Application of Poisson Underreporting Model to Examine Crash Frequencies at Signalized Three-Legged Intersections

S. S. P. Kumara and H. C. Chin

Before effective remedial treatments can be implemented at hazardous intersections, it is often necessary to identify the causal factors affecting accident frequency. However, a problem often encountered in safety studies is the underreporting of accidents. This biased reporting may affect the selection of the parsimonious model. This study analyzes the factors affecting road accident frequency at three-legged signalized intersection approaches in Singapore, with special emphasis to underreporting. The annual accidents at 104 three-legged signalized intersections are modeled as the sum of observed Poisson outcome of accident reports. The reporting mechanism is introduced as a probit equation. The model shows that several geometric, traffic, and traffic control factors significantly affected the accident occurrence. The total approach volumes, left-turn volumes, existence of unprotected left-turn slip roads, number of signal phases per cycle, use of permissive right-turning phase, and sites with sight distances less than 100 m or greater than 300 m appear to increase accident occurrence. However, the presence of right-turn channelization, provision of an acceleration section for left turning, existence of a surveillance camera, availability of median railings, and presence of an approach gradient greater than +5% may reduce the occurrence of intersection approach accidents. Moreover, the study shows that the reporting rate may drop because current law enforcement requires that only injury accidents and accidents at intersections in a residential area be reported.

Singapore is a heavily motorized country with one of the most advanced transportation systems in Asia. Although the level of road safety in Singapore is good by Asian standards, the accident rates are still unacceptable. Annual accident statistics show that more than 30% of the crashes in Singapore occur at signalized intersections, making these one of the most hazardous location types in the road network. Several major factors influence accident frequency at intersections, including traffic characteristics, traffic control measures, and geometric design elements at the site. A common approach to understanding the effects of these factors in accident causation is to develop an accident prediction model, that is, to establish a relationship between accident frequency and the geometric and traffic characteristics of the intersection. Although these models have been developed in many countries (*1–3*), few studies involve such models in Singapore. Although a reliable data system is essential to develop a good representative model, one particular problem in accident studies is that of underreporting, which may hide true information about accidents.

According to Singapore law, motorists involved in noninjury traffic accidents are not required to lodge any police report but may make reports with the General Insurance Association, which deals with accident claims. However, traffic accident reports must be lodged with traffic police when there is injury, or when the accident involves a government vehicle, a foreign vehicle, or a pedestrian or cyclist, or when there is damage to government property. An injury accident is generally defined as one in which (*a*) at least one party involved in the accident is conveyed to a hospital from the accident scene by an ambulance, (*b*) any party conveyed to a hospital by other means becomes hospitalized or is given outpatient medical leave of 3 or more days, or (*c*) no one is conveyed to a hospital from the scene but a latent injury developed subsequently in any involved party entails hospitalization or outpatient medical leave of 3 or more days.

Even if injury accidents are reported according to the law, many injury accidents may not be reported, especially slight-injury accidents. This appears to be true in many countries (*4–6*). Rosman has explained that in Western Australia, about 40% to 45% of hospital admission records of road crashes did not have a corresponding police report because of underreporting (*5*). The amount of underreporting is not known in Singapore, but it may be significant. Therefore, the available accident data for analysis may not reflect the complete picture. Hence in studying accident occurrence, the adopted statistical methodology must account for such underreporting effects. This study develops a suitable model to describe crashes at three-legged signalized intersection approaches in Singapore, taking into account the effect of underreporting.

METHODOLOGY

A relationship between recorded accidents and possible explanatory variables can be made through count data models such as Poisson regression. But with the prevalence of misspecifications, such as underreporting, it is necessary to modify the parent Poisson regression model to improve the quality of parameter estimation. A misspecification is a violation of assumptions in parent Poisson model or specific problems related to data. It is necessary to identify and treat these misspecifications to improve the quality of parameter estimation. These misspecifications affect not just one assumption but many at a time. Possible countermeasures for the presence of underreporting are discussed in the following.

S. S. P. Kumara, Transportation Research Group, School of Civil Engineering and the Environment, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom. H. C. Chin, Department of Civil Engineering, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260.

Transportation Research Record: Journal of the Transportation Research Board, No. 1908, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 46–50.

