but these drivers may be central to reducing the road toll. In this Section two sampling alternatives to random samples are considered. They are stratified sampling and choice-based sampling.

### 5.1.1. Estimating Latent Variable Models with Stratified Samples.

To demonstrate the effect that stratified samples exert on the parameter estimates from LVMs, the example cross-tabulation of crash involvement by age is used. Rather than treating the probability terms contained in Tables 6 and 7 as estimates of the true probabilities, here it is assumed that they exactly represent the true population probabilities. To avoid confusion, Table 6 is repeated as Table 13, with this change in status explicitly recognised. In this Section another table of age Vs crash involvement is constructed based on a stratified sample. LVM parameter estimates obtained from Table 13 and this other table are then compared.

TABLE 13

POPULATION PROBABILITY DISTRIBUTION ($P_{rg}$) FOR CRASH INVOLVEMENT BY AGE

| | Age < 25 years | Age ≥ 25 years | Total |
|---|---|---|---|
| crash involvement | 0.018 | 0.018 | 0.036 |
| no crash involvement | 0.396 | 0.568 | 0.964 |
| total | 0.414 | 0.586 | 1.000 |

Before considering stratified sampling two characteristics of Tables 6 and 13 should be highlighted. Firstly, under random sampling we would *expect* the sample distribution to equate with the population distribution. That is, we would *expect* that Tables 6 and 13 would be the same, as indeed they are (by coincidence!) in *actuality*. A more formal way of writing this is $E(\hat{P}_{ij}) = P_{ij}$ [in words, the expected values of the probability estimates ($\hat{P}_{ij}$) obtained from the sample equate to the true population probabilities]. Secondly, from Table 13, given that an individual is less than 25 years old, the probability of he or she being involved in a crash is[10]:

$$\frac{\text{Probability of age < 25 years } \textit{and} \text{ being involved in an accident}}{\text{Probability of age < 25 years } \textit{and} \text{ being involved in an accident} + \text{Probability of age < 25 years } \textit{and} \text{ no accident involvement}}$$

[10] Note that the expression below is a 'long hand' version of equation (29).

or, 0.018/(0.018 + 0.396) = 0.044. Similarly, for those aged 25 years or more the probability of crash involvement is 0.018/(0.018 + 0.568) = 0.030. These are just the probability terms in Table 7, but their current use is as population values.

In stratified sampling the population under study is segmented into groups based on values of the *explanatory variables* and each group is then sampled separately. Often a uniform sample is drawn from each of the strata, however, this is not necessary and the sample drawn from each stratum can be of any size. Suppose then that a stratified sample was drawn based on the defined age groups, with equal samples (5,000 observations) being drawn from both groups. The expected numbers of observations in each cell from this stratified sample are shown in Table 14. These were calculated by multiplying the conditional probabilities of Table 7 (now treated as conditional probability terms for the population) by 5000. Because inexperienced drivers are involved in relatively more crashes, the number of crashes observed in the stratified sampling scheme will tend to be greater than the number observed in the random sample (cp. Tables 14 and 5).

The probability terms, $P_{rg}$, calculated from Table 14 are shown in Table 15. These terms could have been derived directly from the population probability distribution. As an example, given that 50% of the observations are characterised by an age of less than 25 years, the probability from the stratified sample of observing an age < 25 years *and* a crash involvement is 0.5 x 0.018/(0.018 + 0.396) = 0.022.

It has long been known that stratified samples do not effect the parameter estimates obtained from logit models and other LVMs (e.g. Bishop 1975). This result is relatively easy to demonstrate. Recall that the logit probabilities are defined by $P^{*}_{rg} = P_{rg}/P_{+g}$. Referring to the example of Tables 5 and 14, in the random sample if an individual is aged less than 25 years the probability of being involved in a crash is 0.018/(0.018 + 0.396) = 0.044 and

TABLE 14

EXPECTED NUMBER OF OBSERVATIONS IN EACH CELL FROM A STRATIFIED SAMPLE FOR CRASH INVOLVEMENT BY AGE

| | Age < 25 years | Age ≥ 25 years | Total |
|---|---|---|---|
| crash involvement | 220 | 150 | 370 |
| no crash involvement | 4780 | 4850 | 9630 |
| Total | 5000 | 5000 | 10000 |

TABLE 15

BASIC PROBABILITY ESTIMATES, $\tilde{P}_{rg}$, OBTAINED FROM THE STRATIFIED SAMPLE SHOWN IN TABLE 14

| | Age < 25 years | Age ≥ 25 years | Total |
|---|---|---|---|
| crash involvement | 0.022 | 0.015 | 0.037 |
| no crash involvement | 0.478 | 0.485 | 0.963 |
| total | 0.500 | 0.500 | 1.000 |

in the stratified sample this probability is 0.022/(0.022 + 0.478) = 0.044. Other conditional probability terms for the stratified sample are shown in Table 16. It can be seen that the probabilities used to estimate the logit model are unaffected by stratified sampling. Consequently, the parameter estimates from a logit model will be the same regardless of whether the sample was drawn on a random or stratified basis.

5.1.2. Estimating Latent Variable Models with Choice-Based Samples.

The major problem with applying random sampling to a population with the characteristics displayed in Table 13, however, is not related to the amount of information collected on individuals aged less than 25 years. Rather, very little information is collected on crashes across all age groups. For example, if the sample size was restricted to 1000 individuals, rather than the 10,000 individuals used in Tables 2 and 5, we would expect to observe only 36 crashes including only 11 crashes by persons aged less than 25 years. Clearly, once more realistic, multiple, levels of categorization were applied, many cells with zero entries would emerge.

In response to this type of problem in recent years methods have been devised to estimate LVMs using choice-based samples. In choice-based sampling the population under study is segmented into groups based on values of the *response variable* and each group is then sampled separately. It is important that sampling within each group is random. As with stratified sampling, samples drawn from each group can be of varying sizes.

To illustrate the effect of choice-based sampling, assume that from the population described by Table 13 a choice-based sample is drawn with 5,000 observations collected on individuals being involved in road crashes and 5,000 observations collected on individuals with no crash involvement. We would *expect* the number of observations collected in each cell to be as shown in Table 17. The probability estimates, $P_{rg}$, calculated from Table 17 are shown in Table 18. These terms could have been derived

TABLE 16

CONDITIONAL PROBABILITY ESTIMATES, $\tilde{P}^{*}_{rg}$, OBTAINED FROM THE STRATIFIED SAMPLE SHOWN IN TABLE 14

| | Age < 25 years | Age ≥ 25 years |
|---|---|---|
| crash involvement | 0.044 | 0.030 |
| no crash involvement | 0.956 | 0.970 |
| total | 1.000 | 1.000 |

TABLE 17

EXPECTED NUMBER OF OBSERVATIONS IN EACH CELL FROM A CHOICE-BASED SAMPLE FOR CRASH INVOLVEMENT BY AGE

| | Age < 25 years | Age ≥ 25 years | Total |
|---|---|---|---|
| crash involvement | 2500 | 2500 | 5000 |
| no crash involvement | 2054 | 2946 | 5000 |
| total | 4554 | 5446 | 10000 |

TABLE 18

BASIC PROBABILITY ESTIMATES, $\tilde{P}_{rg}$, OBTAINED FROM THE CHOICE-BASED SAMPLE SHOWN IN TABLE 17

| | Age < 25 years | Age ≥ 25 years | Total |
|---|---|---|---|
| crash involvement | 0.250 | 0.250 | 0.500 |
| no crash involvement | 0.205 | 0.295 | 0.500 |
| total | 0.455 | 0.545 | 1.000 |

TABLE 19

CONDITIONAL PROBABILITY ESTIMATES, $\tilde{P}^{*}_{rg}$, OBTAINED FROM THE CHOICE-BASED SAMPLE SHOWN IN TABLE 17

| | Age < 25 years | Age ≥ 25 years |
|---|---|---|
| crash involvement | 0.550 | 0.459 |
| no crash involvement | 0.450 | 0.541 |
| total | 1.000 | 1.000 |

directly from the population probability values by the relation $\tilde{P}_{rg} = 0.5\ P_{rg}/P_{r+}$ where the $\tilde{P}_{rg}$ are the expected cell probabilities from the choice-based sample. For example, $\tilde{P}_{0,-1} = 0.5[0.018/(0.018 + 0.018)] = 0.250$.

Unlike the case of stratified sampling, the logit parameter estimates obtained from a choice-based sample will be different from those obtained from a random sample. The conditional probability terms from the choice-based sample are shown in Table 19. It can be seen that these are quite different from those derived using the random sample (shown in Table 4). However, simple weights can be applied to the choice-based probability terms so that they become identical to the random sample probability terms.

These weights are given by:

$$\omega_r = \frac{P_{r+}}{\tilde{P}_{r+}} \qquad (38)$$

In words,

the probability of response r being observed in a randomly drawn sample

---

the probability of response r being observed in the choice-based sample

In the example of Tables 2 and 17, these weights are 0.036/0.5 for crash involvement (r = 0) and 0.964/0.5 for no crash involvement (r = 1). As a demonstration that these weights do result in the random sample and adjusted choice based sample probability estimates being equal in our example $\text{adj}(\tilde{P}_{0,-1}) = 0.25(0.036/0.5) = 0.018$ which is the same as $P_{0,-1}$. The other weighted probability estimates from the choice-based sample are shown in Table 20. The $\omega_r$ weights may also be applied when estimating logit models from choice-based sample data, so that the parameter estimates will replicate those that we would expect to

TABLE 20

BASIC PROBABILITY ESTIMATES, ADJ($\tilde{P}_{rg}$), OBTAINED FROM THE CHOICE-BASED SAMPLE SHOWN IN TABLE 17 AFTER WEIGHTING

| | Age < 25 years | Age ≥ 25 years | Total |
|---|---|---|---|
| crash involvement | 0.018 | 0.018 | 0.036 |
| no crash involvement | 0.396 | 0.568 | 0.964 |
| total | 0.414 | 0.586 | 1.000 |

obtain from a random sample. A formal proof of this is contained in Manski and Lerman (1975). It can be seen that these weights require very little extra information. All that is required is an estimate of the number of people in the population of interest falling into each response category.

Choice-based sampling would seem to have enormous potential for application in road crash research for three reasons:

1. Many phenomena of interest to the road crash researcher occur very rarely. This means that even if a random sample is very large only a few observations will be collected on the response of interest. It is under these circumstances that choice-based sampling offers significant cost savings.

2. Often very good information exists on the response variable for the population. For instance, from the mass crash data records we know the proportion of the population involved in crashes, categorised by type if necessary. Thus the need to calculate the weights, $\omega_r$, presents no impediment to using

choice-based samples.

3. In road crash research most of the data that has been collected relates to the crashes themselves or the type of driver behaviour that cause crashes. Relatively little data has been collected on normal driver behaviour or on travel that does not involve a crash. This is despite Johnson and Perry's (1980) prognosis that these latter matters are critical to an understanding of how crashes occur. That more research be devoted to normal driving behaviour was the major conclusion to emerge from Johnson and Perry's review. It is evident from the examples in this paper that the LVM framework *forces* the analyst to not only study the behaviour of interest, but also the antithesis of that behaviour. To estimate the model of crash involvement, for instance, data was not only required on those individuals who had been involved in a crash but also some data was needed on individuals with no crash involvement. With choice-based sampling new samples on normal driver behaviour or non-crash travel may be used to profitably *supplement* the data bases already in existence. For instance, to estimate the model of crash involvement, data on individuals who had been involved in a crash could be gleaned from the mass crash data tapes, with a new survey being conducted just to obtain some information on individuals who had not been involved in a crash during the study period.

### 5.1.3. Sample Sizes Required for the Estimation of Latent Variable Models.

Before departing sampling issues it is worth mentioning approximate sample sizes needed to estimate the LVMs reviewed to this point. Limited evidence, mainly acquired in the transport planning area, suggests that reliable parameter estimates can be obtained with as few as 50 to 70 observations and sample sizes in the range of 200 to 500 are more than adequate provided sampling is controlled to yield a reasonable spread of observations. These small sample sizes combined with the sampling techniques mentioned means that data collection costs for LVMs tend to be relatively

low.

## 5.2. INTRODUCING ENDOGENOUS VARIABLES INTO LATENT VARIABLE MODELS.

In Section 4.3 brief reference was made to a simultaneous model system consisting of two discrete dichotomous variables. In this Section a recursive model system is considered with one of the continuous explanatory variables in the LVM being specified as a function of a second set of variables.

Mathematically the system considered is:

$$y_q^* = Z_q\delta + \gamma V_q + \epsilon_q \quad (39a)$$

$$V_q = X_q\beta + \eta_q \quad (39b)$$

where $y_q^*$ is an unobservable continuous variable with an observable dichotomous outcome such that $y_q = 0$ if $y_q^* < 0$ and $y_q = 1$ otherwise, $Z_q$ is a vector of variables determining the state of $y_q^*$, $V_q$ is a continuous variable determining the state of $y_q^*$, $X_q$ is a vector of variables determining $V_q$, $\delta$ and $\beta$ are parameter vectors, $\gamma$ is a parameter, and $\epsilon_q$ and $\eta_q$ are error terms. Often it is convenient to assume that the $\epsilon_q$ are normally distributed leading to the probit form for the latent variable model.

Two methods are available for estimating equation system (39a+b). Combining (39a) with (39b) we have:

$$y_q^* = Z_q\delta + \gamma(X_q\beta) + \epsilon_q + \gamma\eta_q \quad (40)$$

and this can form the LVM to be estimated. Using this method it is impossible to separate the estimates of $\gamma$ and $\beta$. Alternatively, equation (39b) may be estimated using OLS and the predicted values of $V_q$ used in the latent variable model.

An obvious application of this model system in the road crash research field is for equation (39a) to be a model of crash

involvement with $V_q$ representing exposure (say, distance travelled). Very little information on exposure is available, but some data can be extracted from home interview travel surveys conducted as part of transport studies and further data may become available from the exposure study funded by Federal Office of Road Safety. This data would allow estimation of an exposure model. Amongst the explanatory variables ($X_q$) included in the exposure model would be locational factors, access to motor vehicles, and socio-economic descriptors. The next step is to use the predicted exposure values as an explanatory variable in the model of crash involvement.

