Pergamon

Accid. Anal. and Prev., Vol. 30, No. 4, pp. 425–433, 1998
© 1998 Elsevier Science Ltd. All rights reserved
Printed in Great Britain
0001-4575/98 $19.00 + 0.00

PII: S0001-4575(97)00122-X

HETEROGENEITY CONSIDERATIONS IN ACCIDENT MODELING

MATTHEW G. KARLAFTIS[1]* and ANDRZEJ P. TARKO[2]

[1]School of Engineering, Hellenic Air-Force, 106 Themistokleous Str., Athens, 106-81, Greece
and [2]School of Civil Engineering, Purdue University, 128 Civil Engineering bldg., West Lafayette,
IN 47907-1284, USA

(Received 2 May 1997; in revised form 24 October 1997)

Abstract—Panel data sets are becoming readily available and increasingly popular in safety research. Despite its advantages, panel data raises new specification issues, the most important of which is heterogeneity, which have not been addressed in previous studies in the safety area. Based on a county accident data set, the present analysis extends prior research in a significant direction. There is an explicit effort to control for cross-section heterogeneity that may otherwise seriously bias the resulting estimates and invalidate statistical tests. Because common modeling techniques such as the fixed and random effects models, developed to account for heterogeneity, are impractical for count data, this study uses cluster analysis to overcome this. First, observations are disaggregated into homogeneous clusters. Then, separate negative binomial models including a time trend factor are developed for each group. The results clearly indicate that there are significant differences between the models developed, and that separate models describe data more efficiently than the joint model. © 1998 Elsevier Science Ltd. All rights reserved

Keywords—Accident modeling, Clustering, Heterogeneity, Negative binomial models

INTRODUCTION

Road safety modeling attracts considerable research interest because of its wide variety of applications. Traffic engineers, for example, are interested in identifying those factors that influence accident frequency and severity to improve highway design and to provide a safer driving environment. State Departments of Transportation (DOT) and other agencies wish to identify high accident areas to promote a variety of safety treatments. Traditionally, these problems have been addressed in a multiple or a Poisson and negative binomial regression framework. Owing to the increasing availability of a combination of cross-section and time-series data (referred to as panel data) in the safety area, these models may not accurately depict the qualitative and quantitative relationship between accidents and the various variables used.

The increasing interest in safety and accident investigation was promoted by the Federal government with the requirement for a safety management system (SMS) after passing the Intermodal Surface Transportation Efficiency Act (ISTEA) of 1991. This legislation required that the states identify and implement all opportunities to improve roadway safety in the planning, construction, maintenance and operations phases. Even though the SMS is not Federally required as of 1996, the Federal Highway Administration (FHWA) encourages the states to pursue its development. Most states that we are aware of, continue to work on the development of statewide SMSS, suggesting the need for improving on existing empirical models for accident measurement.

This paper focuses on the development of improved accident models, with panel data, that can be used in a wide variety of applications such as estimating crashes for urban and suburban arterial sections (Bowman et al., 1994), crash rates for different median types (Knuiman et al., 1993), identifying injury and fatal crashes as a function of various roadway charactreristics (Hadi et al., 1993), and identifying areas and spots for safety treatment (Tarko et al., 1996). The remainder of this paper is organized as follows. In the next section we provide the background necessary for the development of the models presented in this paper.

Following this, we present and discuss the data and methodology that were used, as well as the estimated models. The final section of the paper contains concluding remarks.

*Corresponding author. e-mail: karlafti@compulink.gr

BACKGROUND

Much of the early work in the empirical analysis of accident data was done with the use of multiple linear regression models (a comprehensive review of the earlier work appears in Hadi et al., 1993). These models suffer from several methodological limitations and practical inconsistencies which have been pointed out repeatedly in the literature (see for example Lerman and Gonzales, 1980; Ivan and O'Mara, 1997). To overcome these limitations, several authors used Poisson regression models which are a reasonable alternative for events that occur randomly and independently over time.

The Poisson model has a number of advantages over the normal regression model when dealing with *count* data (as, for example, accident count data), First, linear regression assumes a normal distribution of the dependent variable, an assumption which does not hold with count (accident) data. The Poisson model, on the other hand, recognizes the discrete nature of accident counts. Second, linear regression may produce negative estimates for the dependent variable, which is incorrect for accident counts. Despite its advantages, Poisson regression assumes equality of the variance and mean of the dependent variable. However, it is quite common to have the variance of the data substantially higher than the mean (Dean and Lawless, 1989). This phenomenon, called "overdispersion," leads to invalid *t*-tests of the estimated parameters. The restriction on the variance imposed by the Poisson model can be overcome with the use of negative binomial regression, which allows the variance of the dependent variable to be larger than the mean.

