Using Logit modelling to measure the safety of New Zealand drivers

Ron Veltman M.Ag.Econ Transport Statistician

New Zealand Transport Agency

Ron.Veltman@nzta.govt.nz

Abstract

Measuring Safety is akin to quantifying risk. A logit model has been developed that determines the percentage probability of a driver being involved in a crash incident in the following year, conditional upon certain driver characteristics and driving behaviour in the current year.

The logit model was regressed using the total New Zealand driver population, and is applied to each and every individual current New Zealand driving license holder. The model is designed to be applied annually, and a five class histogram of driver probabilities is derived from the results.

The probability risk score is conditional upon the drivers' gender, age, driving experience and traffic offending.

The "most at Risk" class contains those drivers whose probability scores exceed 0.7632%. That driver group is categorized as being "Recidivist" and is analysed as such to determine common characteristics that can be utilized to inform policy.

In applying the model to drivers and tracking the resulting scores to "at fault" drivers involved in serious crashes in the following year, a robust "predictive" validation of the model was discovered. Crashes in the main involved both band 1 and band 2 "recidivist" drivers already identified in the year prior. Analysis of the recidivist group reveals that the predominant problem cohort is young males aged generally between 16 and 24 years of age, and speed is the predominant behaviour.

Key words: Driver License Register, Basic driver characteristics, Traffic offending.

1. Introduction

A long held ambition of the New Zealand Transport Agency has been to objectively identify individually the least safe drivers amongst the population of New Zealand driver license holders. The objective behind such an ambition is to isolate, analyse and profile the resulting "recidivist" group of drivers in such a way that the design of a range of strategies for reducing crash incidents can include provision for specifically targeting recidivist driving behaviour.

In contemplating the technical challenge of quantifying the safety, or otherwise, of individual drivers, then given that by definition the concept of safety is synonymous to risk, and risk is fundamentally composed of probability and consequence, application of a simple logit model immediately comes to mind as an obvious method for measuring, and therefore ranking, all New Zealand drivers in terms of safety.

To coin a phrase "safety is no accident" a logit model offers us a conditional probably of a driver being involved in a consequential crash incident or accident, conditional upon certain characteristics pertaining to that driver, and that driver's behaviour.

Logit modelling has traditionally been used in the transport sector for predominantly either estimating public transport patronage (Martin and Rutherford 2003), or for identifying the safety characteristics of certain transport system infrastructure designs (Polus, Shiftan, Lazar 2005). With regard to safety, generally research using logit modelling has concentrated on crash severity (Milton, Shankar, and Mannering 2008).

It seems, rightly or wrongly, that the majority of research into Australasian transport safety has traditionally concentrated on crash incidents, in terms of causation due to road design or the lack of vehicle technology (Watson and Newstead 2009). One could conclude that the focus has been more on the consequence side of the risk ledger in a reactionary way, rather than on the probability side in a proactive way, in our efforts to improve transport safety.

2. Logit model construct

2.1. Logit model form

Introduced by Berkson in 1944, a variation of the Probit or Normit model (Bliss 1934), the Logit model developed takes the following very simple form:

$$P_i = E(Y = 1 | X_i) = \frac{1}{1 + e^{-L_i}} \quad (1)$$

Where P_i is the conditional probability (log odds ratio), and

Where the Logit term $$L_i = \alpha_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} \quad (2)$$

And where X_{1i} = Driver gender (Male, Female)

X_{2i} = Driver age (whole years)

X_{3i} = Driver experience (whole years license held)

X_{4i} = Driver traffic offences (derived score in whole numbers)

Alpha and beta coefficients are derived using iterative Maximum Likelihood Estimation.

Of note is the way in which Logit modelling utilises a combination of ordinal and stochastic exogenous variables, with X_{1i} being the only ordinal variable.

Initial iterations included X_{5i} = Driver crash history (derived score in whole numbers) which was later removed due to insignificance.