For model formulation, Winkelmann first modified the parent Poisson mean function with the addition of unobserved heterogeneity (*7*). Heterogeneity refers to the presence of site-specific but unobserved factors, such as pavement surface effects, including rutting, skid resistance, and type of surface material, and driver population effect and age distribution. Hence the conditional mean number of actual accidents (y_i^*) at an intersection approach i following the modified Poisson model will be

$$E\left(y_i^* \middle| x_i \epsilon_i\right) = \exp(x_i\beta + \epsilon_i) \tag{1}$$

where

β = vector representing parameters to be estimated,
x_i = vector representing the explanatory variables, and
ϵ_i = unobserved heterogeneity normally distributed across the individual data with mean zero and standard deviation σ^2.

As suggested by Winkelmann (*7*), the reporting mechanism may be given as

$$c_i = \gamma z_i + u_i^r \tag{2}$$

where

γ = vector representing parameters to be estimated,
z_i = vector representing explanatory variables related to the reporting mechanism, and
u_i^r = disturbance term.

If C_i^* denotes the event that an event that happens is actually recorded, then assuming that an accident j is reported and hence $C_i^* = 1$ if the net utility of doing so is positive, that is, $c_i = \gamma z_i + u_i^r > 0$. Therefore, $C_i^* = 1$ if and only if $C_i > 0$. Hence the number of reported counts is given by

$$y_i = \sum_{j=1}^{y_i^*} C_i^* \tag{3}$$

If $y_i^*|\epsilon_i$ is Poisson distributed, then $y_i|\epsilon_i$ is also Poisson distributed with a modified mean, and the final version of the modified Poisson mean for reported accidents combining Equations 1 and 2 is given by

$$\tilde{\lambda}_i = \exp(x_i\beta + \epsilon_i) \times \Phi\left(\frac{z_i\gamma}{\sqrt{1-\rho^2}} + \frac{\rho\epsilon_i/\sigma}{\sqrt{1-\rho^2}}\right) \tag{4}$$

where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution and ρ is the correlation between ϵ_i and u_i. The underreporting may be either endogenous where the correlation between variables relating accident occurrence and reporting is possible or exogenous otherwise. Hence the insignificant ρ indicates that the reporting state is independent from the accident occurrence state. Winkelmann derived the probit for the endogenous underreporting model as (*7*)

$$g\left(y_i \middle| x_i, z_i; \beta, \gamma, \rho, \sigma\right) = \int_{-\infty}^{\infty} \exp\left[-\exp(x_i\beta + \epsilon_i)\Phi\left(\frac{z_i\gamma + \rho\epsilon_i/\sigma}{\sqrt{1-\rho^2}}\right)\right] \times \left[\exp(x_i\beta + \epsilon_i)\Phi\left(\frac{z_i\gamma + \rho\epsilon_i/\sigma}{\sqrt{1-\rho^2}}\right)\right]^{y_i} \times \frac{1}{y_i!\sigma}\Phi(\epsilon_i/\sigma)\,d\epsilon_i \tag{5}$$

The parameters of this model are usually estimated by using the maximum likelihood method.

The exogenous underreporting model can be determined by forcing the correlation coefficient ρ in the endogenous model to zero. This produces a Poisson regression probit for the exogenous underreporting model:

$$g\left(y_i \middle| x_i, z_i; \beta, \gamma, \rho, \sigma\right) = \int_{-\infty}^{\infty} \exp\left[-\exp(x_i\beta + \epsilon_i)\Phi(z_i\gamma)\right] \times \left[\exp(x_i\beta + \epsilon_i)\Phi(z_i\gamma)\right]^{y_i} \times \frac{1}{y_i!\sigma}\Phi(\epsilon_i/\sigma)\,d\epsilon_i \tag{6}$$

The parameters of the exogenous model also can be estimated by the maximum likelihood method.

The suitability of the model depends on the distribution of available data. Vuong proposed a test statistic for nonnested models (*8*), which appears to have some power to distinguish between the parent Poisson model and the modified Poisson underreporting model. The Vuong statistic (V) is distributed as standard normal distribution, and hence its value may be compared to the critical value for standard normal distribution. Therefore it may be taken that $V > 1.96$ distinctly favors the Poisson underreporting models, whereas $V \le 1.96$ distinctly favors the parent Poisson model; otherwise the test is indecisive. A decision rule that may be used in selecting the appropriate model, which is based on the Vuong statistic V and t-statistic of the parameter ρ of the endogenous Poisson underreporting model, is shown in Table 1. Literature that uses the underreporting models is limited but not rare. Winkelmann and Zimmermann applied a similar modeling approach to data on labor mobility (*9*).