Much literature exists on accident analysis using Poisson and negative binomial regression models. For example, Bowman et al. (1994) and Knuiman et al. (1993) used negative binomial regression models with panel data to develop equations to estimate crashes for urban and suburban arterial sections (Bowman et al., 1994) and crash rates (Knuiman et al., 1993) with different median widths and types as the independent variables. In a similar study, Hadi et al. (1993) used negative binomial modeling with panel data (large number of roadway sections from 1988–1991) from the Florida Department of Transportation to identify injury and fatal crashes as a function of various roadway geometric characteristics. More recently, Ivan and O'Mara (1997) used this type of regression model to determine the geometric factors which affect accidents in the state of Connecticut. In a relatively different line of research, Tarko et al. (1996) used negative binomial regression with area wide panel data from Indiana (92 counties from 1988–1993) to systematically analyze safety problems in counties. The developed models allowed the researchers to establish a method to prioritize counties with safety problems using infrastructure, demographic and traffic characteristics.

State-of-the-art accident analysis models, such as the ones reviewed above, have typically been developed using cross-section or panel data sets, but have not accounted for the presence of heterogeneity, which is likely to exist in such data. Heterogeneity refers to the presence of persistent site-specific (or area-specific) but unobserved factors. If not accounted for, heterogeneity may lead to biased coefficient estimates for the models (Greene, 1991).

In this paper we introduce a methodology which attempts to recognize the possible existence of heterogeneity in count data models and develop a framework to account for it. The example model is to be developed using county accident data from Tarko et al. (1996). Heterogeneity occurs in a structural model based on panel data if the differences among cross-sectional units are not appropriately reflected in the existing explanatory variables. If this heterogeneity is not accounted for, it is captured by the error term leading to biased parameter estimates. To account for heterogeneity, two general modeling formulations exist: the fixed effects and the random effects models. Unfortunately, as we will explain in the next section, these two formulations do not lend themselves to forecasting, and are thus of limited practical importance in safety analyses. In an attempt to overcome part of this problem, we utilize a stratification scheme which allows accidents in counties of similar socioeconomic, traffic, and infrastructure characteristics to be analyzed and compared with each other. This scheme allows for safety inferences to be made for comparable groups of counties, without creating a situation in which the number of accidents for a small rural county are compared with those of a highly urbanized county. It is expected that analysing homogeneous groups of counties will be a positive step toward recognizing and minimizing the heterogeneity problem. It is worth noting that the same approach can be used to model accident counts and crash rates at individual intersections and on highway sections.

DATA AND METHODOLOGY

Data

In order to provide additional insights into the effects that various infrastructure, socioeconomic and traffic characteristics have upon county crashes, the analysis in this paper employs cross-section time-series (panel) data. The data were obtained from the

Indiana police records for 14 crash categories. The data contain observations for the 92 counties of Indiana for 6 years (1988–1993). For each of those counties and years there is a set of potential explanatory variables that we used to build the crash models (Table 1). In essence, the data comprise observations for 92 counties in Indiana. For each county and for each year, we have observations for various crash categories along with information on the various macroscopic explanatory variables. This annual data for each county pooled together forms our (panel) data set.

Analyses based on panel data sets posses several advantages over analyses based upon cross-section or time-series data (Hsiao, 1993). From a statistical perspective, by increasing the number of observations, panel data have higher degrees of freedom and less collinearity (particularly in comparison with time-series data), thus improving the efficiency of the parameter estimates. Moreover, panel data allow a researcher to answer questions that cannot be adequately addressed in time-series or cross-section data. For example, suppose we find from a cross-section of counties that, on average, vehicle miles traveled (VMT) raise fatal crashes by 20%. If we have a homogeneous population of counties, it means that each county's fatal crashes increases 20%. This could also be an implication for a sample of heterogeneous counties. However, an alternative explanation in a sample of heterogeneous counties is that the VMT have no effect on (4/5) of the counties and raise fatal crashes 100% on (1/5) of the counties. Although we cannot distinguish these hypotheses in a cross-section sample, it is possible to discriminate between them by identifying the effect of VMTs on a cross-section of time-series for the different counties.