2.2. Data Sets Utilised

The base data set utilises the full New Zealand driver register. All licenses issued up to and including 31 December 2007, less those belonging to all license holders that had died up to and including 31 December 2007, were used in the model. After the removal of duplicate license identifiers, due in the main to change of name, and the removal of foreign or “pseudo” licenses, the population of New Zealand drivers modelled was 3,475,986.

The population included any license holder not residing in New Zealand during 2007, and ignored the possible surrendering of a New Zealand license for a foreign replacement.

Two further secondary sets of data were used: All traffic offence charges for the calendar year ended 31 December 2007, including speed camera offences, and all reported crash incidents with driver identifiers for the year 2008. Offences and crashes involving non New Zealand License holders were removed.

Linkage between the datasets was via driver license identification number.

Sufficient time had elapsed for all data to be complete, meaning any potential data input lag was not an issue in preparing all data sets.

2.3. Variable construction

Looking at the logit function (2), one might immediately notice the absence of any exposure variable, and logically that variable would be driven distance in kilometres. Such a variable would be entered as an additional independent variable, rather than maybe expressing offences according to a ratio of distance (offence rate per 1000 km driven).

There is much difficulty in deriving vehicle distance as it is, and an impossibility to derive driver distance. Where vehicle distance can be derived, and linked to ownership, it is clear that a driver could drive several vehicles during a year, as could a vehicle be driven by several drivers, both privately and commercially. It is also reasonable to expect most households have access to more than one vehicle, and there is also the issue of vehicle sales or change of ownership to contend with.

Driver experience, at a stretch, could be considered somewhat as a proxy exposure variable and this is why it is included. There is a presumption in the function that the more years a driver has a license, then the more likely is that driver to drive regularly. Exceptions would be a generation of middle aged or retired female license holders who are regularly driven by their spouses. The accumulation of distance driven over time, if driven safely, might be considered no riskier in comparison to someone who holds a license but never drives.

Given the objective behind this exercise, that being a ranking of drivers, then the issue regarding exposure might not be as crucial as one would expect.

2.3.1. Dichotomous dependant variable

Each and every driver involved in a reported crash incident during the year 2008 was coded 1 otherwise 0.

On average 11% of crash incidents are indeterminate by the police in establishing the driver at fault. Those considered possible victims of an incident might have had a lack of experience, or a lack of perception regarding a potential crash, and consequentially were unable to defensively avoid the crash. In order to maximise the number of records coded 1, mere involvement was therefore used as the coding criteria.

Of the 3,475,986 drivers in the data set, 13,627 (0.4%) were identified as being involved in a crash incident during 2008. In comparison to the driver data set, this is a very low number. With only 2008 crash incidents used alone, there is a consequential potential for convergence in parameter estimation to not occur. A second dependant variable constructed with using both 2008 and 2009 crashes was built. With the inclusion of 2009, of the 3,475,986 drivers, 25,023 (0.7%) were coded 1 as being involved 0 otherwise. It was found that this variable was no different from the 2008 only crash variable and it was abandoned.

An issue to be contemplated with assignation of driver crash involvements is that the data is reliant on “reported” crashes only. The potential for estimation bias without “unreported” crashes being identified (Ye and Lord 2010) is therefore present. Whilst data exists regarding the incidence of non-reporting, via hospitalised casualties, it is not clear how such types of data might be included within this model construct.

2.3.2. Driver gender

Of the 3,475,986 drivers in the data set, 53.78% (1,869,391) were Male (coded 1), with the remaining 46.22% (1,606,595) Female (coded 2).

Within the driver register, a small number of license holders are defined as gender indeterminate, and these were recoded as Male. There are also a number of “gender changed” records, and the female record was accordingly selected.

2.3.3. Driver age

Driver age involved simply subtracting Date of Birth from 31 December 2007, converting to years and rounding to a whole number. As is often the case with manually entered register data, several input errors were detected. Only the obvious ones could be corrected, as there is no way to validate any date correction. For example, several 14 years olds were found in the data set (impossible given license age was then 15), and these were "re-aged" as being 15. The incidence of date input error is unknown.