DATA

To develop a mathematical model that correlates accident frequencies to causal factors, one needs to select a representative set of intersections that possess a wide variety of geometric and traffic characteristics. In Singapore, a large number of these intersections are in city areas and are characterized by high traffic volumes. A total of 104 three-legged signalized intersections were selected for the model development, and these represent about 40% of such intersections in Singapore. The resulting data give a well-defined sample of city intersections to study, and all street classifications, including principal, minor, collector arterials, and local streets are represented. Each intersection was divided into separate approaches, and accident and

TABLE 1 Decision Rule for Model Selection

	t-Statistic of ρ	
Vuong Statistic	<2	>2
V≤1.96	Parent Poisson or NB model	Parent Poisson or NB model
V>1.96	Exogenous Poisson underreporting model	Endogenous Poisson underreporting model

NB = negative binomial.

other data were taken at each approach for 1-year intervals for 9 years, from 1992 to 2000. A total of 2,808 observations were provided by the 312 intersection approaches for 9 years.

For the period from 1992 to 2000, three-legged intersections contributed about 15.7% of total reported road accidents in Singapore. To identify the accidents related to the considered intersection approaches, it was necessary to define the junction area. The junction area is normally defined as the area within 100 m of all approaches from the center of the junction. The annual accident data from 1992 to 2000 were extracted from National Road Accident Data Base.

Accident models are said to be more representative if there exists a closer relationship between accidents and exposure (*2*). As the yearly accident data are considered, it is necessary to obtain annual average daily traffic (AADT) data for intersection approaches. However, as such data are not available, the average daily traffic (ADT) data were assumed to be the same as AADT. These ADT values were obtained from Land Transport Authority (LTA) Singapore. Finally the annual number of vehicles entering the approach were estimated by modifying the ADT values according to the number of registered vehicles in that year with reference to the ADT values available for 2000 following Kulmala (*2*) and Poch and Mannering (*3*).

Information relating to traffic control measures and geometric design elements at the sites obtained either from LTA records or through site measurements. Table 2 presents summary statistics of the variables considered for these intersection approaches.

TABLE 2 Summary Statistics of Variables for Three-legged Signalized Intersection Approaches

Variables	Min.	Max.	Mean	SD
Annual accident frequency	0	6	0.2896	0.6933
Traffic data in AADT				
Total approach volume	1093	66,378	12,922.3	9,497.2
Total left turn volume	0	21,312	2,742.0	3,713.2
Right turn volume	0	26,514	2,068.36	3,222.1
Geometric data				
Sight distance (m)	49.15	515.63	244.11	154.26
Length of left turn slip road (m)	0	103.46	8.6615	15.77
Number of approach lanes	1	6	2.8109	0.9506
Median width (m)	0	29	1.9540	2.7813
Existence of +5% gradient*	0	1	0.2051	0.4039
Horizontal curve*	0	1	0.3590	0.4798
Right turn channelization*	0	1	0.3237	0.4690
Left turn channelization*	0	1	0.2058	0.4043
Left turn slip road*	0	1	0.2981	0.4575
Acceleration section on* left turn slip road	0	1	0.0128	0.1125
Other variables				
Number of signal phases	2	4	2.9423	0.3348
Surveillance camera*	0	1	0.1154	0.3195
Median railings*	0	1	0.2788	0.4485
Permissive right turning phase*	0	1	0.7356	0.4411
Law enforcement*	0	1	0.2895	0.4158
Residential intersection*	0	1	0.2596	0.4385

*Indicator variables

RESULTS AND DISCUSSION

The use of repeated annual data for each intersection approach may create serial correlation issues because of the use of annual expansion factors to estimate traffic volumes and so forth. However, the Poisson underreporting model assumes that annual accident frequencies are independent observations. To test for possible correlation problems from one year to the next, the accident data were first split into several sets, and a comparison of coefficient estimates by using the likelihood ratio test was conducted, as discussed by Poch and Mannering (*3*). No significant differences in coefficients were observed. Thus it can be concluded that serial correlations from year to year do not significantly affect the model results.

The endogenous underreporting model is the first natural selection. The Akaike information criterion is used to select the most suitable model among a set of competing models (*10*). The endogenous model gives an insignificant correlation coefficient ρ ($p = 0.9991$, $t = 0.001$), which means that the unobserved approach specific effects are not correlated with the specific disturbances of underreporting. Therefore, reporting may be exogenous. On the basis of the Vuong statistic (*8*), the exogenous probit underreporting model is a better representative model to its parent Poisson model ($V = 5.73 > 1.96$) following the criteria given in Table 1. The significant factors and the summary statistics of the model are shown in Table 3.