Further, in comparison with cross-sectional data, panel data raises new specification issues that have generally not been considered in the accident literature. The most important of these is heterogeneity bias (Hausman and Taylor, 1981). As alluded to above, heterogeneity refers to the differences across cross-sectional units that may not be appropriately reflected through explanatory variables included in the model. If one does not correct for the heterogeneity across cross-sectional units, the estimated parameters are biased since they capture part of the heterogeneity. Indeed, as noted by Greene (1991), cross-section heterogeneity should be the central focus of panel data analysis.

Table 1. Independent variables

Variable type	Code
Number of licensed drivers (thousands)	license
Number of registered vehicles (millions)	veh
Population (millions)	POP
Median family income (thousands)	inc
Total road mileage (thousands of miles)	mil
Proportion of state roads in total mileage	st_mil
Proportion of city roads in total mileage	city_mil
Total VMT (1000 miles/day)	VMT
Proportion of rural roads in total VMT	rur_VMT
Proportion of urban roads in total VMT	ur_VMT

In general, heterogeneity has long been recognized as a specification problem in econometrics, and several models and methodologies have been developed to account for it. Nevertheless, these types of remedies have not been used in safety research. This may be a result of two things. First, the fixed and random effects specifications developed to account for heterogeneity problems in panel data can be used only in part, as we will see in the next section, in problems involving count data (such as the safety and accident related problems). Second, the collection and emergence of panel data sets in safety research is fairly recent. As a result, efforts will be made in the near future to account for panel data problems in safety research. The contribution of this paper is in the direction of incorporating panel effects in safety related research.

In contrast to previous studies, this paper explicitly recognizes the possible existence of heterogeneity and introduces a two-step procedure for overcoming it: first, we use cluster analysis to group counties in homogeneous groups; second, we utilize separate negative binomial regression models for each cluster of counties to quantify the effects of several characteristics on county-wide accidents. It is expected that this methodology will outperform some of the state-of-the-art specifications, by developing models over homogeneous groups which account for potential heterogeneity that exists between groups, and thus the developed models will have improved parameter estimates and reduced error variance.

The methodology

The usual Poisson regression framework assumes that the response (dependent) variable is Poisson-distributed. Specifically, for county i, the number of a certain type of crashes is given by the following Poisson probability mass function:

$$p(Y=y_i)=\frac{e^{\lambda_i}\lambda_i^{y_i}}{y_i!},\ y=0,1,\ldots, \qquad (1)$$

where: y_i = the number of a certain type of crashes at county i for a given time period; λ_j = in this model, it is both the mean and variance of y_i.

To develop a crash model using this approach, λ_i is expressed as a function of some of the explanatory variables presented in Table 1. The relationship

Table 2. Descriptive statistics for the independent variables

	Number of counties	Variable[a]											
		ur-VMT		rur-VMT		pop		mil		st-mil		city_mil	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
All counties	92	0.19	0.21	0.81	0.22	0.06	0.1	0.99	0.44	0.12	0.029	0.12	0.11
Cluster 1 (urban)	6	0.839	0.061	0.161	0.161	0.69	0.16	3.074	0.462	0.074	0.022	0.525	0.101
Cluster 2 (suburban)	29	0.377	9.141	0.623	0.138	0.27	0.1	1.999	0.509	0.094	0.014	0.379	0.14
Cluster 3 (rural)	57	0.04	0.09	0.96	0.06	0.041	0.034	0.923	0.268	0.131	0.028	0.109	0.074

[a]Table I contains the full name of each variable

between λ_i and the explanatory variables is of the following form:

$$\ln \lambda_i = \boldsymbol{b}'\boldsymbol{x}_i, \qquad (2)$$

where: $\mathbf{b}$ = vector of parameters to be estimated; $\mathbf{x}_i$ = vector of erogenous variables for county i.

The Poisson model has been criticized because of its implicit assumption that the variance of the crash process at individual locations equals its mean. When this assumption is violated (the mean and the variance are not equal), the efficiency of the parameter estimates is lost, and the t-statistics are corrupt since they are based on biased standard errors.

Essentially, the Poisson model was proposed based on the empirical indications that the process of crash counts at individual locations is Poisson (Nicholson, 1985). If some explanatory factors are missing, as is frequently the case, the incomplete model is forced to absorb the overdispersion (over Poisson variance) causing overoptimistic estimation of the model efficiency, yielding corrupt t-statistics. To overcome this limitation, researchers have commonly used the negative binomial model. This is an extension of the Poisson model that allows for the variance of the process to differ from the mean. One way this model arises is as a modification of the Poisson model in which λ_i is specified as follows:

$$\ln \lambda_i = \boldsymbol{b}'\boldsymbol{x}_i + \epsilon_i, \qquad (3)$$

where: ϵ_i = random error representing the effect of omitted explanatory variables.