It should be recognised that the same variable might appear in the $X_q$ and $Z_q$ vectors. For instance, it is possible that an individual's age will affect both the probability of being involved in a crash and the level of exposure. Is the high incidence of road crashes among the young in part due to higher levels of exposure? If so, how much? These are questions which may be addressed by the model system. Place an age variable in the kth position of the $Z_q$ and $X_q$ vectors. Then the impact of age on exposure is measured by $\beta_k$, the direct impact of age on the probability of crash involvement measured by $\delta_k$, and the indirect impact of age on the probability of crash acting through exposure is measured by $\gamma\beta_k$. Note that by just including age in a model of crash involvement, disregarding exposure, the parameter estimated will represent $(\delta_k + \gamma\beta_k)$; that is, the direct and exposure related age effects will be confused.

## 5.3. SAMPLE SELECTIVITY MODELS.

Attention in this Section is switched from the analysis of discrete response variables to the analysis of continuous response variables. In particular, the analysis of continuous response variables in non-random samples is examined. It is shown that LVMs can be linked to analyses of data from non-random samples so as to permit the results obtained from such samples to be applied to the entire population. Statistically the techniques considered fall into cells (h) - (j) of Table 1. Firstly, samples are studied

where the continuous variable of analysis interest is only observed within a limited range. Secondly, more elaborate non-random samples are studied. Both these examinations are placed within the context of regression analysis of the continuous variable. Thirdly, the role of LVMs in weighting summary statistics extracted from non-random samples is disclosed.

### 5.3.1. The Truncated Regression and Basic Tobit Models.

The type of non-random sample examined in this section is characterised by data points on the continuous variable of interest enveloped within a range that is less than the range of values exhibited by the population as a whole. For ease of exposition by way of concrete example, it is assumed that the analysis task is to discern factors contributing to the cost of property damage only (pdo) road crashes and the data source is official crash data tapes. Legislation requires pdo road crashes to be reported only if they exceed a specified cost level. It seems reasonable, therefore, to presume that no data points will be available on crashes with a cost value less than the legislative reporting limit. That is the data on crash costs will be attenuated.

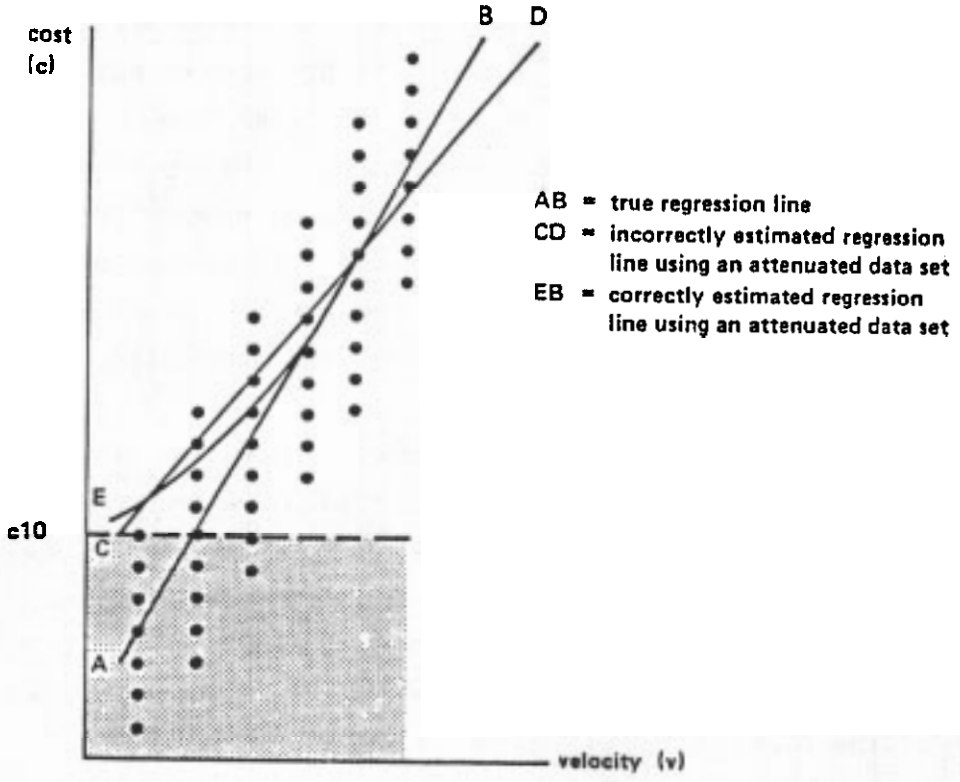
Data from a hypothetical sample containing information on crash costs is displayed in Figure 3. For illustrative purposes it is assumed that only one factor contributes to crash costs, impact velocity, but the arguments presented below are readily extended to include analyses with multiple explanatory variables. If all the data in Figure 3 is observed a linear regression model might be fitted of the form:

$$c_q = \beta_1 + \beta_2 v_q + \xi_q \qquad (41)$$

where $\beta_1$ and $\beta_2$ are parameters to be estimated, $c_q$ is the cost of the qth crash, $v_q$ is impact velocity and $\xi_q$ is a disturbance term with an expected value of zero and which is independent of $v_q$. $c_q$ and $\xi_q$ are assumed to be normally distributed. With a random sample unbiased estimates of $\beta_1$ and $\beta_2$ can be obtained by OLS. The resulting regression relationship between $c_q$ and $v_q$ is shown as

FIGURE 3

ILLUSTRATION OF REGRESSION SPECIFICATION ERROR WHEN SAMPLE POINTS WITH A C VALUE LESS THAN SOME CONSTANT ARE UNOBSERVED (FROM BERK 1983)




(Note: Shaded area represents the set of observations, lying within the population of analysis interest, but which are not captured within the available data set)

line AB in Figure 3. Problems, however, occur in estimating $\beta_1$ and $\beta_2$ when the sample is non-random.

Suppose, initially, that crashes costing less than $c_{10}$ are unobserved. Simply fitting the model of equation (41) to the remaining data, using OLS, will lead to biased estimates of $\beta_1$ and $\beta_2$. The fitted OLS regression line is shown as line CD. The true regression line, passing through the expected values of $c_q$ for each value of $v_q$, given the attenuated data set, is shown by line EB. Essentially the difference between lines CD and EB is due to specification error inherent in CD from forcing a linear relationship between $c_q$ and $v_q$ when a non-linear relationship is appropriate.

Two major consequences stem from the non-recognition of the attenuated data set shown in the non-shaded section of Figure 3. Firstly, the slope of the estimated regression line CD (i.e. the estimated $\beta_2$ value) will be less than the slope of the population regression line (i.e. the true $\beta_2$ value). This can be seen from a comparison of line AB and CD. The result, for the example given, is that the estimated effect of impact velocity on the cost of crashes will be less than the true population effect. Clearly the estimated $\beta_2$ value should not be used to infer the effect of impact velocity on crash costs for the population. The model therefore lacks *external validity*. Secondly, even if interest is restricted to road crashes with a cost greater than $c_{10}$ the estimate of $\beta_2$ will still be biased. The reason is that $\xi_q$ will be positively correlated with $v_q$. It can be seen from Figure 3 that for low values of $v_q$ there is a tendency for $\xi_q$ to be negative, while for high values of $v_q$ the converse holds. The result is violation of one of the assumptions required for OLS to yield an unbiased estimate of $\beta_2$ and the destruction of the *internal validity* of the model.

Before indicating how the correct regression line can be estimated from the attenuated data set, it is important to recognise two general classes of samples that might produce the set of observations in the non-shaded section of Figure 3. If data

points with $c_q$ values less than $c_{10}$ are totally unobserved the data set is termed '*truncated*'. Alternatively, if for data points with $c_q < c_{10}$ values for the explanatory variables can still be obtained, with $c_q$ values remaining unobserved, the resulting sample is said to be a *censored* sample.

For censored samples, unbiased estimates of $\beta_1$ and $\beta_2$ can be obtained using the basic Tobit model. This model was first discussed by Tobin (1958) and further developed by a number of researchers, principally, Amemiya (1973), Fair (1977), Goldberger (1964,1981), and Heckman (1976b). The attenuated nature of the sample is appropriately recognised in a Tobit model through inclusion of an extra regressor. From equation (41), for the censored data set characterised by $c_q > c_{10}$:

$$E(c_q|c_q > c_{10}) = - c_{10} + \beta_1 + \beta_2 v_q + E(\xi_q|\xi_q > c_{10} - \beta_1 - \beta_2 v_q) \quad (42)$$

where the E(.) term can be read as the expected value of $\xi_q$ given that $\xi_q$ is greater than $c_{10} - \beta_1 - \beta_2 v_q$. A two-stage method is available for estimating this model. Basically the method involves estimating a latent variable probit model that an observation will exceed the threshold value, $c_{10}$, then using output from this model to form the E(.) term in equation (42). Equation (42) can then be estimated using OLS. When the sample is truncated this method is unavailable since observations where $c_q < c_{10}$ are completely eliminated from the sample. Nevertheless the true regression line can be estimated from truncated samples using a full information maximum likelihood approach. Details are given in Hausman and Wise (1976, 1977).

5.3.2. Generalised Sample Selection Models with Censored Data.

The previous Subsection concerned the correct estimation of regression models when the sample was conditioned on the dependent variable exceeding a single threshold value. More complicated sample selection processes are dealt with in this section. The inclusion or exclusion of sample points is assumed to depend on a

host of factors, only some of which are known by the analyst. It is shown that even from these sorts of samples analyses can be conducted that will permit general conclusions to be drawn.

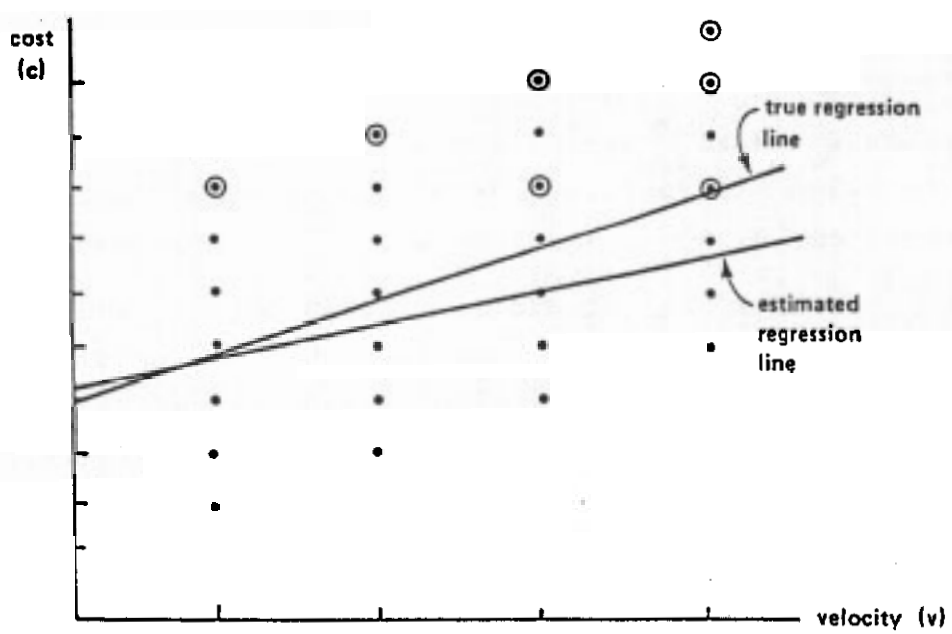
A good example of a potential application in the road crash sphere is in the analysis of insurance data. Here is a rich data source (Searle 1980). In the past, however, researchers have exhibited a reluctance to utilise this data, in large part, because of its unrepresentativeness; data is only available on those individuals who have decided to claim om insurance. This is an example of self-selection bias. Other types of selection bias are when an administrator decides to include or exclude certain observations (administrator selection) and attrition in panel data.

To analyse this, more general, sample selectivity problem assume that selection in a sample is conditioned by an unobservable variable $y_q^*$. When $y_q^*$ exceeds a threshold value an observed outcome is that observation q is included in the sample; otherwise observation q is excluded. As in Section 4 this threshold value can be specified as 0. In the insurance example, $y_q^*$ might be interpreted as propensity to claim on insurance. The analyst will be aware of some factors that condition the propensity of individual q to claim on insurance, but not all of the factors.

Following through the insurance example, suppose again that the analysis task is to examine the factors contributing to crash damage costs. Figure 4 illustrates the likely situation where $y_q^*$, the propensity to claim on insurance (and thus be included in the sample), is positively correlated with $c_q$, crash damage costs. Circled observations lying within the population of interest (all crashes) are nevertheless excluded from the sample due to no insurance claim being made. It can be seen that the application of the regression model of equation (41) to the given (insurance) sample depicted in Figure 4, using OLS, will yield biased estimates of $\beta_1$ and $\beta_2$. Statistically this can be attributed to correlation between the error term $\xi_q$ and $v_q$. From Figure 4, for low values of $v_q$ there is a tendency for $\xi_q$ to be

FIGURE 4

ILLUSTRATION OF ESTIMATED AND TRUE REGRESSION LINES WHEN INCLUSION OF SAMPLE POINTS DEPENDS ON THE VALUE OF A SELECTION VARIABLE




(Note: Circled observations lie within the population of analysis interest, but are not captured within the available data set)

smaller than for large values of $v_q$. As before both the internal and external validity of the model will be undermined. However, now it is difficult to determine whether the biased OLS estimates will understate or overstate the true causal effects.

Mathematically the regression equation for the observed data points is:

$$E(c_q|y_q^*) = \beta_1 + \beta_2 v_q + E(\xi_q|y_q^* > 0) \quad (42)$$

In turn $y_q^*$ can be segmented into an 'analyst known' portion and an unknown portion:

$$y_q^* = Z_q\delta + \epsilon_q \quad (43)$$

where $Z_q$ contains a list of factors that the analyst knows will affect the propensity to claim on insurance and $\delta$ is a vector of associated parameters. From the discussion, the observation will be included in the sample (i.e. a claim on insurance will be made) if:

$$Z_q\delta + \epsilon_q > 0$$

and will be excluded otherwise. The former case is denoted by $y_q = 1$ (when observation q from the population is included in the sample) and the latter by $y_q = 0$. Equation (42) can now be re-expressed as:

$$E(c_q|y_q^* > 0) = \beta_1 + \beta_2 v_q + E(\xi_q|\epsilon_q > -Z_q\delta) \quad (44)$$

The last term on the RHS of equation (44) can be shown to equal (see Heckman 1976b, Barnard 1986b, Cain 1975, and Muthen and Joreskog 1983 among others) $\lambda\{\phi(Z_q\delta)/\Phi(Z_q\delta)\}$ where $\phi$ is the density function of the standard normal, $\Phi$ is the distribution function of the standard normal and $\lambda$ is an estimate of the covariance between $\epsilon_q$ and $\xi_q$, so:

$$E(c_q | y_q^* > 0) = \beta_1 + \beta_2 v_q + \lambda \frac{\phi(Z_q\delta)}{\Phi(Z_q\delta)} \qquad (45)$$

A two-staged estimation procedure is to:

1. Estimate a probit model of the probability of an observation from the population being included in the sample (i.e. of an insurance claim being made). From this model estimates for the parameter vector $\delta$ are obtained.