As shown in Table 3, 11 traffic and geometric variables are found to be significant. Accident occurrence is significantly affected by the total approach volume ($t = 6.610$, $p < 0.0001$) and left-turn volume ($t = 2.392$, $p = 0.0168$). In Singapore, drivers drive on the left side of the road. Increased volume implies greater interaction between vehicles and possibly more conflicts. Accidents will therefore increase because of greater exposure. Furthermore, as volume increases, there may be fewer gaps in the traffic for turning and merging maneuvers.

The model also indicates that the provision of right-turn channelization improves safety ($p < 0.0001$). This exclusive lane lowers the differential in speeds among vehicles on the approach and hence possible reduction in conflicts. In their study of intersection accident frequencies, Poch and Mannering concluded that without an exclusive lane, right turning vehicles are required to slow in a lane shared with speed-maintaining vehicles may cause speed differentials that tend to cause accidents (*3*).

The model results also show that the unprotected right-turn phase may increase accident occurrence ($p < 0.0001$). This unprotected phase allows the right-turning vehicles to proceed with the turn when there is a significant gap in the opposing traffic stream. The need to make a judgment adds to a driver's workload and may explain why the risk is higher.

Surprisingly, the model indicates that approaches on gradients with +5% or more reduce the accident occurrence by about 46% ($p = 0.0004$), all other factors being constant. Almost all the approach gradients are less than 8% at the intersections considered. Higher grades may give rise to a safer environment because of reduced speed.

According to the model, the uncontrolled left-turn slip road increases accident occurrence ($p = 0.1375$). The uncontrolled lane, which allows left-turning vehicles to merge into the cross-traffic

TABLE 3 Estimation Results for Exogenous Probit Underreporting Model

Independent Variable	Estimated Coefficient β	*p*-Value
Parameters of Poisson model		
Constant	−8.3599	<0.0001
Total approach volume in thousand AADT	0.6310	<0.0001
Left turn volume in thousand AADT	0.1843	0.0168
Sight distance 1 (1 if <100 m, 0 otherwise)	0.1303	0.3714
Sight distance 2 (1 if >300 m, 0 otherwise)	0.1974	0.0799
Right turn channelization (1 if yes, 0 otherwise)	−0.4983	<0.0001
Left turn slip road (1 if yes, 0 otherwise)	0.1837	0.1375
Acceleration section on left turn slip road (1 if yes, 0 otherwise)	−0.3695	0.0112
Surveillance camera (1 if yes, 0 otherwise)	−0.1897	0.2112
Median railings (1 if yes, 0 otherwise)	−0.1466	0.1884
Permissive right turn phase (1 if yes, 0 otherwise)	0.4985	<0.0001
Gradient greater than +5% (1 if yes, 0 otherwise)	−0.4642	0.0004
Parameters of probit reporting model		
Law enforcement that only injury accidents are to be reported (1 if yes, 0 otherwise)	−0.2135	0.1619
Residential intersection (1 if yes, 0 otherwise)	−0.3282	0.0148
σ	1.0825	<0.0001
Vuong statistic	5.72	
Log likelihood at convergence	−1811.75	
AIC	−3597.50	
ρ^2	0.12	

AIC = Akaike information criterion.

stream, increases the likelihood of accidents. However, by providing an acceleration section in the left-turn lane, drivers may be able to merge more easily. This explains a reduction in accidents of about 37% ($p = 0.0112$) when such lanes are provided, everything else being constant.

Surprisingly, the model indicates that a sight distance greater than 300 m is associated with higher accidents ($p = 0.0799$). When the sight distance is large, drivers with greater freedom to maneuver will drive at higher speeds, resulting in an increase in risk of perception and judgment errors because of premature observation making (*2*). Moreover, accident frequency appears to increase by about 13% when the sight distance is smaller than 100 m. This should be considered as a general trend variable as the significance level is compared less to other variables. Short sight distances put drivers in a critical situation and reduce a driver's time to judge traffic and operational conditions at the intersection. In a study on safety at rural three- and four-arm junctions, Kulmala also concluded that the short sight distances may increase accident occurrence (*2*).

Median railings are installed to prevent pedestrians from crossing the road away from designated pedestrian crossings. From the model, the presence of median railings appears to reduce accident occurrence ($p = 0.1884$). This should also be considered as a general trend variable as the significance level is compared less to other variables. The examination of accident records revealed that there are fewer pedestrian-related accidents at intersections at which there are median railings.