Greene (1991) presents the negative binomial probability mass function which can be written as:

$$p(Y = y_i) = \frac{\Gamma\left(\dfrac{1}{a_i} + y_i\right)}{\Gamma\left(\dfrac{1}{a_i}\right) y_i!} \left(\frac{1}{1 + a_i \lambda_{it}}\right)^{\frac{1}{a_i}} \left(1 - \frac{1}{1 + a_i \lambda_{it}}\right)^{y_i}, \qquad (4)$$

and

$$\frac{\mathrm{Var}(y_i|\boldsymbol{x}_i)}{E(y_i|\boldsymbol{x}_i)} = 1 + a_i E(y_i|\boldsymbol{x}_i),$$

where: $\Gamma(.)$ = gamma function; a_i = rate of "overdispersion."

Until quite recently, incorporating heterogeneity (or panel effects) in either the Poisson or negative binomial frameworks was not possible. In 1984, Hausman et al. (1986) developed a formulation which accounts for the existence of panel effects in the Poisson and negative binomial models. The fixed effects Poisson model, for example, can be written as:[1]

$$\ln \lambda_{it} = k_i + \boldsymbol{b}'\boldsymbol{x}_{it}, \qquad i = 1, \ldots N, \qquad t = 1, \ldots, T_i \qquad (5)$$

where: k_i = is a 1×1 parameter representing the average effect of excluded variables specific to the ith county; N = number of counties; T_i = time period for an observation in the ith county;

This model is estimated by considering the *conditional* joint distribution

$$f(y_{i1}, \ldots, y_{iT} | \sum_T y_{iT}).$$

The resulting density is a function of the $\mathbf{b}$ alone, which is then estimated using maximum likelihood. Unfortunately, since the fixed effects are conditioned out, *not* computed, the fixed effects parameters, the marginal effects, and the predicted values *cannot* be estimated for this model. The same problem exists for the fixed and random effects negative binomial model.

[1] We present here only the initial part of the formulation which might be of interest to the reader. Readers interested in the panel effects formulation at a more detailed level are urged to refer to the original paper (Hausman et al., 1986), or to Greene (1991) for a more simplified, yet thorough, explanation of the estimation process.

When heterogeneous data are pooled, this shortcoming of the fixed effects Poisson, and the fixed and random effects negative binomial models makes their use impractical in the area of safety analysis. The ability to obtain marginal effects and to predict accident rates is an indispensable component of safety analyses. And, while the fixed and random effects models cannot be of practical assistance in accounting for heterogeneity, we need to analyze accidents in a manner that minimizes the bias associated with this phenomenon.[2] To achieve this, we first perform a cluster analysis whose goal is to create clusters which are as internally homogeneous and as externally heterogeneous as possible. When data are pooled together and a single model is estimated over this data, the implicit assumption is made that counties are homogeneous. For each group the slopes (of the models) accurately capture the effects of the independent variables on accident counts. Then, we estimate different negative binomial models for the different groups of counties, yielding the appropriate slopes, marginal effects, and predicted accident counts for each group. That is, we accept the Negative Binomial model structure as correct, but attempt to adjust for heterogeneity by estimating different models over homogeneous groups of data. As a result, these models will now have improved parameter estimates and reduced error variance.

CLUSTER ANALYSIS

Cluster analysis is a statistical technique which groups items together on the basis of similarities or distances (dissimilarities). Hierarchical cluster analysis, using Ward's minimum distance algorithm (Hartgen and Segedy, 1987), was employed to cluster the 552 observations based on the independent variables in Table 1. Ward's model is a common and straightforward model which uses minimum Euclidean distance between observations. In this method, the distance between two clusters is the square root of the ANOVA sum of squares between the two clusters, added up over all variables. As observations are grouped into clusters, Ward's method minimizes the within-cluster sum of squares, while maximizing the between-clusters sum of squares, thus creating clusters which are internally as homogeneous and externally as heterogeneous as possible.

The choice of how many clusters to use is largely up to the analyst, so for the purposes of this analysis we considered between two and six clusters. It is important to note that the number of clusters was chosen after a thorough search. Counties were initially clustered based on all the independent variables, then on several subsets (combinations) of independent variables, and finally on different subsets of the accident categories. The results remained very stable throughout this process. Regardless of the variables on which we based the cluster analysis, most counties remained in the same (or neighboring) clusters.