2. Use the estimated parameter vector $\hat{\delta}$ to form the term $\frac{\phi(Z_q\hat{\delta})}{\Phi(Z_q\hat{\delta})}$

3. Apply OLS to equation (45) replacing $\frac{\phi(Z_q\delta)}{\Phi(Z_q\delta)}$ with $\frac{\phi(Z_q\delta)}{\Phi(Z_q\hat{\delta})}$ to obtain estimates for $\beta_1$, $\beta_2$ and $\lambda$. The estimates of $\beta_1$ and $\beta_2$ will be unbiased and can be used to draw inferences concerning all crashes and not just crashes about which insurance claims are made.

To estimate the probit model some information is needed on those individuals not in the sample under analysis. To profitably use insurance data, for instance, some information is needed on those individuals not making claims. The range of information collected from such individuals, however, will normally be substantially less than that collected from those individuals included in the analysis. Further, the sampling discussion in Section 5.1 suggests that this extra information to estimate the probit model can be collected at relatively little cost.

The technique outlined above can be used with any unrepresentative sample. Many of the detailed psychometric and ergonomic road crash related studies are potentially unrepresentative due to the time demands placed on participants and the voluntary nature of

participation. So too are samples of crash countermeasures where these have been placed at identified 'blackspots'. Concern has already been raised in the crash literature over the unrepresentativeness of countermeasure data (Hauer 1980). The techniques outline offer the possibility of correcting biases that may exist in these data sources.

## 5.4. WEIGHTING SURVEY STATISTICS.

Often surveys are conducted for less elaborate purposes than applying the regression type analyses that have formed a major thrust of this report. Sometimes all that is needed from a survey is summary statistics of the incidence of certain phenomena within the population. For example, information may be required on the incidence of certain kinds of driver behaviour.

Typically to tackle this sort of problem a random sample is drawn from the population of interest. If the survey is representative, frequency tabulations and like summary statistics can be calculated from the data and expanded to a population level by applying a single weight equal to the inverse of the sampling ratio. What is the best course of action, however, if, despite best intentions, the survey turns out to be unrepresentative? That is, what should be done if upon receipt of the survey returns representative checks reveal (despite the sample being drawn randomly) certain discrepancies between the socio-demographic composition of the sample and that known to exist (e.g. from census data) for the population? This situation is not uncommon because even random surveys suffer from refusals and 'no contacts'.

The methods introduced in this report suggest the following solution. First estimate a probit (or logit) model of the probability of an observation from the population being included in the sample. The data for this model could come from a synthetically constructed set based on the known population distribution and the sample distribution. Next calculate, using data on each person in the sample, the probability of that person being included in the sample. More precisely, these probabilities represent the

probability of sample inclusion of individuals from the population *with identical characteristics* to that person. The inverse of these probabilities are the optimal weights to expand the sample.

To clarify this assume that a sample and population can be perfectly described by age distribution and that only two age categories exist. Say, for the sake of consistency, these categories are age $< 25$ years and age $\geq 25$ years. Total numbers falling into each age category in the population of interest are shown in Table 21. Suppose a 50% sample was drawn from this population and per chance all those sampled aged 25 years or more responded to the survey but only half of those sampled aged less than 25 years responded to the survey. The sample age distribution is shown in Table 22. From this information we can calculate that the probability of an individual from the population with an age less than 25 years being included in the sample is 0.25 and the probability of an individual from the population aged more than 25 years being included in the sample is 0.50. The inverse of these rates are, respectively, 4.0 and 1.0. It can be seen that by applying these weights the population distribution is replicated and therefore we can place confidence in survey statistics, such as perhaps the number of crashes, for which the population distribution is unknown. With just one population parameter the probabilities are easy to calculate and there is no need to resort to a model. However, with many variables conservation of effort suggests that a modelling approach should be adopted. It should be evident from Section 4 of this report that the probabilities used above are exactly those probabilities estimated in a LVM of sample inclusion.

## 6. SOME EXAMPLES OF THE USE OF LATENT VARIABLE MODELS IN ROAD CRASH RESEARCH.

Throughout this report sparing use has been made of specific applications of LVMs in road crash research. In this Section we choose to redress this imbalance by highlighting three empirical studies that have used LVMs in road safety related analyses. Two of the studies were undertaken by the author as part of the current

TABLE 21

POPULATION FREQUENCY AGE DISTRIBUTION

| Age < 25 years | Age ≥ 25 years |
|---|---|
| 41,400 | 58,600 |

TABLE 22

SAMPLE FREQUENCY AGE DISTRIBUTION

| Age < 25 years | Age ≥ 25 years |
|---|---|
| 10,350 | 29,300 |

A.R.R.B. project funded by the Federal Office of Road Safety. The third study was conducted by researchers at Northwestern University.

## 6.1. BICYCLE COMMUTING USE AND PERCEIVED SAFETY.

In Section 4.2.2 an analysis framework was developed for studying the influence of perceived safety on choice of bicycle for commuting. A theoretical justification for applying multinomial logit to this choice analysis was provided in Section 4.2.2. In this Section empirical results from a study conducted into this issue are outlined. Further details on the study are to be found in Barnard (1986a).

The data for the study of bicycle commuting use were from a survey commissioned by the Australian Road Research Board and the South Australian Department of Transport conducted in the eastern and north-eastern suburbs of Adelaide in 1981 (Barnard 1981). Data from this survey has been used extensively by a number of researchers (e.g. Barnard (1987). Clarke et al. (1985), Wigan (1982)). In the study summarised here only a small part of the survey was utilized, namely, the journey to work (JTW) questionnaire which was satisfactorily answered by 219 full time workers.

The JTW questionnaire consisted of seven parts, each part concentrating on a particular main method of travel to work - car driver/motor cycle, car passenger, car pool, taxi, bus, bicycle and walk. The respondent was first asked to supply his usual main method of travel to work, then for other methods of travel used in the last three months. Finally, respondents were asked if there were any other modes they had considered using to travel to work. The percentage reponses for chosen and alternative modes are shown in Table 23.

For each mode mentioned responses were sought to the detailed questions about the use of that mode to travel to work. Information was obtained on times taken to travel by each mode, parking costs, vehicle operating expenses, and so on. The bicycle

TABLE 23

FREQUENCY DISTRIBUTION OF COMMUTING MODE USE AND PERCEIVED AVAILABILITY

| Mode | Per cent of respondents for whom this is the usual method of travel to work | Per cent of respondents for whom this is an alternative method of travel to work |
|---|---|---|
| Car driver | 65 | 16 |
| Car passenger | 6 | 41 |
| Car pool | 2 | 2 |
| Bus | 15 | 54 |
| Bicycle | 5 | 8 |
| Walk | 7 | 10 |

questions are reproduced in Figure 5. Eleven respondents chose bicycle as the usual method to travel to work, and it represented an alternative travel method for a further eighteen respondents. These findings are broadly consistent with those obtained from the 1977 Metropolitan Adelaide Data Base Study (Pak Poy and Associates (1978)). It is obvious that most people do not even consider bicycle as a method of travelling to work.

Of particular interest here is responses to the question concerning perceived bicycle safety. Perceived bicycle safety was measured on a 1 to 5 point scale and measured the individual's assessment of being involved in minor and major bicycle crashes. The anchor points on this scale were (1) almost no chance of having any sort of crash if a bicycle were to be used to travel to work for one year and (5) large chance of having a serious crash involving personal injury. The frequency distribution of these responses is shown in Table 24. Astonishingly, 27.6% of those who had used or considered using a bicycle to travel to work, felt that if they were to travel this way every day for one year, there was a large chance of being involved in a serious personal injury crash. In contrast only 10.3% felt that there was almost no chance of being involved in any sort of crash.

The bicycle travel safety ratings were further analysed through application of linear regression in an effort to discern systematic variation in individual ratings. It should be noted that the use of regression assumes that the ratings are interval scaled. Strictly, the ratings are ordinal scaled; however, available evidence suggests that mostly such ratings can safely be treated as though interval scaled (e.g. Kim (1975), Labovitz (1970)) and indeed this practice is common in sociological studies. Regression results are displayed in Table 25. As can be seen the only significant factor found to explain individual safety ratings is time spent bicycling on main roads. Interestingly, time spent travelling on side streets, bikeways and in parks although adding to exposure, apparently was not seen as increasing the crash risk.

FIGURE 5

BICYCLE QUESTIONS FROM JOURNEY-TO-WORK QUESTIONNAIRE

- ONLY ASK IF USUAL OR ALTERNATIVE MODE IS BICYCLE -

33. About how long does it normally take to get to work by bicycle?

(minutes) ☐☐

34. Could you next estimate about how many minutes of this trip are spent travelling along main roads, along streets, along bikeways, & through parks?

Minutes

(i) Main roads ☐☐

(ii) Side streets ☐☐

(iii) Bikeways ☐☐

(iv) Parklands ☐☐

35. How long does it usually take you to walk from where you park your bicycle to where you actually work/are educated? minutes ☐☐

36. Please indicate using the numbers 1 to 5 how safe you feel riding your bike?

1 = almost no chance of having any sort of accident

2 = some chance of having only a minor accident

3 = large chance of having a serious accident involving personal injury

4 = slight chance of having a serious accident involving personal injury

5 = large chance of having a serious accident involving personal injury

number ☐

TABLE 24

FREQUENCY DISTRIBUTION OF RESPONSES TO PERCEIVED BICYCLE SAFETY

| Perceived safety category (on year time reference) | Per cent of responses |
|---|---|
| 1. Almost no chance of having any sort of accident | 10.3 |
| 2. Some chance of having only a minor accident | 17.2 |
| 3. Large chance of having only a minor accident | 20.7 |
| 4. Slight chance of having a serious accident involving personal injury | 24.1 |
| 5. Large chance of having a serious accident involving personal injury | 27.6 |

Note: sample size = 29.

TABLE 25

PERCEIVED BICYCLE SAFETY REGRESSION RESULTS

| Variable Name | Variable Definition | Parameter Estimate | T-Statistic |
|---|---|---|---|
| BKROADS | Time spent travelling on main roads (minutes) | 0.0825 | 2.68 |
| BKSTREETS | Time spent travelling on side streets (minutes) | 0.0038 | 0.11 |
| BKWAYS | Time spent travelling on bikeways (minutes) | -0.0707 | -0.97 |
| BKPARKS | Time spent travelling in parklands (minutes) | -0.0511 | -0.96 |
| DENSEZN | Dummy variable with value 1 if workplace is located in CBD and zero otherwise | 0.1554 | 0.32 |
| CONSTANT | | 2.9135 | 5.64 |

Notes:

1. Dependent variable = perceived bicycle travel safety rating.

2. $R^2 = 0.34$.

The final piece of analysis concerns the influence of perceived bicycle safety and provision of bicycle facilities on the decision to bicycle to work. Following the model specification in Section 4.2.2, it is assumed that each individual associates with each available commuting mode a level of utility, this utility being a function of modal attributes. The modal utility functions, identified by equation (20), were specified in their non-random components as:

$$y^*_{q1} = a_1 + \delta_1(\text{INVTT})_{q1} + \delta_2(\text{OVTT})_{q1} + \delta_3(\text{TCOSTINC})_{q1}$$

$$y^*_{q2} = a_2 + \delta_1(\text{INVTT})_{q2} + \delta_2(\text{OVTT})_{q2} + \delta_3(\text{TCOSTINC})_{q2} + \delta_4(\text{DIVTIME})_{q2}$$

$$y^*_{q3} = a_3 + \delta_1(\text{INVTT})_{q3} + \delta_2(\text{OVTT})_{q3} + \delta_3(\text{TCOSTINC})_{q3}$$

$$y^*_{q4} = a_4 + \delta_2(\text{OVTT})_{q4} + \delta_5(\text{BKWAYS})_{q4} + \delta_6(\text{BKSAFETY})_{q4}$$

$$y^*_{q5} = a_5 + \delta_2(\text{OVTT})_{q5} + \delta_7(\text{WKPARKS})_{q5} \qquad (46)$$

where the subscripts 1, . . . , 5 refer to the modes car driver, car passenger/car pool, bus, bicycle and walk, respectively, with the variable definitions shown in Table 26.

Estimation results from a multinomial logit (MNL) mode choice model, defined by equations (23) and (46), using the JTW data, are shown in Table 26. These results hold few surprises. Travel time was categorised into in-vehicle and out-of-vehicle times. The former refers to time spent travelling in a car or a bus. The latter refers to wait time and time spent walking to and from a car or public transport, time spent bicycling, or travel time for the walk mode. From Table 26 the parameter estimate for OVTT is about twice that of IVTT suggesting that commuters in choosing a mode, tend to negatively weight out-of-vehicle time about double in-vehicle time. The parameter estimates attached to INVTT, OVTT

TABLE 26

MNL MODEL OF JOURNEY TO WORK MODE CHOICE

| Variable Name | Variable Definition | Parameter Estimate | T-Statistic |
|---|---|---|---|
| INVTT | In-vehicle travel time | -0.0531 | -1.92 |
| OVTT | Out-of-vehicle travel time | -0.1185 | -3.68 |
| TCOSTINC | Travel costs divided by income | -0.0838 | -2.36 |
| DIVTIME | Driver diversion time to pick up and drop car passenger, 0 for non-car passenger modes | -0.1736 | -2.22 |
| BKWAYS | Time spent travelling on bikeways, 0 for non bike modes | 0.5132 | 2.17 |
| BKSAFETY | Perceived bicycle safety rating, 0 for non-bike modes | -0.6620 | -1.68 |
| WKPARKS | Per cent of travel time spent walking through parklands, 0 for non-walk modes | 7.226 | 2.08 |
| CARDRIVER | Constant equal to 1 if car driver, 0 otherwise | -0.1986 | -0.25 |
| CARPASS | Constant equal to 1 if car passenger, 0 otherwise | -2.119 | -2.48 |
| CARPOOL | Constant equal to 1 if car pool, 0 otherwise | 1.625 | 1.27 |
| BUS | Constant equal to 1 if bus, otherwise | -0.4701 | -0.64 |
| BIKE | Constant equal to 1 if bicycle, otherwise | -0.337 | 0.02 |

Notes: $\rho^2$ = 0.52. per cent of choices correctly predicted 83% (without model 45%).

and the travel cost variable TCOSTINC imply that out-of-vehicle time is valued at 141% and in-vehicle time at 63% of the wage rate.