Figure 1: Histogram of Age

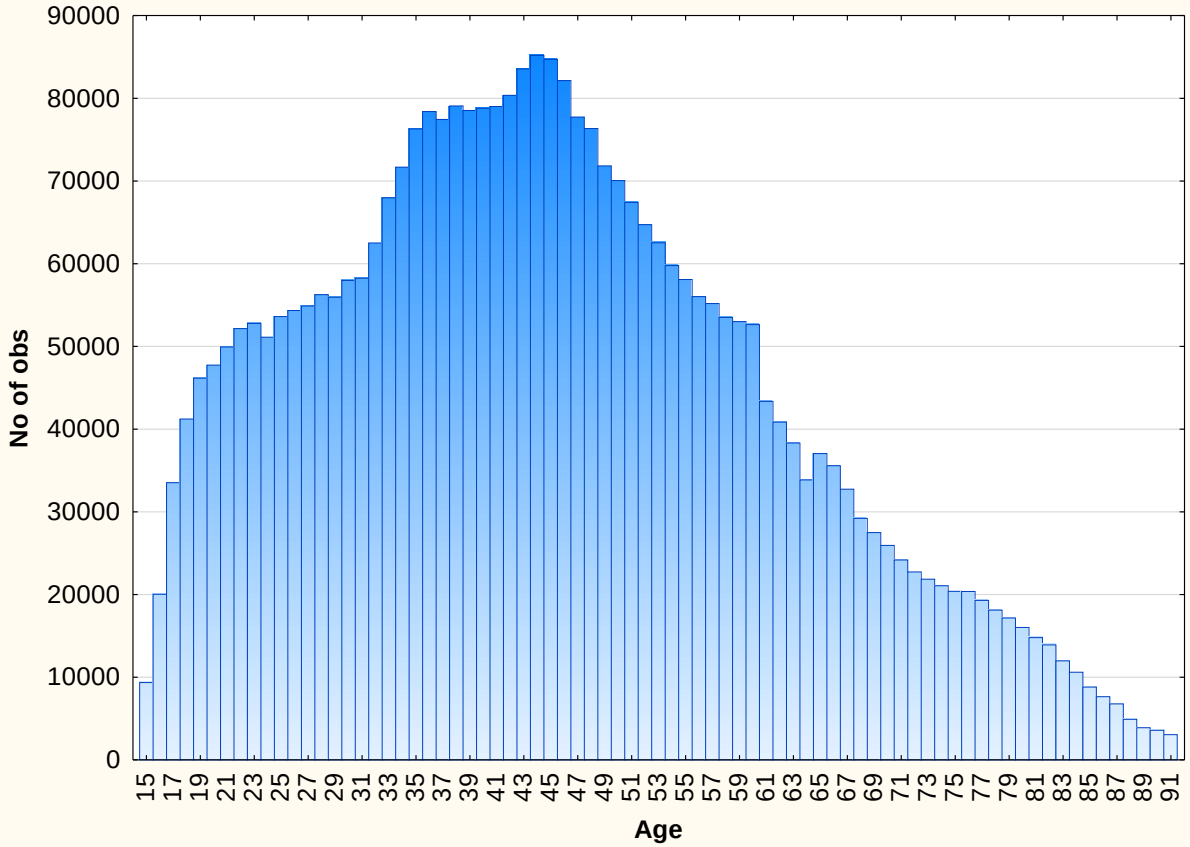


Table 1: Age Statistics

	N	Mean	Min	Max	STD
Age	3,475,986	45	15	91	16.6

2.3.4. Driver experience

Driver experience was calculated as License Issue Date subtracted from 31 December 2007, converted to years and rounded to whole years. Zero years indicates a full year is yet to be achieved.

License issue date is defined as the date at which the license holder gained their last license class, and is therefore not the full period a license has been held. This then defines experience as the number of years having past since the last official driver license training and testing was undertaken.