The analysis yielding the most robust results gave us three general groups (clusters) of counties.[3] As Table 2 indicates, the first group includes counties with relatively large and dense urban areas, group 2 consists of moderately urbanized counties, while group 3 is a collection of rural counties. After we assigned counties to different clusters, two questions were addressed. First, are the counties within each cluster homogeneous with respect to the variables considered? Second, are the systems of various clusters significantly different from each other with respect to the same variables?

The first question is answered with the minimal coefficient of variation criterion used in the clustering process (Johnson and Wichern, 1992). The final groupings showed a high degree of homogeneity as indicated by low coefficients of variation values associated with individual groups relative to the overall values. To answer the second question, we used pairwise t-tests. The results from this analysis indicated that, with the exception of household income, all other variables are significantly different between all groups at the 99% significance level.

MODEL ESTIMATION

As previously mentioned, the available data is a set of 14 different crash categories for 92 counties in Indiana. In this paper we develop four negative binomial models for crashes involving aged drivers.[4] The first model uses the pooled data from all 92 counties. The remaining models use data that correspond to the urban (six counties), suburban (29

[2] Besides heterogeneity, serial correlation is also commonly observed in panel data. In this (count/panel data) framework we are not aware of an explicit methodology to account for this problem. In the models estimated in this paper we have included a simple time trend to, at least in part, recognize this problem. Extensive treatment of serial correlation in time-series count models can be found in Johansson (1996).

[3] We should note that *regardless* of what variables were used in this clustering process, the same counties were clustered in the same groups in the three cluster case.

[4] For the purposes of this research we used the exact same methodology for 14 other crash categories. Since the main purpose of this paper is to demonstrate the potential of the proposed methodology, we only report the results related to accidents involving aged drivers. Results for the other crash categories are available upon request from the authors.

Table 3. Estimation results for the negative binomial models

Dependent variable:	Aged driver (all counties)		Aged driver (urban)		Aged driver (suburban)		Aged driver (rural)	
	Coefficient estimates (*t*-ratios)		Coefficient estimates (*t*-ratios)		Coefficient estimates (*t*-ratios)		Coefficient estimates (*t*-ratios)	
Independent variable	NB[a]	MEs[b]	NB	MEs	NB	MEs	NB	MEs
Intercept	3.02		6.02		3.65		3.01	
	(10.74)		(10.08)		(11.35)		(10.79)	
VMT	0.00032	0.019	0.000048	0.091	0.000035	0.0093	0.00023	0.014
	(2.05)		(1.38)		(0.38)		(2.01)	
POP	12.05	1083.8	0.18	351.85	5.41	1438.2	12.01	1083.1
	(4.87)		(0.25)		(4.49)		(4.34)	
inc	−0.0012	−0.11	0.0047	7.88	−0.0099	−2.66	−0.0011	0.11
	(−0.23)		(0.22)		(−1.55)		(−0.23)	
mil	0.85	76.89	0.25	484.11	0.86	229.03	0.85	76.80
	(7.21)		(2.03)		(6.27)		(7.21)	
st-mil	−0.31	−27.32	−0.46	−884.61	1.33	354.3	−0.30	27.02
	(−0.34)		(−0.28)		(2.1)		(−0.34)	
city_mil	3.73	336.08	0.57	1089.4	3.22	859.44	3.75	335.6
	(4.08)		(2.13)		(5.48)		(4.12)	
ur-VMT	1.54	138.88	0.42	795.59	0.86	230.75	1.56	137.55
	(4.78)		(1.43)		(2.97)		(4.80)	
time	−0.075	−0.68	−0.095	−18.11	−0.043	−1.15	−0.076	0.55
	(−2.79)		(−1.98)		(−1.41)		(−2.73)	
α[c]	1.65		1.30		1.41		1.667	
	(11.47)		(1.97)		(8.012)		(11.51)	
Summary statistics								
No. observations	552		36		174		342	
L(O)	−8142.99		−6893.22		−12117.53		−8044.24	
L(B)	−2918.18		−662.41		−1747.82		−1619.67	

[a] Negative binomial model.
[b] Marginal Effects.
[c] Overdispersion parameter.

counties), and rural (57 counties) clusters, as determined in the previous section.