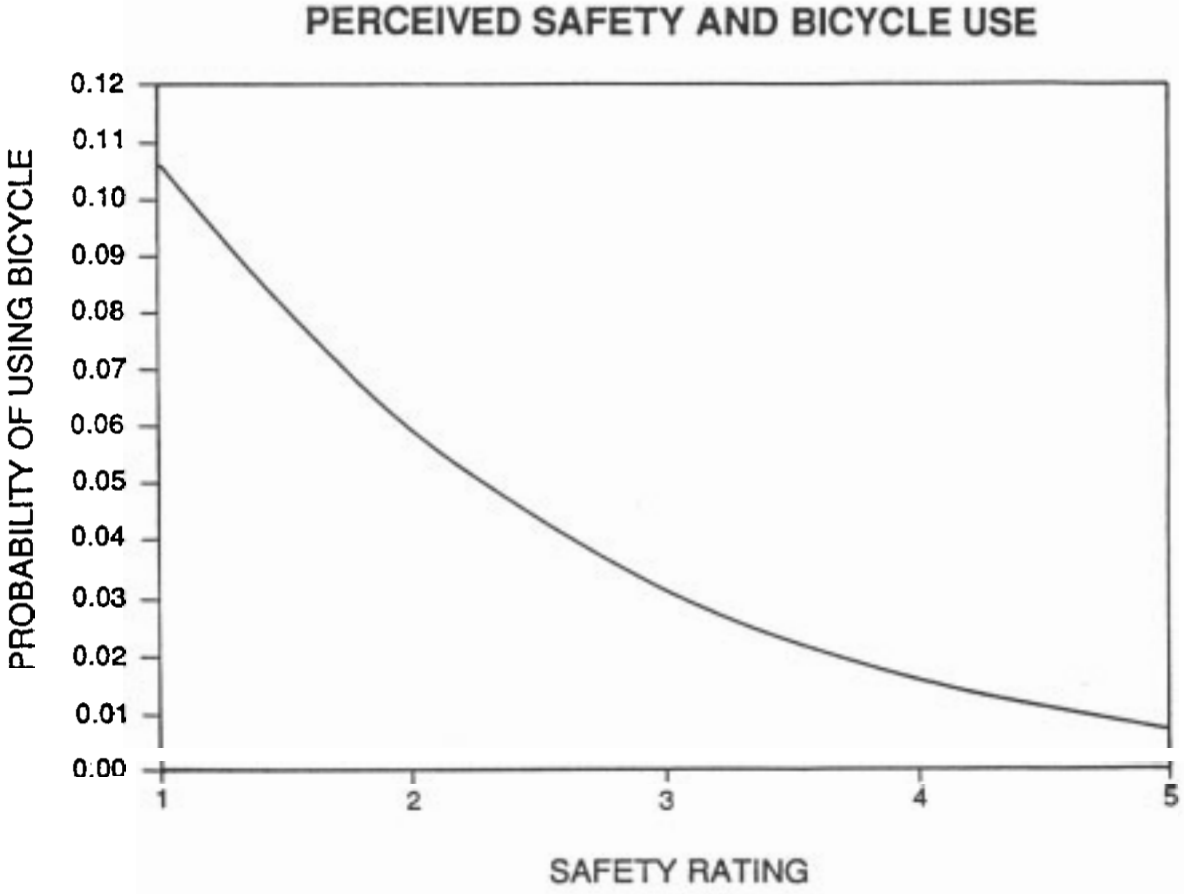
Of pivotal importance to bicycle planning is the parameter values associated with the variables BKWAYS and BKSAFETY. The positive parameter estimate for BKWAYS indicates that the provision of bikeways increases the probability of choosing bike as the commuting mode. From the regression analysis, however, the intrinsic attractiveness of bikeways appears not to be related to better safety. As anticipated, the perceived safety rating has a negative impact on the probability of bicycling to work.

Having estimated utility functions for each mode, predictions can be made of the effect on utility and hence mode choice of changes in the travel environment. For example, there may be a desire to analyse the effect of altering perceptions of bicycle safety for those who currently believe that there is a large chance of being involved in a serious personal injury crash if bicycle was used as the commuting mode for one year (category 5 in the scale) to believing that there would be almost no chance or only a slight chance of having a minor crash (categories 1 and 2 in the scale). Such changes in perceptions may be brought about through a public relations campaign or through physical changes to the travel environment that result in real improvements in bicycle safety. Using the estimated utility functions and equation (23), Figure 6 graphs the predicted probability of choosing bicycle against the perceived safety rating. By comparing the predicted probabilities of using a bicycle with a safety rating of 5 (approximately, 0.01) with the predicted probabilities of using a bicycle given a safety rating of 1 or 2 (approximately, 0.11 and 0.06, respectively) and recalling that 28% of respondents gave bicycle a safety rating of 5, it can be seen that bicycle use would substantially increase if its perceived safety could be improved.

In summary research reported in this note has shown:

(i) that as a commuting option bicycles have achieved very low market penetration,

Figure 6


PERCEIVED SAFETY AND BICYCLE USE
PROBABILITY OF USING BICYCLE
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
1
2
3
4
5
SAFETY RATING

(ii) that bicycles lack safety credibility as a commuting mode and this impacts on their use,

(iii) that perceived safety is related to conditions of bicycle use, particularly the amount of time spent on main roads, and

(iv) the provision of bikeways increases the amount of bicycle commuting.

These conclusions apply only to the survey area and must be tempered by the small sample size.

## 6.2. THE EFFECT OF SEAT BELT WEARING ON ROAD CRASH INJURIES.

In Section 4.2.3 a theoretical model was developed of the probability of being injured at various levels of severity given that a road crash had occurred. The model type that resulted from the theory presented was ordered logit. Recently this model has been empirically implemented (Barnard 1989) using injury data on vehicle drivers from the Adelaide In-Depth Accident Study (Road Accident Research Unit 1979).

The essential feature of the model developed in Section 4.2.3 is of a continuous variable, representing the severity of the crash, crossing a number of threshold levels, determined by the ability of the vehicle occupant to withstand collision forces. Crash severity and occupant injury threshold levels jointly determine the level of injuries sustained. Furthermore, just as the utility associated with a commuting mode could be specified as a function of modal attributes, so can crash severity and occupant threshold levels be specified as functions of other observable variables. In the empirical implementation of the model crash severity was specified as a function of change in vehicle velocity on impact, vehicle mass and seat belt use and tollerance capacity a function of age, sex and state of inebriation. These variables were selected on the basis of an understanding of the physics of road crashes and results obtained from laboratory based

biomechanical research.

6.2.1. Explanatory Variables of Injury Severity.

Biomechanical studies of automobile collisions have shown that the single most important variable contributing to crash severity is the change in vehicle velocity on impact. As it relates directly to the change in momentum, this variable best reflects the forces acting on vehicle occupants' bodies during a collision. During the collision period the occupants' initial velocities must be changed to the new velocity of the vehicle compartment, resulting in occupant contact with the vehicle interior and/or with a restraint system. The occurrence and nature of the injuries sustained will largely depend upon the de-acceleration time history of the collision (Krishnan et al. 1983).

Following Marquardt (1974), Carlson (1979), Hutchinson (1983) and Krishnan et al. (1983), from momentum considerations the change in vehicle velocity on impact can be calculated as:

$$\Delta v_c = \frac{(v_c^2 + v_o^2 + 2v_c v_o \cos(\theta))^{0.5}}{1 + m_c/m_o} \qquad (47)$$

where $v_c$ is the initial velocity of the case vehicle, $v_o$ is the velocity of the other vehicle, $m_c$ and $m_o$ are the masses of the vehicles and $\theta$ is the angle of alignment of the vehicles at the point of impact.[1] The change in velocity on impact increases as the initial velocities of the vehicles increase and as the ratio $m_c/m_o$ increases.

To aid appreciation of the index, hypothetical crash

[1]The formula of equation (47) takes into consideration the change in velocity along both the x and y axes. Some of the studies mentioned only included the velocity change in one direction.

configurations are depicted in Figures 7 and 8. Both figures involve a two-vehicle collision between a Honda Civic of mass 690 kg and a Ford Fairmont of mass 1333 kg. The vehicles are assumed to be both travelling at 30 km/h immediately prior to impact. In the head-on crash $\theta = 0$, so that the relative collision velocity is 60 km/h and the change in velocity for the Honda Civic is -40 km/h. If both cars were travelling at 40 km/h the change in velocity of the Honda on impact would be -53 km/h. The corresponding changes in velocity for the Ford Fairmont are -20 km/h and -27 km/h. For the crash depicted in the second diagram the change in velocity on impact for the vehicles is exactly half that of the Figure 7 crash. The reduction in impact velocity change is entirely due to the different alignment of the vehicles, measured by the $\cos(\theta)$ term in equation (47).

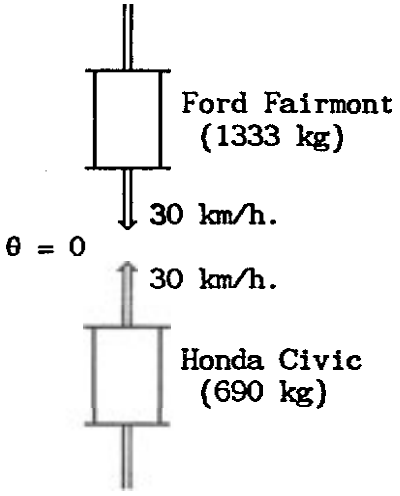
A feature of equation (47) is that the change in vehicle velocity on impact is contingent upon the *ratio* of vehicle masses. This means, given initial velocities and vehicle alignment, the velocity change for a collision involving two equivalently weighted light vehicles will be the same as for two equivalently weighted heavy vehicles. An understanding of collision physics, however, leads to the conclusion that the severity of the crash will be greater for light vehicle occupants than for massive vehicle occupants. Larger vehicles provide a protective effect to occupants. The larger the vehicle the greater is the proportion of initial vehicle energies that can be absorbed by metal deformation without intrusion into the occupant compartment. There is also more room in the vehicle cabin for the occupant to travel without striking an object.

In addition to the main effects of impact velocity change and vehicle mass, the severity of the crash to the occupant will be affected by seating position and restraint usage. The driver, in particular, is likely to be more vulnerable than other vehicle occupants because of the close proximity to the body of the steering apparatus. The role of seat belts in reducing injuries has been well documented (e.g. Lave and Webb 1970, Trinca 1980, Layton and Weigh 1983).

FIGURE 7

EXAMPLE OF IMPACT VELOCITY CHANGE CALCULATION:

HEAD-ON COLLISION

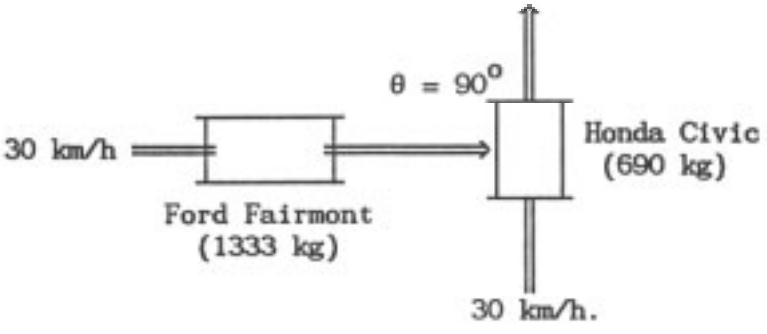



$$\Delta v_{Honda} = \frac{[30^2 + 30^2 + 2(30)(30)]}{1 + \frac{690}{1333}} = 40 \text{ km/h.}$$

FIGURE 8

EXAMPLE OF IMPACT VELOCITY CHANGE CALCULATION:

SIDE-ON COLLISION




$$\Delta v_{Honda} = \frac{[30^2 + 30^2 + 2(30)(30)\cos(90)]}{1 + \frac{690}{1333}} = 28 \text{ km/h}$$

There is less evidence on the factors contributing to the different capacities of individuals to tolerate collision forces. Age, sex and state of intoxication are the variables used in the current analysis.

6.2.2. The Adelaide In-Depth Accident Study.

The data source for the research is the Adelaide in-depth accident study. This study, sponsored by the Office of Road Safety and the Australian Road Research Board, obtained an 8% sample of crashes in the Adelaide Metropolitan Area, to which an ambulance was called, during the period March 1976 - March 1977. The inclusion criterion means that the sample consists of a non-random subset of the total crash population.

Crash researchers invariably work with non-random samples. Typically the sample inclusion criterion is that damages should exceed a specified monetary amount (currently $1000 in South Australia). In practice many crashes with monetary losses well in excess of the specified monetary amount go unreported. It is likely that the Adelaide in-depth sample inclusion criterion is not very different to normal reporting criterion. Ambulances tend to be called routinely to crashes of even moderate severity. In the sample used for modelling an ambulance was called, but was not required, in over 50% of crashes.

The Adelaide in-depth sample was found to be representative of the population of crashes, characterised by mass crash data records, for a number of key variables (Road Accident Research Unit 1979). In total the study collected information on 494 vehicle crash involvements.

The available sample size from this source was small compared to that available from mass crash data tapes. It is, however, of a higher quality, with the data on each crash reflecting the combined on-site talents of an engineer, psychologist and a medical officer. In selecting this data source we were particularly

mindful of having accurate information on the velocities of vehicles just prior to impact and vehicle alignment when calculating the velocity change index of crash severity. This information is not included on Australian mass crash data records. We were also aware that the statistical techniques adopted in this study may be successfully applied, and stable parameter estimates obtained, with as few as 70 observations (Tye et al. 1982, p. 27)

No automobile mass information was collected in the Adelaide data; however, information on vehicle make and model was obtained. The information needed to calculate impact velocity change was derived by linking the Adelaide data with a data set on automobile characteristics, collected under the Dimensions of Automobile Demand Project sponsored by the National Energy Research Development and Demonstration Program (NERRDP) (Hensher 1986).