The existence of surveillance camera reduces the accident occurrence ($p = 0.2112$). This should also be considered as a general trend variable. In this instance, the camera appears to have improved safety because drivers tend to be more cautious, as observed by Retting et al. (*11*).

The model indicates a high tendency of underreporting ($p = 0.0148$) when the intersection is in a local or residential area in Singapore. This implies that the reporting rates are also known to vary by geographic location of the intersection. The high reporting rate in urban areas may be due to the greater population size. Alsop and Langley also indicated that the reporting rate is high in urban areas, where traffic densities are generally higher (*4*).

The model illustrates that law enforcement has some effect on accident reporting ($p = 0.1619$). This should be considered as a general trend variable, but it was observed that overall it increases the total quality of the model. The legal requirement may hide most slight-injury and especially property-damage accidents and their valuable details. Slight-injury accidents might also be regarded as property-damage accidents, and these may not be reported. However the more severe the injury, the greater the likelihood of reporting as suggested by Alsop and Langley (*4*). As a solution to these problems, it is important at the least that reports are made to insurance companies for property-damage accidents. This requires that traffic police records be combined with insurance company records to achieve a reliable data source for property-damage accidents. Further, it is important to combine police-recorded accident data with hospital data to create an accurate and more representative database for injury accidents.

CONCLUSION

Underreporting is a major problem in pursuing accident analysis research because accident prediction models mainly depend on the recorded accidents and their recorded details of the causes. This study used Poisson underreporting models to analyze causal factors affecting intersection approach accident frequencies when there is a reason to believe that some accidents that have occurred may not have been recorded. The parameters identified in the model are more accurate and useful, hence proper measures can be taken to improve both safety and reporting. The model shows that total approach volumes, left-turn volumes, existence of unprotected left-turn slip roads, number of signal phases per cycle, use of permissive right-turning phase, and sites with sight distances less than 100 m or greater than 300 m appear to increase accident occurrence. However, the presence of right-turn channelization, provision of an acceleration section for left turning, existence of a surveillance camera, availability of median railings, and presence of an approach gradient greater than +5% may reduce the occurrence of intersection approach accidents. According to the model, underreporting in Singapore appears to be affected by the legal requirement that only injury accidents are to be reported and the geographic location of the intersection.

REFERENCES

1. Pickering, D., R. D. Hall, and M. Grimmer. *Accidents at Rural T-junctions.* Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1986.
2. Kulmala, R. *Safety at Rural Three- and Four-Arm Junctions: Development and Application of Accident Prediction Models.* Julkaisija-Utgivare, Finland, 1995.
3. Poch, M., and F. Mannering. Negative Binomial Analysis of Intersection Accident Frequencies. *Journal of Transportation Engineering*, Vol. 122, No. 2, 1996, pp. 105–113.
4. Alsop, J., and J. Langley. Underreporting of Motor Vehicle Traffic Crash Victims in New Zealand. *Accident Analysis and Prevention*, Vol. 33, No. 3, 2001, pp. 353–359.
5. Rosman, D. L. The Western Australian Road Injury Database (1987–1996): Ten Years of Linked Police, Hospital and Death Records of Road Crashes and Injuries. *Accident Analysis and Prevention*, Vol. 33, No. 1, 2001, pp. 81–88.
6. Ibrahim, K., and D. T. Silcock. The Accuracy of Accident Data. *Traffic Engineering and Control*, Vol. 27, 1992, pp. 492–497.
7. Winkelmann, R. *Count Data Models: Econometric Theory and Application to Labour Mobility*. Springer-Verlag, Berlin, 1997.
8. Vuong, Q. Likelihood Ratio Tests for Model Selection and Non-Nested Hypothesis. *Econometrica*, Vol. 57, 1989, pp. 307–334.
9. Winkelmann, R., and K. F. Zimmermann. *Poisson Logistic Regression*. Department of Economics, University of Munich, Munich, Germany, 1993.
10. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. In *Second International Symposium on Information Theory* (B. V. Petrov and B. F. Csaki, eds.), Academiai Kiado, Budapest, Hungary, 1973, pp. 276–281.
11. Retting, R. A., A. F. Williams, C. M. Farmer, and A. F. Feldman. Evaluation of Red Light Camera Enforcement in Oxnard, California. *Accident Analysis and Prevention*, Vol. 31, No. 2, 1999, pp. 169–174.

The Safety Data, Analysis, and Evaluation Committee sponsored publication of this paper.