The independent variables considered in the final model specification are taken from previous research (Tarko et al., 1996) and Akaike's information criterion (AIC).[5]

AIC $=\chi^2+2q$, where χ^2 is the goodness-of-fit χ^2 (chi-square), and q represents the number of unknown parameters solved for in the model being fitted. In general, the smaller the value of AIC, the better the model. Starting with a full set of independent variables, a stepwise procedure was used to select the best model based on minimizing the value of AIC.

The estimation results for the negative binomial models associated with the crash category mentioned above are presented in Table 3. The results are in agreement with a priori hypotheses and previous research (Tarko et al., 1996). In general, higher VMTs are associated with an increased number of accidents, as are population, total road mileage, the proportion of city mileage, and the proportion of urban roads in total VMT. The effects of the proportion of state roads depends on the data used to estimate the model. In the case of suburban counties the coefficient sign is positive and statistically significant, while in all others cases it is negative (but statistically not significant). This type of macro model (with macroscopic independent variables) can be used to systematically analyze safety problems in counties. For example, by using the values of the independent variables, a researcher can estimate the expected value of a given type of accident in a given county, and observe if the actual number of accidents is higher or lower than the expected.[6] Further, the same methodology can be used to to estimate expected number of accidents in some future point, at an aggregate level.

The time variable, represented by a simple linear time trend with the value of 1 for 1988, 2 for 1989, and so on, was included to capture a potential change in the overall accident level with time. The high *t*-statistics on this variable indicate that there is indeed a change in the overall level of accidents, holding all

[5] In general, the smaller the value of AIC, the better the model. Starting with a full set of independent variables, a stepwise procedure was used to select the best model based on minimizing the value of AIC.

[6] For a detailed explanation on how these models can be used for identifying areas and spots for safety treatment see Tarko et al. (1996).

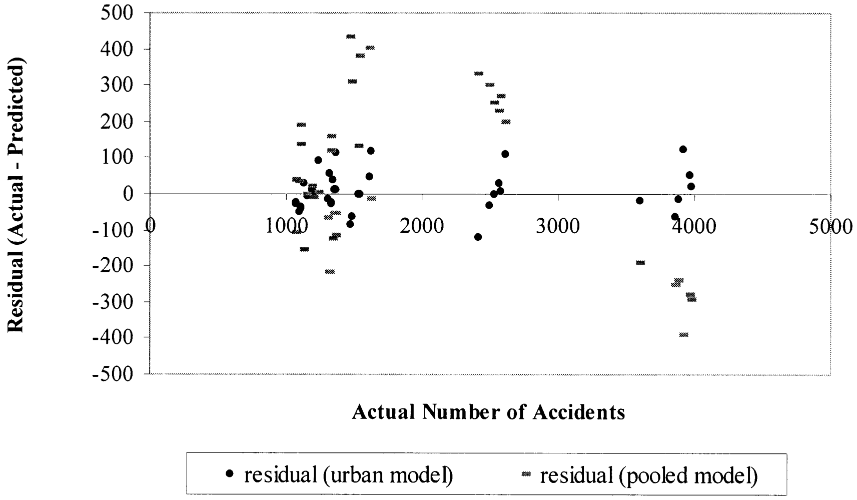

Fig. 1. Actual versus residual accident counts for nurban counties (1988–1993).

other variables constant. It is interesting to note that the overall accidents involving aged drivers in Indiana counties tends to decrease over time for all types of counties.[7]

The coefficient *a*, which captures the rate of overdispersion is highly significant, indicating that the negative binomial model is the appropriate model for the data. What is very important to note in these results is the difference in magnitude, and occasionally sign of the coefficient estimates. This result, coupled with the vastly different magnitudes of marginal effects for the variables, seems to suggest that heterogeneity is present in the sample even after accounting for the effects of VMT, population, income, etc. and is accounted for with the help of the cluster analysis. The marginal effects are very often overlooked in the literature, even though they are very important both in terms of evaluating the results of the model, and in terms of the practical use of the model. The marginal effects (Table 3) show the change in the dependent variable (accidents involving aged drivers) due to a one unit increase in some exogenous factor (independent variable). This is similar to the interpretation of the usual linear regression coefficients. For example, each additional 1,000,000 increase in population increases the accidents involving aged drivers by 1083 (for the pooled model).