The current research is confined to an examination of two-vehicle collisions. In the Adelaide data 303 vehicles were involved in collisions of this type. Occupants of these vehicles in total numbered 561 persons. Automobile-only two-vehicle collisions involved 196 automobiles, occupied by 394 persons. The distribution of AIS injuries sustained by vehicle occupants in two-vehicle collisions and in automobile-only two-vehicle collisions are displayed in Table 27. Unfortunately for this analysis, no injuries were observed in AIS classes 5 or 6 (notably there were no fatalities). Other relevant summary statistics for the data are shown in Table 28.

### 5.2.3. Injury Severity Model Estimation Results.

Four models were developed using the Adelaide/NERRDP data. These related to (i) two-vehicle collisions, all occupants, (ii) two-vehicle collisions, drivers only, (iii) automobile-only two-vehicle collisions, all occupants, and (iv) automobile-only two-vehicle collisions, drivers only. The form for $y_q^* - t_q^*$ used in these models was:

TABLE 27

DISTRIBUTION OF AIS CLASSIFIED INJURIES

(a) Two-vehicle collisions.

| AIS Class | Proportion of Sample |
|---|---|
| 0 | 0.46 |
| 1 | 0.34 |
| 2 | 0.14 |
| 3 | 0.05 |
| 4 | 0.01 |

(b) Automobile-only two-vehicle collisions.

| AIS Class | Proportion of Sample |
|---|---|
| 0 | 0.44 |
| 1 | 0.39 |
| 2 | 0.14 |
| 3 | 0.03 |
| 4 | 0.01 |

TABLE 28

SUMMARY SAMPLE STATISTICS

| Variable Description | Sample Statistic |
|---|---|
| **(a) Two-vehicle collisions** | |
| average number of vehicle occupants | 1.85 persons |
| average speed on impact | 36 km/h |
| average collision impact velocity | 53 km/h |
| average velocity change on impact | 27 km/h |
| average vehicle mass | 1160 kg |
| percent of head-on collisions | 6% |
| percent of side-on collisions | 81% |
| percent of rear-end collisions | 13% |
| percent of occupants wearing a seat belt | 64% |
| percent of occupants aged more than 60 years | 7% |
| percent of female occupants | 42% |
| percent of drivers intoxicated | 16% |
| **(b) Automobile-only two-vehicle collisions** | |
| average number of vehicle occupants | 2.01 persons |
| average speed on impact | 36 km/h |
| average collision impact velocity | 55 km/h |
| average velocity change on impact | 28 km/h |
| average vehicle mass | 1134 kg |
| percent of head-on collisions | 5% |
| percent of side-on collisions | 88% |
| percent of rear-end collisions | 7% |
| percent of occupants wearing a seat belt | 66% |
| percent of occupants aged more than 60 years | 8% |
| percent of female occupants | 48% |
| percent of drivers intoxicated | 17% |

$$y_q^* - t_q^* = \delta_0 + \beta_0 + \delta_1 \text{VELCHNG} + \delta_2 \log(\text{MASS}) + \delta_3 \text{SBELT}$$

$$+ \delta_4 \text{DRIVER} + \delta_5 \text{INTOX} + \beta_1 \text{AGE60} + \beta_2 \text{FEMALE}, \qquad (48)$$

where VELCHNG is the velocity change on impact (= $\Delta v$), MASS is the mass of the case vehicle (= $m_c$), 'log' denotes the natural logarithm, SBELT is a binary variable taking value 1 if in the investigator's judgement a seat belt was certainly or probably worn and 0 otherwise, DRIVER is a binary variable if the occupant was seated in the driving position and 0 otherwise, INTOX is a binary variable if the occupant was slightly, moderately or severely intoxicated and 0 otherwise, AGE60 is a binary variable taking the value 1 if the occupant was aged more than 60 years and 0 otherwise, and FEMALE is a binary variable taking the value 1 if the occupant was a female and 0 if the occupant was a male.

A number of features of the model implied by equations (10) and (48) should be noted. First, the parameters $\delta_0$ and $\beta_0$ cannot separately be estimated. The model estimates a single constant term equal to $\delta_0 + \beta_0$. A consequence of this is that the separate indices for $y_q^*$ and $t_q^*$ cannot be recovered, only the combined index, $y_q^* - t_q^*$. This inability is not a cause for concern, however, because we are only interested in knowing how the independent variables VELCHNG, MASS, SBELT, DRIVER, INTOX, AGE60 and FEMALE affect the probability of sustaining an injury at AIS level i, i = 0, 1, 2, 3, 4. Second, for the drivers-only models, DRIVER is a constant so that $\delta_4$ cannot be estimated. The parameter $\delta_4$ is absorbed into the constant term and other parameter estimates remain unbiased. Third, intoxication data were only collected for drivers. This term was therefore omitted in the 'all occupants' models.

Estimation results for the four models are shown in Tables 29 - 32. All included variables, except the absolute mass and seat belt variables in the 'automobile-only driver-injuries' model, are statistically significant at the 2.5% level using a one-tailed T-test, and even these variables are significant at the 5% level

TABLE 29

ORDERED LOGIT MODEL OF OCCUPANT INJURY SEVERITY
IN TWO-VEHICLE CRASHES

| Variable Mnemonic | Parameter Estimate | Standard Error | T-Statistic |
|---|---|---|---|
| CONSTANT | 3.27464 | 1.2670 | 2.59 |
| VELCHNG | 0.06411 | 0.0071 | 8.97 |
| log(MASS) | -0.75606 | 0.1652 | -4.58 |
| DRIVER | 0.46227 | 0.2134 | 2.17 |
| SBELT | -0.30994 | 0.0956 | -3.24 |
| AGE60 | 0.95468 | 0.2893 | 3.30 |
| FEMALE | 0.78249 | 0.2041 | 3.83 |
| $k_1$ | 2.05903 | 0.1346 | 15.30 |
| $k_2$ | 3.74015 | 0.2411 | 15.51 |
| $k_3$ | 6.40987 | 0.6249 | 10.26 |

| | |
|---|---|
| Number of observations | 561 |
| Log Likelihood at $\bar{\delta} = 0$ | -660.98 |
| Log Likelihood at convergence | -557.48 |
| $R^2$ | 0.36 |

TABLE 30

ORDERED LOGIT MODEL OF OCCUPANT INJURY SEVERITY
IN AUTOMOBILE-ONLY TWO-VEHICLE CRASHES

| Variable Mnemonic | Parameter Estimate | Standard Error | T-Statistic |
|---|---|---|---|
| | 4.04485 | 3.2690 | 1.24 |
| | 0.05813 | 0.0103 | 5.65 |
| | -0.86774 | 0.4418 | -1.96 |
| | 0.55297 | 0.2437 | 2.27 |
| | -0.28259 | 0.1075 | -2.63 |
| | 1.16618 | 0.3363 | 3.47 |
| | 1.15146 | 0.2423 | 4.75 |
| | 2.10255 | 0.1543 | 13.63 |
| | 3.92270 | 0.3037 | 12.92 |
| | 5.65502 | 0.6583 | 8.59 |

| | |
|---|---|
| Number of observations | 394 |
| Log Likelihood at $\bar{\delta} = 0$ | -651.01 |
| Log Likelihood at convergence | -408.84 |
| $R^2$ | 0.24 |

TABLE 31

ORDERED LOGIT MODEL OF DRIVER INJURY SEVERITY
IN TWO-VEHICLE CRASHES

| Variable Mnemonic | Parameter Estimate | Standard Error | T-Statistic |
|---|---|---|---|
| CONSTANT | 3.46723 | 1.5760 | 2.20 |
| VELCHNG | 0.06221 | 0.0098 | 6.34 |
| log(MASS) | -0.72935 | 0.2127 | -3.43 |
| SBELT | -0.31620 | 0.1252 | -2.53 |
| INTOX | 1.10009 | 0.3335 | 3.30 |
| AGE60 | 0.90783 | 0.3694 | 2.46 |
| FEMALE | 0.61840 | 0.2883 | 2.15 |
| $k_1$ | 2.13719 | 0.1927 | 11.09 |
| $k_2$ | 3.75491 | 0.3104 | 12.10 |
| $k_3$ | 6.21678 | 0.6771 | 9.18 |

| | |
|---|---|
| Number of observations | 303 |
| Log Likelihood at $\vec{\delta} = 0$ | -373.96 |
| Log Likelihood at convergence | -302.71 |
| $R^2$ | 0.42 |

TABLE 32

ORDERED LOGIT MODEL OF DRIVER INJURY SEVERITY
IN AUTOMOBILE-ONLY TWO-VEHICLE CRASHES

| Variable Mnemonic | Parameter Estimate | Standard Error | T-Statistic |
|---|---|---|---|
| CONSTANT | 6.07375 | 4.7140 | 1.29 |
| VELCHNG | 0.06644 | 0.0152 | 4.37 |
| log(MASS) | -1.14107 | 0.6420 | -1.78 |
| SBELT | -0.27665 | 0.1519 | -1.82 |
| INTOX | 1.12041 | 0.4069 | 2.75 |
| AGE60 | 0.96767 | 0.4426 | 2.19 |
| FEMALE | 0.97779 | 0.3525 | 2.77 |
| $k_1$ | 2.10935 | 0.2255 | 9.35 |
| $k_2$ | 3.61398 | 0.3714 | 9.73 |
| $k_3$ | 5.30452 | 0.7399 | 7.17 |

| | |
|---|---|
| Number of observations | 196 |
| Log Likelihood at $\bar{\delta} = 0$ | -234.31 |
| Log Likelihood at convergence | -203.79 |
| $R^2$ | 0.31 |

using this test. Parameter estimates attached to the two dominant indices of crash severity, VELCHNG and MASS, are of the anticipated sign. The sign of the parameter estimate attached to VELCHNG is positive signifying that, ceteris paribus, an increase in the initial speeds of the vehicles, *or* a decrease in the mass of the case vehicle *relative* to the other vehicle involved in the collision, will increase the likelihood that occupants of the case vehicle will be injured. Vehicle mass exerts a further influence on the probability of being injured through the MASS variable. This variable is measuring the degree of protection offered by travelling in a vehicle of larger *absolute* mass. The negative sign on this variable indicates that the probability of injury decreases as the absolute mass of the case vehicle increases.

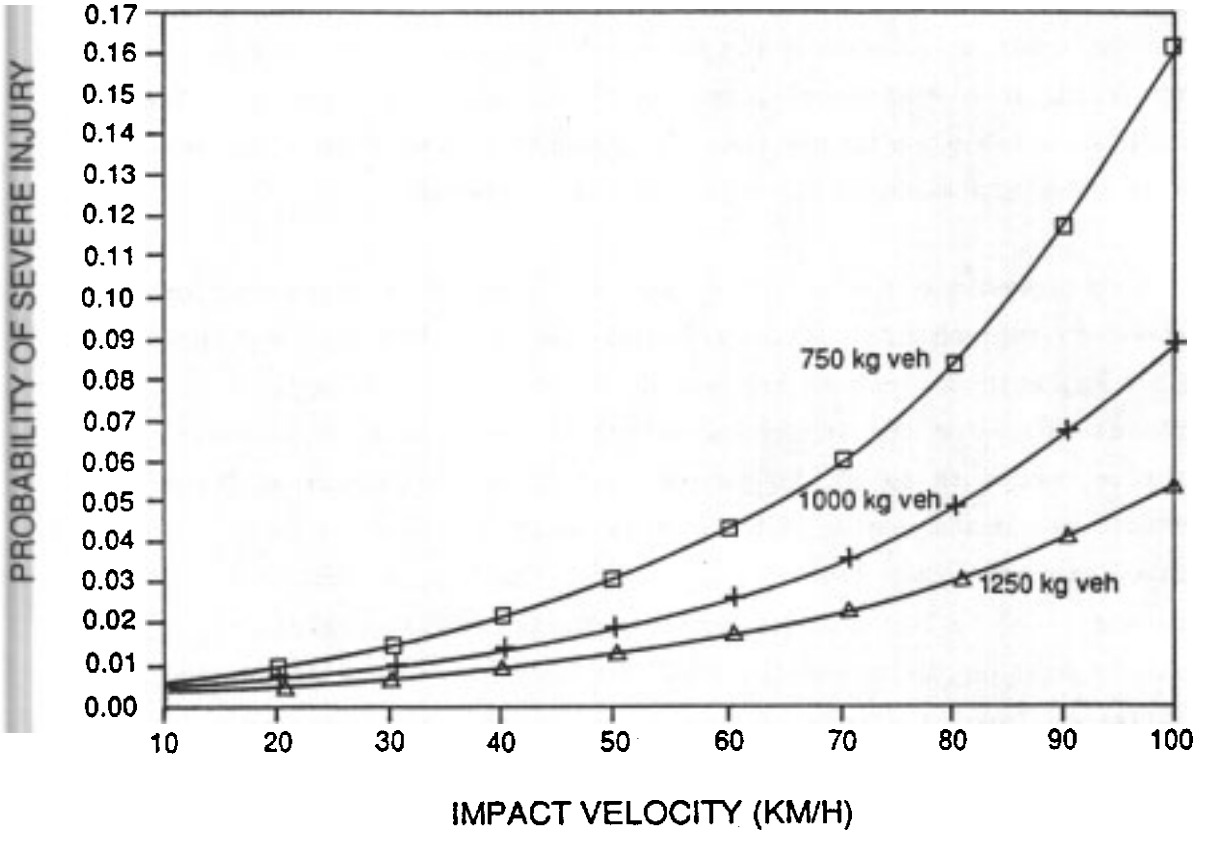
To appreciate the effect of vehicle size on injuries consider a head-on two-vehicle collision between a small (650 kg) and large (1300 kg) automobile both travelling at 30 km/h. The model predicts that for the injurious impact of the collision on small vehicle occupants to be the same as the injurious impact on large vehicle occupants one of the vehicles would have had to be travelling more than four times as fast; that is, at 120 km/h instead of 30 km/h. A diagrammatic depiction of the effect of vehicle mass on the probability of sustaining a severe injury in collisions involving various impact velocities with a second vehicle of mass 1000 kg is shown in Figure 9.[12]

Models were also estimated with impact velocity, the numerator of equation (47), included separately from the ratio of vehicle masses, the denominator of equation (47), and the logarithm of absolute vehicle mass. The separate terms were all statistically

[12]The variable levels assumed in Figures 9 - 12, excepting those explicitly set in each Figure, are $m_o$ = 1000 kg, $m_c$ = 1000 kg, SBELT = 1 (seat belt worn), DRIVER = 0 (occupant not seated in driver's position), INTOX = 0 (occupant sober), AGE60 = 0 (occupant younger than 60 years of age) and FEMALE = 0 (occupant is a male).

Figure 9

## INJURY SEVERITY BY VEHICLE WEIGHT



significant at the 2.5% level. However, the overall goodness of fit measures for these models and the models of Tables 29 - 32 were virtually identical, lending support to the use of the VELCHNG variable, derived from a theoretical consideration of the physics of two-vehicle collisions, as an index of crash severity. Models which omitted either of the mass effects or the impact velocity term were significantly inferior to the model in which all these terms were included. The correlation between log(MASS) and VELCHNG was always less than 0.35.

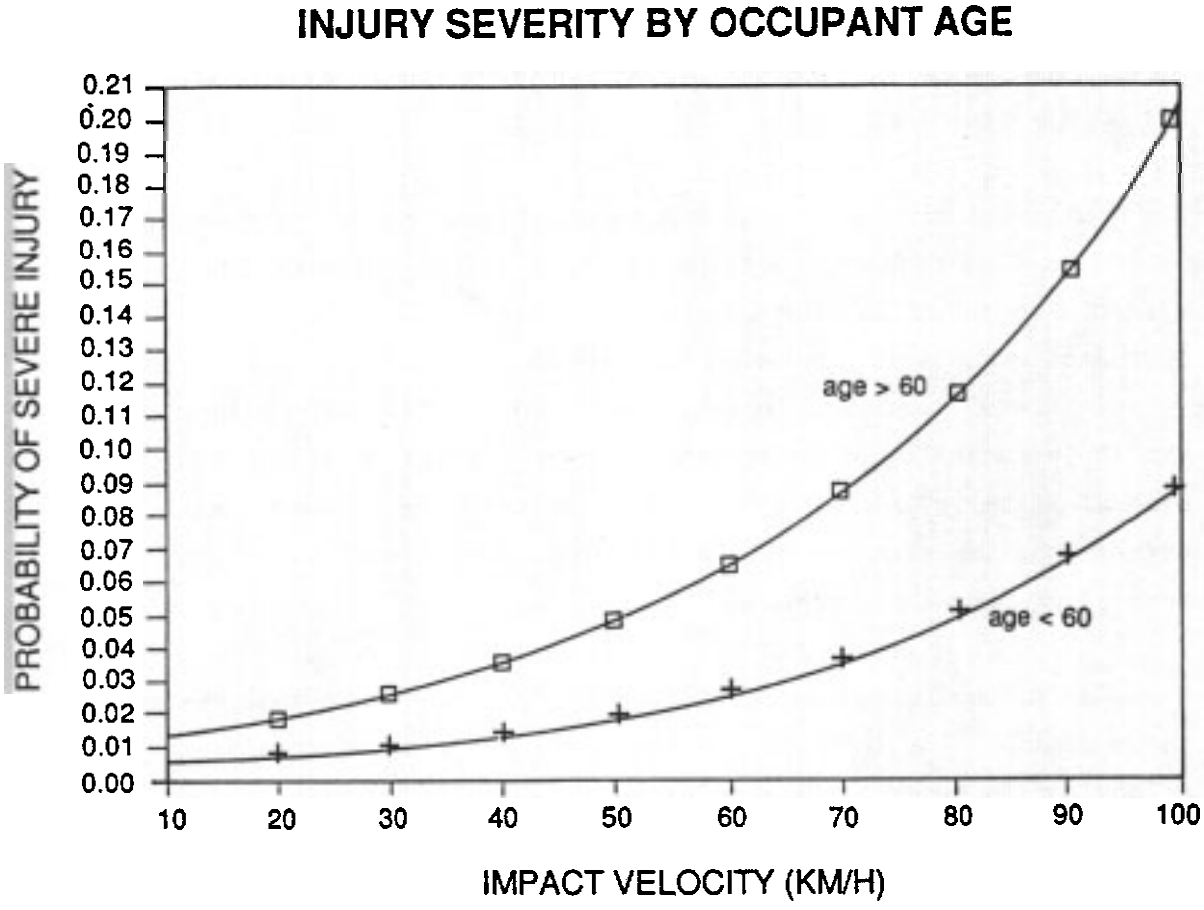
Parameter estimates attached to the remaining set of (binary) variables also took the anticipated signs. The positive sign attached to the AGE60 variable signifies that the elderly have an increased probability of being injured in a collision, of a given severity level, between two vehicles. The relationship between age and the probability of sustaining a severe injury is quantified diagrammatically in Figure 10. It is interesting to note that, ceteris paribus, females have a higher probability of being injured in road crashes than do males.

Again, having estimated the model, 'what if' type analyses can be conducted. For instance, it may be of interest to predict the effect on the severity of road crash injuries of an aging in the population or a change in the distribution of vehicle weights such as occurred as a result of the energy crisis of the early 1970s. A topic that has attracted considerable interest is the effect of restraint use in reducing road crash injuries. From the model of Table 31 we can calculate the effect of seat belts in reducing severe injuries for drivers involved in two-vehicle crashes, by the following steps:

(i) For individual q in our sample calculate the value of $\bar{y}^*_q - \bar{t}^*_q$ on the assumption a seat belt was not worn, as:

$$\bar{y}^*_q - \bar{t}^*_q = 3.467 + 0.062(\text{VELCHNG}_q) - 0.729(\log(\text{MASS}_q)) + 0.908(\text{AGE60}_q) + 0.618(\text{FEMALE}_q) + 1.100(\text{INTOX}_q)$$

Figure 10


INJURY SEVERITY BY OCCUPANT AGE
PROBABILITY OF SEVERE INJURY
0.21
0.20
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
age > 60
age < 60
10
20
30
40
50
60
70
80
90
100
IMPACT VELOCITY (KM/H)

Note that the values for all variables, but the seat belt variable, are those actually pertaining to individual q. SBELT is set to zero irrespective of whether or not a seat belt was actually worn. For example if individual q was a sober male aged less than 60 years driving a car of mass 1000kg and involved in a collision where the change in velocity was 35 km/h, assuming a seat belt was not worn $\bar{y}^*_q - \bar{t}^*_q$ may be calculated as 0.601.

(ii) Calculate the probability of individual q being severely injured on the assumption a seat belt was not worn as:

$$\text{Prob}(y_q = 3) + \text{Prob}(y_q = 4) = 1 - \frac{\exp(3.740 - \bar{y}^* + \bar{t}^*)}{1 + \exp(3.740 - \bar{y}^*_q + \bar{t}^*_q)},$$

with 3.740 being the threshold constant that separates injury severity classes 2 and 3. For the example individual:

$$\text{Prob}(y_q = 3,4) = 1 - \frac{\exp(3.740 - 0.601)}{1 + \exp(3.740 - 0.601)} = 0.04.$$

(iii) Repeat steps (i) and (ii) for all individuals in the sample and sum the probabilities to obtain the expected number of severely injured crash victums on the assumption of no seat belt wearing.

(iv) For individual q in our sample calculate the value of $\bar{y}^*_q - \bar{t}^*_q$ on the assumption a seat belt was worn, as:

$$\bar{y}^*_q - \bar{t}^*_q = 3.467 + 0.062(\text{VELCHNG}_q) - 0.729(\log(\text{MASS}_q)) + 0.908(\text{AGE60}_q) + 0.618(\text{FEMALE}_q) + 1.100(\text{INTOX}_q) - 0.316.$$

Again note that the values for all variables, but the seat belt

variable, are those actually pertaining to individual q. SBELT is set to one irrespective of whether or not a seat belt was actually worn. For the example individual, assuming a seat belt was worn, $\vec{y}^{*}_{q} - \vec{t}^{*}_{q} = 0.285$.

(v) Calculate the probability of individual q being severely injured on the assumption a seat belt was worn, following step (ii). For the example individual the new probability value is:

$$\text{Prob}(y_q = 3,4) = 1 - \frac{\exp(3.740 - 0.285)}{1 + \exp(3.740 - 0.285)} = 0.03.$$

(vi) Repeat steps (iv) and (v) for all individuals in the sample and sum the probabilities to obtain the expected number of severely injured crash victums on the assumption of universal seat belt wearing.

(vii) Compare the figures derived in steps (iii) and (vi) to gauge the impact of seat belt wearing in reducing severe car injuries.

Provided the model has been correctly estimated, the value obtained in step (iii) should correspond to the number of severe injuries that would actually be observed if no one wore seat belts. Likewise the value obtained in step (vi) should correspond the number of severe injuries that would actually be observed if everyone wore seat belts. This use of the model, therefore, provides an excellent measure of seat belt effectiveness in reducing injuries. Applying this measure to the Adelaide data set it was estimated that the wearing of seat belts should lead to a 33% reduction in the number of severe injuries and a 12% reduction in the number of minor injuries.

The estimates obtained from this study on the effect of seat belts in reducing severe and minor injuries are considerably lower than the estimates of 60% and 30%, respectively, quoted by the United States National Highway Traffic Safety Administration (cited

in Arnould and Grabowski 1983). The latter of my estimates, however, is in reasonable accord with an estimated 44% reduction in the number of fatalities from seat belt wearing derived from the work of Layton and Weigh (1983), based on Queensland data, and an estimated reduction of 30% in severe and fatal crashes by Krishnan (1983) based on two U.S. data sets.[13] In making these comparisons it is important to note that the current research was restricted to an analysis of urban road crashes.

A pictorial representation of the effectiveness of seat belts in reducing the probability of severe injuries for a sober male vehicle occupant, younger than 60 years of age, at various impact velocities is shown in Figure 11.

Finally, as shown in Figure 12, inebriated drivers were more likely to be injured than non-inebriated drivers. Other studies (e.g. Raymond 1974, Johnson 1978) have concluded that those affected by alcohol are more likely to be involved in crashes in general, particularly, severe crashes. The current study adds to the store of knowledge on the relationship between alcohol and road safety by concluding that given a crash of a set severity level those under the influence of alcohol are about two and one-half times more likely to sustain a severe (AIS classes 3 and 4) injury than a sober occupant.[14]

---

[13]Layton and Webb calculate that fatalities were reduced by 37% as a result of legislation making seat belt wearing compulsory, if fitted, after allowing for the proportion of vehicles in the fleet without seat belts fitted. We have adjusted this estimate upwards to also allow for the proportion of the travelling population not using a seat belt when available (see Lay 1984).

[14]Dr. Max Lay has drawn to my attention research by Waller et al. (1986) which reached a similar conclusion about the effect of alcohol on injury severity. Waller and his colleagues concluded that alcohol involved drivers were 3.85 times more likely to be killed than sober drivers, once differences in vehicle deformation

Figure 11

## INJURY SEVERITY BY SEAT BELT USAGE

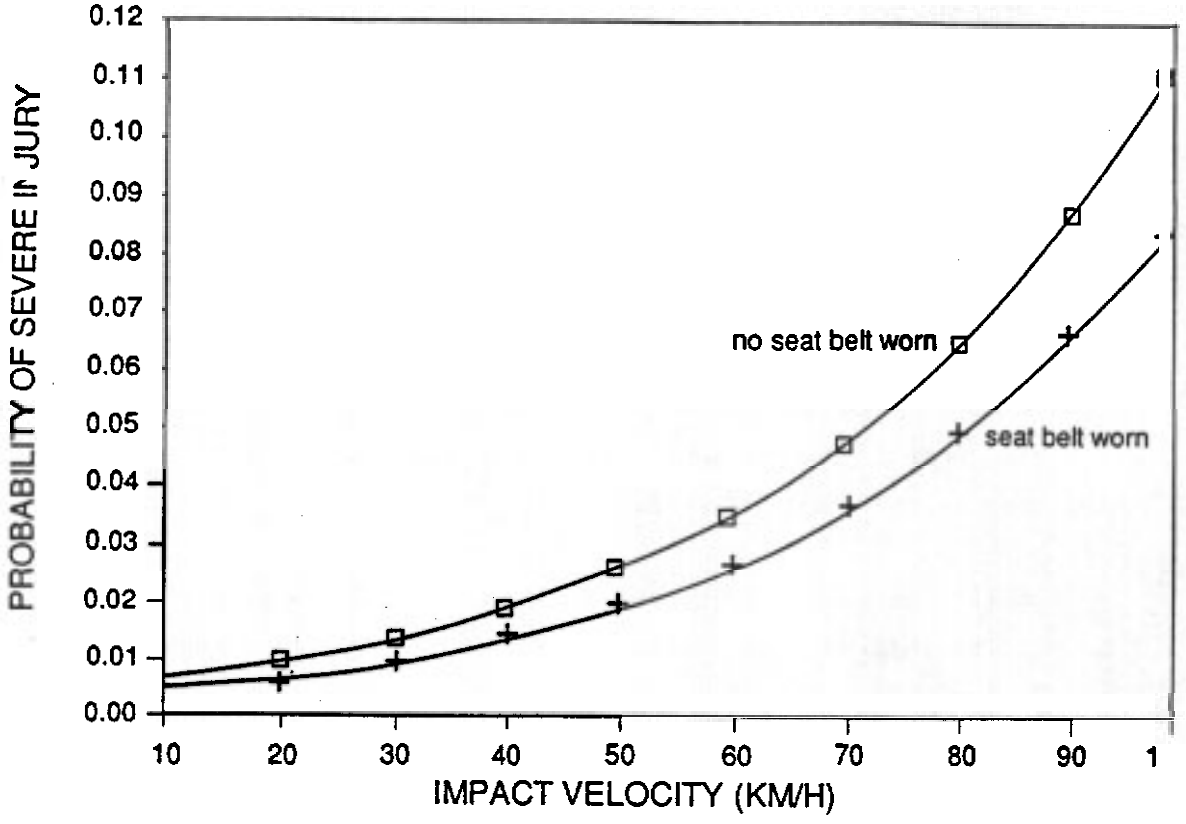
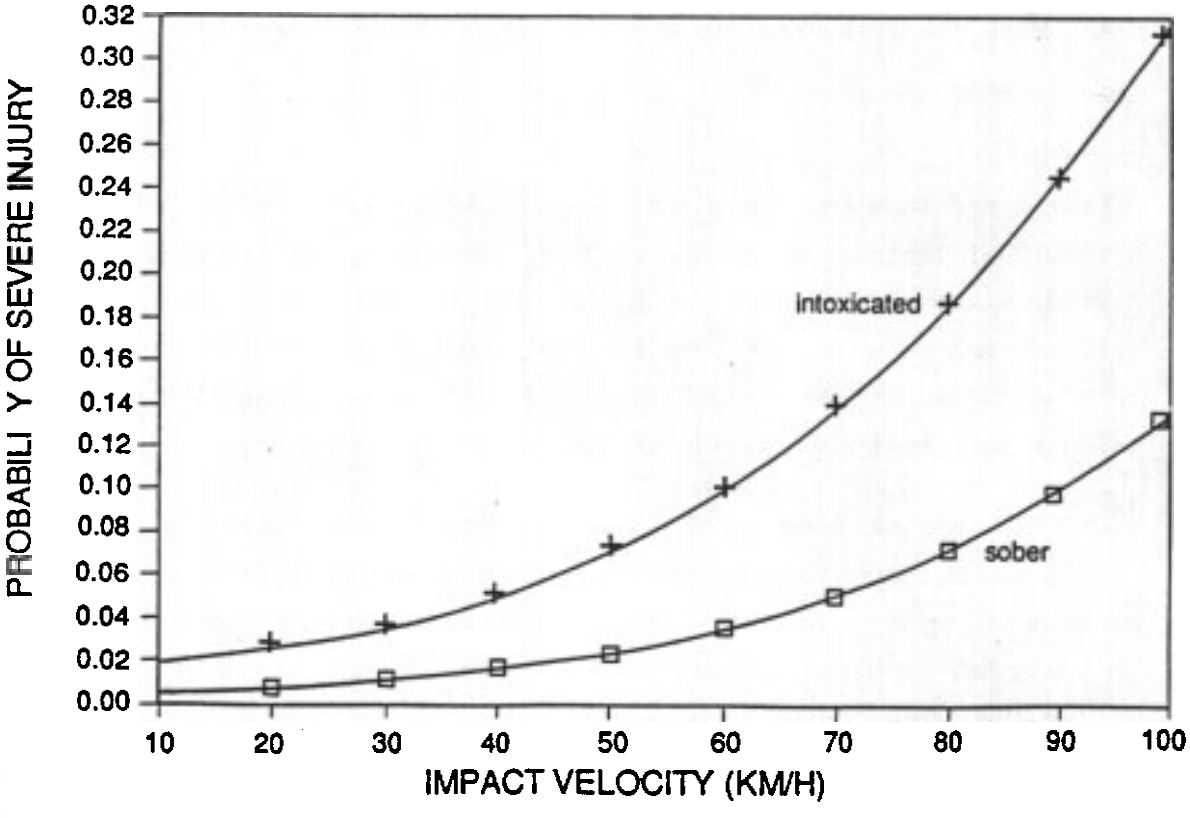