It is worth noting that, had the models been estimated without first clustering the counties, the results would have been seriously skewed toward the rural counties (which are predominant in Indiana). This suggests that clustering counties in homogeneous groups before analysing them could yield results that are more accurate and could lead to more valid policy recommendations concerning safety. That is, the clustering technique can be used to improve parameter estimation and reduce error variance in safety modeling. As a final note, we graphed actual accident counts versus the residual values (where the residuals are defined as actual count minus predicted count) from each model for each county category. Figure 1 depicts the model residuals for the urban counties for 1988–1993. It is apparent that the residuals from the pooled model give predictions that are much worse than the dedicated urban model. Figure 2 shows the same information for the suburban counties. In the case of rural counties both models gave very similar predictions, so we only graphed the residuals from the pooled model (Fig. 3). In sum, Figs. 1 and 2 graphically demonstrate the need for separate models for the different county categories, as well as the improvement in accident prediction obtained with the combined use of clustering techniques and negative binomial regression.

It is also important to note that the significance of many of the coefficient estimates in the pooled negative binomial model is lost in some of the disaggregate models. The t-statistic of the VMT coefficient drops from 2.05 in the pooled model to 0.39 in the model for the suburban counties. This result seems to strongly suggest that, in the pooled model, the efficient estimates for some of the variables are skewed toward the direction of the rural counties, thus biasing the results for the urban and suburban counties.

As can be seen from Table 3, the coefficients for the pooled and rural models differ from those of the urban and suburban models. To examine this point even further, we performed a χ^2-test of the equality of the coefficient between the different models. The results of this test strongly rejected the null hypothesis of coefficient equality between the pooled and the

[7] This results does not hold universally, since for same accident categories, such as fatalities, the trend was positive.

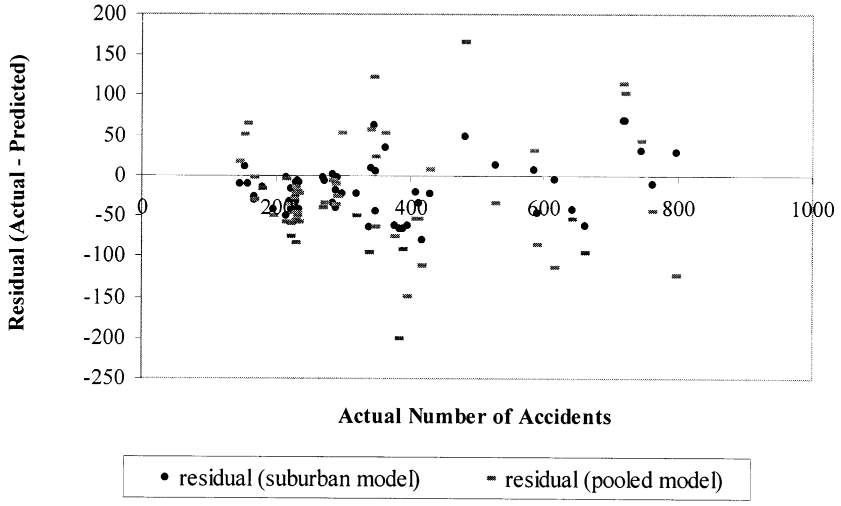

Fig. 2. Actual versus residual accident counts for urban counties (1988–1993).

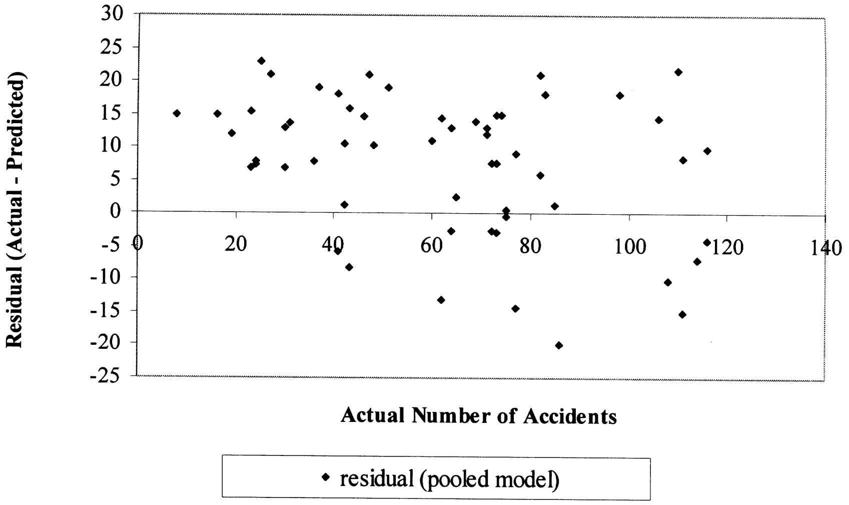

Fig. 3. Residual accident counts for rural counties (1988–1993).

urban and suburban models at the 99% level of significance, indicating that separate models should be developed.