Figure 12

## INJURY SEVERITY BY STATE OF INTOXICATION



The two main measures used to assess the overall goodness of fit of the ordered logit model are the value of the log-likelihood function at convergence and $R^2$. The latter of these is analogous to the $R^2$ of regression analysis and is given by:

$$R^2 = \frac{\sum_q (\hat{\tau}_q - \bar{\tau})}{\left[\sum_q (\hat{\tau}_q - \bar{\tau})\right] + \frac{\pi^2}{3} Q} \quad (49)$$

where $\hat{\tau}_q = \hat{y}^*_q - \hat{t}^*_q$, $\bar{\tau} = \sum_q^Q \hat{\tau}_q / Q$, and Q is the sample size.[15] In assessing the fit of these models it must be emphasized that indeterminacy pervades the road crash injury process. Even in a low severity crash a fragment of flying glass can cause severe

---

and crash type had been taken into account. Their estimate of alcohol induced injury severity is slightly higher than mine. Their study, however, did not control for seat belt usage differences between sober and inebriated drivers, sex differences and age differences. These additional factors are taken into account in the current study.

[15]It will be noted that the $R^2$ of equation (49) and the $R^2$ of regression analysis are equivalent except the sum of squared residuals and the total sum of squares are both estimates rather than actual values. The $R^2$ of equation (49) is therefore an estimate of the true $R^2$. To fully utilise this estimate knowledge of the sample distribution of the true $R^2$ is required, but the sample distribution of the true $R^2$ is presently unknown (McKelvey and Zavoina 1975).

injury, possibly death, if it becomes lodged in a vulnerable point of the human body. No model can hope to capture this level of detail. In explaining between 24% - 42% of the variation in road crash injuries, using only a limited set of variables, the models do surprisingly well.

## 6.3. A MODEL OF CRASH INVOLVEMENT.

Section 4.2.1 contained a discussion of one method to model crash involvement using a binary logit model. In this Section a more advanced method of modelling crash involvement is outlined. The method is known as the Cox proportional hazzards model. Although this method falls into the general family of methods covered in this report, being closely related to the models introduced in Section 5.3, it has not been specifically reviewed. It is highlighted here because of its particular suitability for predicting the type of trips that are likely to result in a crash. The discussion is largely based on a paper by Jovanis and Chang (1986).

The starting point for an appreciation of the suitability of the Cox proportional hazzards model for analysing crash trip data is a recognition that a trip occupies a certain amount of time. Ceteris paribus, the longer the trip the more likely a crash will occur. The likelihood of a crash is not only related to exposure (trip duration) but also conditions faced on the trip. These conditions refer to characteristics of the environment, vehicle, road, and driver. For instance, a larger risk factor would be associated with a trip involving crossing three single-lane bridges than a trip involving no such crossings.

Suppose each trip was potentially infinite in length so that all trips were terminated by crashes. The probability of an individual trip of duration T surviving to time t is:

$$S(t) = \text{Prob}\{T \geq t\} = \int_t^{\infty} f(x)\, dx \qquad (50)$$

where S(t) is the survival function and f(t) is the probability density function of T. The probability of a crash occuring at time t given that no crash has occured prior to that time is given by the hazzard function, h(t):

$$h(t) = \frac{f(t)}{S(t)} \tag{51}$$

In the basic Cox model the hazzard rate at time t, conditional on a set of explanatory variables, X, is given by:

$$h(t|X) = h(0|X)\exp(X_q\beta) \tag{52}$$

where $h(0|X_q)$ is the baseline hazzard rate. The probability that a crash occurs at time t is:

$$\frac{\exp(X_q\beta)}{\sum_{q\epsilon R_i} \exp(X_q\beta)} \tag{53}$$

where $R_i$ is the set of trips of duration greater than or equal to t. For the simple case described where a crash inevitably occurs on each each trip and durations are distinct the parameters, $\beta$, can be estimated using equation (53) and maximum likelihood techniques. In reality, however, most trips are short and risk does not accumulate to the extent that a crash results. We may view the data as censored in that we only know that the risk exposure required for a crash involvement to be observed is greater than the duration of the trip. For this more general case parameters may again be estimated by maximum likelihood methods using a sample of crash and non-crash trips. Alternatively, even for this more general case, the $\beta$s may be consistently estimated, albeit inefficiently, using data only on trips involving a crash.

A Cox proportional hazzards model has been used by Jovanis and Chang to study truck crashes in the United States. A sample of more than 1200 truck trips involving collisions was used in

estimating the model. Data was collected on the age, experience, crash record and working hours just prior to the crash trip of the driver, type of truck, cargo weight, type of roadway, number of lanes, weather and lighting conditions, traffic volume, time of year and time of day. Preliminary estimates are contained in Jovanis and Chang (1986). Results are a little disappointing. Not surprisingly, different factors appeared to be responsible for different kinds of truck crashes. Nevertheless, an important finding was that regularly scheduled drivers who took frequent trips appeared to have reduced risk of a crash, particularly if they had a longer number of hours off duty just prior to the trip. Time of day and year factors were also significant determining factors for some types of crashes.

The Cox proportional hazzards model provides a more complete picture of the crash process than the binary latent variable model proposed in Section 4.2.1. The latter provides a 'snapshot' view of the crash whereas the former allows the accumulation of factors over the course of a trip. The cummulative driving difficulty encountered on a trip, for example, may be an important determinant of crash involvement and there are a number of indices available to measure this. The Cox model, however, does not appear to provide the same scope in sampling design / simple estimation packages offered, especially, by the binary logit model. This is an critical issue since, from the results of Jovanis and Chang, the loss of information in only considering crash trips would appear to be severe - estimation of the full model, with censoring, seems to be required.

## 7. SOFTWARE.

A number of commonly used statistical or econometric software packages contain routines to estimate some or most of the models covered in this report. In addition a number of special purpose programs exist to estimate models involving categorical dependent variables. In this section a summary is provided of a selection of software packages that can be used to estimate one or more of the models described in this report.

SPSSX (SPSS Inc. 1983): The origins of SPSS are as a general purpose mainframe statistical package for social scientists. Recently an extended version of this software package, SPSSX, has been released. One of the new routines in SPSSX is a set of procedures for fitting log-linear models. The estimation techniques are designed around aggregate cross-tabulation type data. The procedures can fit binary and multinomial logit models, hierarchical and non-hierarchical log-linear models, quasi-independence log-linear models and log-linear models with structural zeros. An extensive set of data manipulation commands is provided. Further information on SPSSX can be obtained from SPSS Inc., Suite 3300, 444 North Mitchigan Avenue, Chicago, IL 60611, U.S.A. Versions are available for most popular mainframes and the IBM PC.

LIMDEP (Greene 1986): LIMDEP is a package for estimating a variety of econometric models on both cross sectional and time series data. Of the techniques covered in this report, LIMDEP has routines to estimate binary logit and probit, unordered multinomial logit, ordered logit or probit, basic tobit, truncated regression, generalised sample selection models and proportional hazzards models. In addition there is a comprehensive coverage of regression techniques. The data manipulation commands are similar in scope to those found in SPSSX. The package also contains an extensive set of matrix operation facilities. Users can supply their own likelihood function in a FORTRAN subroutine and link this to one of LIMDEP's maximisation routines. The set of LIMDEP commands to estimate the ordered logit model outline in Section 6.2 is shown in Figure 6. Further information on LIMDEP can be obtained from Professor W. Greene, Graduate School of Business Administration, 100 Trinity Place, New York, N.Y. 10006, U.S.A.

HOTZTRAN (Avery and Hotz 1985): HOTZTRAN is a FORTRAN based statistical package which is designed to estimate discrete choice, limited dependent and linear and non-linear regression models where the models may consist of a system of one or more equations. The coverage of model types is similar to LIMDEP. HOTZTRAN is

particularly useful for multiple period (panel data) models where there are several years of data for each individual. Raw data can be read in a number of different ways with a full range of internal transformations available. Although commendably comprehensive the author experienced some difficulty in estimating some of the model types covered in this report using HOTZTRAN and users are cautioned that errors may remain. Further information on HOTZTRAN may be obtained from CERA Economic Consultants Inc., P.O. Box 159, Old Greenwich, CT 06870, U.S.A.

BLOGIT (Crittle and Johnson 1980): BLOGIT is designed specifically to estimate individual observation binary and unordered multinomial logit models. An OLS estimation procedure is also provided. For the estimation of such models BLOGIT has been widely used. This wide use can be attributed to its intial low price (at tape copying cost to academic users), its description in the textbook by Hensher and Johnson (1981) and its distribution by the Australian Road Research Board. Because of price increases, its relatively limited scope and advancing age, BLOGIT probably now compares unfavourably with packages such as LIMDEP for users other than those with a specialised interest in MNL models. It is currently undergoing upgrading. For further information on BLOGIT contact Professor D. Hensher, School of Economics, Macquarie University, N.S.W. 2109, Australia.

GQOPT: It is conceivable that no readily available 'off-the-shelf' software package will exist for estimating some advanced model types. Under these circumstances it is usual for the user to specify the model's likelihood function in a FORTRAN subroutine. As has been pointed out this subroutine could be linked to a package such as LIMDEP. Alternatively a number of programs exist specially devised to maximise user supplied functions. One such package is GQOPT. The user must supply the function to be maximised in a FORTRAN subroutine and, optionally, its first and second derivatives with respect to the parameters to be estimated. A number of algorithms are available for parameter estimation including Davidon-Fletcher-Powell and a quasi-Newton procedure. Further information on GQOPT may be obtained from Professor R.

Quandt, Department of Economics, Princeton University, Princeton, NJ 08540, U.S.A.

## 8. SUMMARY.

This report has attempted to convey an overview of an area of recent advance in econometrics and statistics in the understanding and estimation of quantal response models. It has also attempted to assess the potential applicability of these models in road crash research. The conclusion is that the models appear to be especially suited to the analysis of road crash data. Underlying this conclusion is a recognition of a number of useful features of these models when applied to road crash research. Premier among these are:

(i) that the models possess an intuitively appealing theoretical framework that challenges the analyst to think about crashes in a different way and forces a consideration of normal, non-crash driving behaviour,

(ii) that they can be estimated using individual observations rather than aggregated data, thus providing certain statistical advantages,

(iii) the models offer enormous flexibility and their probabilistic orientation permits utilisation of results from probability theory, and

(iv) the data demands of the models are slight since they can be estimated with small sample sizes and are unaffected by non-random sampling schemes once simple corrective procedures have been instituted.

Computer programs to estimate most of the models mentioned in this report are commercially available. Alternatively general-purpose programs exist for maximisation of user specified likelihood functions. An example is the GQOPT program developed by Richard Quandt. Although most of these packages endeavour to be

user friendly (with varying degrees of success), it must also be recognised that the increased flexibility and realism offered by the class of model reviewed in this report places extra demands on the analyst. **The analyst is required to *think* about the problem** at hand and the most appropriate model specification. In short the models (and software packages) demand a measure of statistical literacy, which is probably not a bad thing.

## REFERENCES

AMEMIYA, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica*, Vol. 41, pp 997-1016.

_____ (1981). Qualitative response models: a survey. *Journal of Economic Literature*, Vol. 19, pp 483-536.

ANDREASSEND, D.C. (1983). A framework for accident costing. *Australian Road Research*, Vol. 13, pp 300-302.

_____ (1985). A framework for costing accidents and accident types. *Accident Analysis and Prevention*, Vol. 17, pp 111-117.

ARNOULD, R.J. and H. GRABOWSKI (1983). Auto safety regulation: an analysis of market failure. *The Bell Journal of Economics*, Vol. 14, pp. 27 - 48.

ATKINS, A.S. (1981). The economic and social costs of road accidents in Australia. Centre for Environmental Studies, CR 21, University of Melbourne, 127pp.

AVERY, R.B. and V.J. HOTZ (1985). *HotzTran User's Manual*. CERA Economic Consultants, Inc. Old Greenwich, Connecticut.

BARNARD, P.O. (1981). The Adelaide travel demand and time allocation study: questionnaire forms, interviewers and coding manuals. Australian Road Research Board Internal Report AIR 352-2, Vermont South, Victoria, pp. 121.

_____ (1986a). Perceived bicycle safety and commuting use: results from the Adelaide Travel Demand and Time Allocation Study. Paper presented at the BIKESAFE Conference, Newcastle, Australia, May.

_____ (1986b). *Modelling Shopping Destination Choices: A Theoretical and Empirical Investigation*. Special Report SR 352-1, Australian Road Research Board, Vermont South, Victoria.

_____ (1987). Use of an activity diary survey to examine travel and activity reporting in a home interview survey. *Transportation*, Vol. 16.

_____ (1989). A model of injuries sustained in two-vehicle collisions on urban roads. Australian Road Research Board Internal Report AIR438-2, Australian Road Research Board, Vermont South, Victoria.

BERK, R.A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review*, Vol. 18, pp 386-398.

BERNDT, E.K., B.H. HALL, R.E. HALL and J.A. HAUSMAN (1974). Estimation and inference in non-linear structural models. *Annals of Economic and Social Measurement*, Vol. 3, pp 653-665.

BISHOP, T., S. FEINBERG and P. HOLLAND (1975). *Discrete Multivariate Analysis*. M.I.T. Press, Cambridge, Mass.

CAIN, G.C. (1975). Regression and selection models to improve non experimental comparisons. In C.A. Bennett and A.A. Lumsdaine (eds), *Evaluation and Experiment, Some Critical Issues in Assessing Social Programs*, Academic Press, New York, pp 297-317.

CAMPBELL, R. and R.J. FILMER (1980). Deaths from road accidents in Australia. Unpublished mimeo.

CARLSON, W.L. (1979). Crash injury prediction model. *Accident Analysis and Prevention*, Vol. 11, pp. 137 - 53.

CARLSON, W.L. (1980). Qualitative analysis of policy decisions. *Accident Analysis and Prevention*, Vol. 12, pp 41-53.

CHIRACHAVALA, T., D.E. CLEVELAND and L.P. KOSTYNIUK (1984). Severity of large-truck and combination-vehicle accidents in over-the-road service: a discrete multivariate analysis. *Transportation Research Record*, No. 975, pp. 23 - 36.

CLARKE, M.I., M.C. DIX and P.M. JONES (1985). *Household Activity - Travel Patterns in Adelaide*. Interim Report to the Director-General of Transport, South Australia, pp. 149.

CRAGG, J.G. and R.S. UHLER (1970). The demand for automobiles. *Canadian Journal of Economics*, Vol. 3, pp 386-406.

CRITTLE, F.J. and L.W. JOHNSON (1980). *Basic Logit (BLOGIT) - Technical Manual*. Australian Road Research Board Technical Manual, ATM No 9, Vermont South, Victoria.

DEACON, R. and P. SHAPIRO (1975). Private preference for collective goods revealed through voting and referenda. *American Economic Review*, Vol. 65, pp 943-955.

De DONNEA, F.X. (1971). *Modal Choice in Dutch Cities*. Rotterdam University Press, Rotterdam.

DOMENCICH, T. and D. McFADDEN (1975). *Urban Travel Demand: A Behavioral Analysis*. North Holland, Amsterdam.

FAIR, R.C. (1977). A note on the computation of the Tobit estimator. *Econometrica*, Vol. 45, pp 1723-1727.

FEINBERG, S.E. (1975). Comment. *Journal of the American Statistical Association*, Vol. 70, pp 521-524.

FIGLEWSKI, S. (1979). Subjective information and market efficiency in a betting market. *Journal of Political Economy*, Vol. 87, pp 75-88.

GARCIA-FERRER, A. and J. del HOYO (1987). Analysis of the car accident indexes in Spain: a multiple time series approach. *Journal of Business and Economic Statistics*, Vol 5, No. 1, pp. 27 - 38.

GOLDBERGER, A.S. (1964). *Econometric Theory*. Wiley, New York.

_____ (1981). Linear regression after selection. *Journal of Econometrics*, Vol. 15, pp 357-366.

GREENE, W.H. (1986). *LIMDEP: User's Manual*. Graduate School of Business Administration, New York University, New York.

HABERMAN, S.J. (1978). *Analysis of Qualitative Data: Volume 1, Introductory Topics*. Academic Press, New York.

_____ (1979). *Analysis of Qualitative Data: Volume 2*, New Developments. Academic Press, New York.

HAUER, E. (1980). Bias-by-selection: overestimation of the effectiveness of safety countermeasures caused by the process of selection for treatment. *Accident Analysis and Prevention*, Vol.12, pp 1131-1137.

HAUSMAN, J.A. and D.A. WISE (1976). The evaluation of results from truncated samples: the New Jersey negative income tax experiment. *Annals of Economic and Social Measurement*, Vol. 5, pp 421-445.

_____ (1977). Social experimentation, truncated distributions and efficient estimation. *Econometrica*, Vol. 45, pp 319-339.

HECKMAN, J. (1976a). Simultaneous equations model with continuous and discrete endogenous variables and structural shift. In S.M. Goldfeld and R.E. Quandt (eds), *Studies in Non-Linear Estimation*, Ballinger Press, Cambridge.

_____ (1976b). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, Vol. 5, pp 475-492.

_____ (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, Vol. 46, pp 931-959.

HECKMAN, J. and R. WILLIS (1976). Estimation of a stochastic model of reproduction: an econometric approach. In N. Terleckyj (ed), *Household Production and Consumption*, National Bureau of Economic Research, New York.

HENSHER, D.A. (1986). Dimensions of automobile demand: an overview of an Australian research project. *Environment and Planning A*, Vol. 18, pp. 1339 - 74.

HENSHER, D.A. and L.W. JOHNSON (1981). *Applied Discrete Choice Modelling*. Croom Helm, London.

HUTCHINSON, T.P. (1983). A bivariate normal model for intra-accident correlations of driver injury with application to the effect of mass ratio. *Accident Analysis and Prevention*, Vol. 15, No. 3, pp. 215 - 24.

JOHNSON, N.L. and S. KOTZ (1970). *Distributions in Statistics: Continuous Univariate Distributions, Volumes 1 and 2*. Houghton Mifflin, Boston.

JOHNSON, I.R. (1978). The implications of alcohol impairment for research into road and traffic system design and management. *Australian Road Research*, Vol. 8, No.4, pp. 57 - 62.

JOHNSON, I.R. and D.R. PERRY (1980). Driver behaviour research - needs and priorities. *Australian Road Research Report No. 108*, Australian Road Research Board, Vermont South, Victoria.

JOVANIS, P. AND H. CHANG (1986). Disaggregate model of highway accident occurence. Paper presented at the 65th Annual Meeting of the Transportation Research Board, Washington D.C.

KERNS, I.B. and H.J. GOLDSMITH (1984). The impact on traffic crashes of the introduction of random breath testing in New South Wales. Proceedings of the Australian Road Research Board Conference, Vol. 12, Part 7, pp. 81 - 95.

KIM, J. (1975). Multivariate analysis of ordinal variables. *American Journal of Sociology*. Vol. 81, pp. 261-298.

KNOX, B. (1983). *QUESTAT Users Guide: Part 1, Reference Manual*. Personal Social Services Research Unit, University of Kent at Canterbury, Canterbury, Kent, England.

KRISHNAN, K.S. (1983). Empirical estimates of seat belt effectiveness in two-car collisions. *Accident Analysis and Prevention*, Vol. 15, No. 3, pp. 227 - 36.

KRISHNAN, K.S., J.V. CARNAHAN and M.J. BECKMANN (1983). An injury threshold model for two-car collisions. *Management Science*, Vol. 29, pp 909-926.

LABOVITZ, S. (1970). The asswignment of numbers to rank order categories. *American Sociological Review*. Vol. 35, pp. 515-524.
KMENTA, J. (1971). *Elements of Econometrics*. Macmillan, New York.

LAVE, L. and W. WEBBER (1970). A benefit-cost analysis of auto safety features. *Applied Economics*, Vol. 2, pp. 265 - 75.

LAY, M.G. (1984). *Source Book for Australian Roads*. Australian Road Research Board, Vermont South, Victoria, 2nd Edition, 551pp.

LAYTON, A.P. (1983). The impact of increased penalties on Australian drink / driving behaviour. *Logistics and Transportation Review*, Vol. 19, No. 3, pp. 261 - 66.

LAYTON, A.P. and J.C. WEIGH (1983). The efficacy of some recent Australian road safety policy initiatives. *Logistics and Transportation Review*, Vol. 19, No. 3, pp. 267 - 78.

LEE, L.F. (1978). Unionism and wage rates: a simultaneous equation model with qualitative and limited dependent variables. *International Economic Review*, Vol. 19, pp 415-433.

MADDALA, G.S. (1977). *Econometrics*. McGraw-Hill, New York.

_____ (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge.

MANSKI, C. and S.R. LERMAN (1977). The estimation of choice probabilities from choice-based samples. *Econometrica*, Vol. 45, pp 1977-1988.

MANSKI, C. and D McFADDEN (eds) (1982). *Structural Analysis of Discrete Data with Econometric Applications*. M.I.T. Press, Cambridge, Mass.

MARQUART, J.F. (1974). Vehicle and occupant factors that determine occupant injury. **Paper presented at SAE Automotive Engineering** Congress, Detroit, Michigan.

McCULLAGH, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society*, Series B, Vol. 42, pp 109-142.

McFADDEN, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed), *Frontiers in Econometrics*, Academic Press, New York.

_____ (1978). Modelling the choice of residential location. In A. Karlquist, L. Lundquist, F. Snickers and J. Weibull (eds), *Spatial Interaction Theory and Residential Location*, North Holland, Amsterdam, pp 75-96.

McKELVEY, R.D. and W. ZAVOINA (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, Vol. 4, pp. 103 - 20.

MISSIAKOULIS, S. (1983). *QUESTAT Users Guide: Part 2, A Guide to the Models and their Statistics*. Personal Social Services Research Unit, University of Kent at Canterbury, Canterbury, Kent, England.

MUTHEN, B. and K.G. JORESKOG (1983). Selectivity problems in quasi-experimental studies. *Evaluation Review*, Vol. 7, pp 139-174.

NERLOVE, M. and J. PRESS (1973). Univariate and multivariate log-linear and logistic models. RAND report R-1306-EDA/NIH.

PAK POY, P.G. and ASSOCIATES (1978). *Metropolitan Adelaide Data Base Study: Phases 1-5.* Reports prepared for the South Australian Department of Transport and Highways Department, Adelaide, Australia.

PEARSON, K. (1900). Mathematical contributions to the theory of evolution in the inheritance of characteristics not capable of exact quantitative measurement, VIII. *Philosophical Transactions of the Royal Society*, Series A, Vol. 195, pp 79-150.

ROAD ACCIDENT RESEARCH UNIT (1979): **Adelaide In-Depth Accident** Study, 1975 - 79, Parts 1 - 10. University of Adelaide, Adelaide, Australia.

SCHMIDT, P. and R.P. STRAUSS (1975). Estimation of logit models with jointly dependent qualitative variables: a simultaneous logit approach. *Econometrica*, Vol. 43, pp 745-755.

SEARLE, B. (1980). Unreported traffic crashes in Sydney. *Proceedings of the 10th Conference of the Australian Road Research Board*, Vol. 10, Part 4, pp 62-74.

SEGAL, S. (1956). *Nonparametric Statics*. McGraw-Hill, New York.

TOBIN, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, Vol. 26, pp 24-36.

TRAIN, K. (1986). *Qualitative Choice Analysis*. M.I.T. Press, Cambridge, Mass.

TRINCA, G. (1980). Medical aspects of seat belt usage. *Journal of Traffic Medicine*, Vol. 8, No. 3, pp 36 - 38.

TYE, W.B., L. SHERMAN, M. KINNUCAN, D. NELSON and T TARDIFF (1982). Application of disaggregate travel demand models. **National** Cooperative Highway Research Program, Report 253, Transportation Research Board, Washington D.C., 207pp.

VIANO, D.C., R.C. HAUT, M. GOLOCOUSKY and K. ABSOLOM (1978). Factors influencing biomechanical response and closed chest trauma in experimental thoracic impacts. Paper presented at the 22nd Conference of the American Automobile Association for Automotive Medicine, Ann Arbor, Michigan.

WALLER, J.A., J.R. STEWART, A.R. HANSEN, J.C. STUTTS, C.L. POPKIN and E.A. RODGMAN (1986). **The potentiating effects of alcohol on** driver injury. Journal of the American Medical Association, Vol. 256, pp. 1461 - 66.

WHITE, K.J. (1978). A general computer program for econometric methods - SHAZAM. *Econometrica*, Vol. 46, pp 239-240.

WIGAN, M.R. (1982). Bicycle ownership, use and exposure in Melbourne 1978-9. **Australian Road Research Board Research Report**, ARR 130, pp. 41.

WINSHIP, C. and R.D. MARE (1984). Regression models with ordinal variables. *American Sociological Review*, Vol. 49, pp. 512 - 25.

WRIGLEY, N. (1979). Developments in the statistical analysis of categorical data. *Progress in Human Geography*, Vol. 3, pp 315-355.

_____ (1981). Categorical data analysis. In N. Wrigley and R.J. Bennett **(eds)**, *Quantitative Geography: A British View*, Routledge and Kegan Paul, London.

_____ (1985). *Categorical Data Analysis for Geographers and Environmental Scientists*. Longman, London.

YULE, G.U. (1900). On the association of attributes in statistics. *Philosophical Transactions of the Royal Society*, Series A, Vol. 194, pp 257-319.

ZLATOPER, T.J. (1984). Regression analysis of time series data on motor vehicle deaths in the United States. *Journal of Transport Economics and Policy*, Vol. XVIII, No. 3, pp. 263 - 74.