CONCLUSIONS

Panel data are becoming increasingly popular in safety analyses. This type of data possesses several advantages over purely cross-section or time-series data. Nevertheless, use of panel data requires that special provisions be made to account for the presence of heterogeneity, which is likely to exist in such data. If not accounted for, heterogeneity may lead to biased coefficient estimates for the models (Greene, 1991). Unfortunately, the fixed and random effects formulations, the most common ways of accounting for heterogeneity, are impractical in this count data setting. First, they do not allow for predicted values to be obtained. Second, since the fixed and random effects parameters are not estimated (Hausman et al., 1986), the marginal effects (crash reduction factors) cannot be estimated. As a result, fixed and random effects count models cannot be used for predictive purposes which is of primary interest to traffic engineers and safety personnel.

In this paper we developed negative binomial models for accidents involving aged drivers. First, we performed cluster analysis to separate counties in Indiana into three separate groups: urban, suburban, and rural counties. Then, we estimated a separate model for each individual group of counties as well as a model for the pooled data. The results clearly suggest that there are indeed significant differences between the models, and that separate models should be developed for different clusters of counties. These models yield improved and possibly more accurate results when compared with the pooled negative binomial model. This result suggests that, in the model that does not account for county differences, the coefficient estimates for some of the variables are skewed toward the rural counties, thus *biasing* the estimated coefficients, especially for the urban and suburban counties.

The methodology presented in this paper can be of assistance in developing improved models over those which do not attempt to account for possible differences in the nature of the counties, spots, or highway sections examined. Panel data sets posses, in general, several advantages over simpler cross-section or time-series data sets. Nevertheless, there are some specification and estimation issues related to panel data that necessitate explicit treatment. Panel data sets are becoming increasingly available in the safety and accident investigation areas, and these specification issues deserve attention and further investigation. The methodology presented in this paper is a step toward this direction.

REFERENCES

Bowman, B. L., Vecellio, R. L. and Miao, J. (1994) Estimating vehicle and pedestrian accidents from different median types. Presented in the 1994 Transportation Research Board Meeting, Washington, DC.

Dean, C. and Lawless, J. F. (1989) Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association* **84**, 406, 467–472.

Greene, W. H. (1991) Econometric Analysis. MacMillan Press, New York, NY.

Johansson, P. (1996) Speed limitation and motorway casualties: a time series count data regression approach. *Accident Analysis and Prevention* **28**, 1, 73–87.

Johnson, R. A. and Wichern, D. N. (1992) Applied Multivariate Statistical Analysis. Prentice-Hall, Englewood Cliffs, NJ.

Hadi, M. A., Aruldhas, J., Chow, L. F. and Wattleworth, J. A. (1993) Estimating safety effects of cross-section design for various highway types using negative binomial regression. *Transportation Research Record* **1500**, 169–177.

Hausman, J. A. and Taylor, D. E. (1981) Panel data and unobservable individual effects. *Econometrica* **49**, 1377–1398.

Hausman, J., Hall, B. and Griliches, Z. (1986) Econometric models for count data with an application to the patents–R & D relationship. *Econometrica* **52**, 4, 909–938.

Hartgen, D. T. and Segedy, J. A. (1987) Peer groups for transit system performance. (Working Paper).: The University of North Carolina at Charlotte, Center for Interdisciplinary Transportation Studies.

Hsiao, C. (1993) Analysis of Panel Data, Cambridge University Press, Cambridge, UK.

Knuiman, M. W., Council, F. M. and Reinfurt, D. W. (1993) The effect of median width on highway accident rates. *Transportation Research Record* **1401**, 70–80.

Lerman, S. R. and Gonzales, S. L. (1980) Poisson regression analysis under alternate sampling strategies. *Transportation Science* **14**, 4, 346–364.

Nicholson, A. J. (1985) The Variability of accident counts. *Accident Analysis and Prevention* **17**, 1, 47–56.

Ivan, J. N. and O'Mara, P. J. (1997) Prediction of traffic accident rates using Poisson regression. Presented in the 1997 Transportation Research Board Meeting, Washington, DC.

Tarko, A. P., Sinha, K. C. and Farooq, O. A. (1996) Methodology for identifying highway safety problem areas. Presented in the 1997 Transportation Research Board Meeting, Washington, DC.