Actual experience could be a function of age. If both age and experience variables are correlated there is potential for multicollinearity to exist within the model specification. With the definition of experience being last class (rather than first class), there is a direct correlation (0.79) between Age and Experience. Multicolinearity is therefore suspected to be present.

A combinatorial variable was considered, but the result would have been difficult to interpret. In the absence of any other available proxy exposure variable, it was decided to accept multicolinearity within the model.

Figure 2: Histogram of Experience

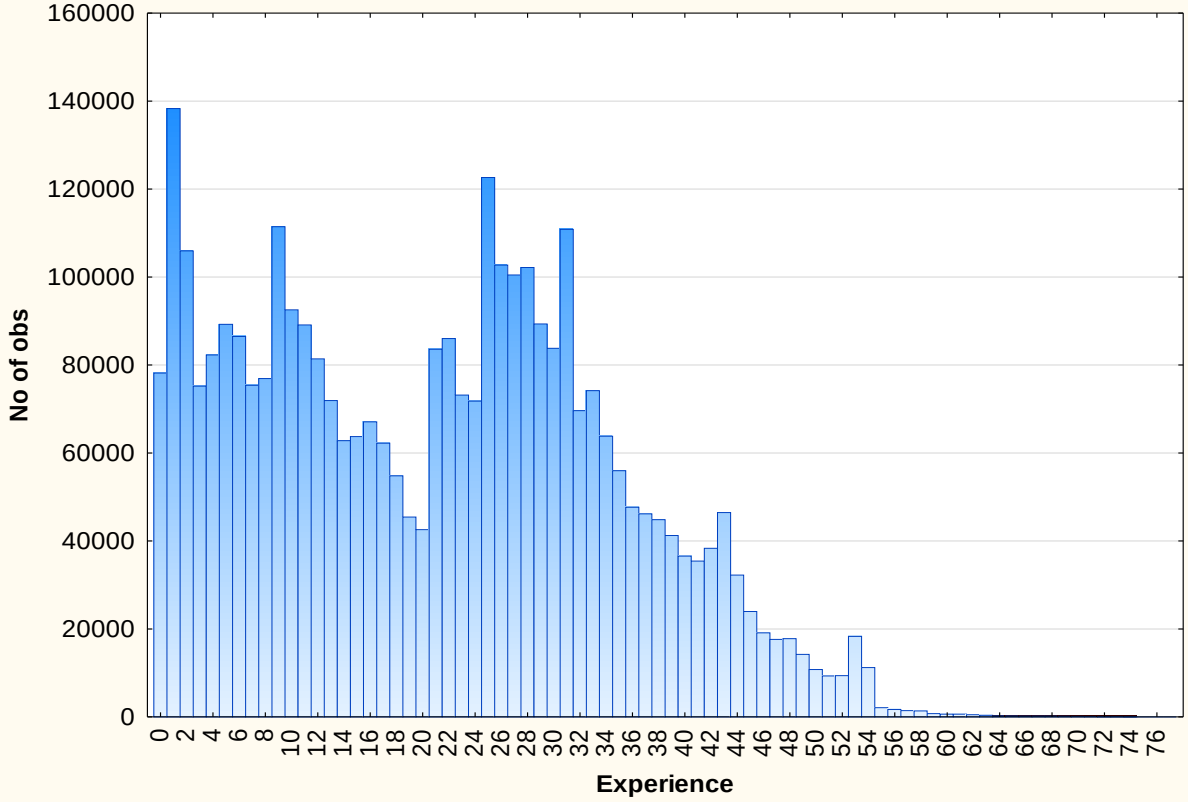


Table 2: Experience Statistics

	N	Mean	Min	Max	STD
Experience	3,475,986	21	0.00	77	13.67

2.3.5. Traffic offences

Each traffic offence has associated with it a precedent code which specifies exactly the nature and description of the charge. By the end of 2007 there were 2,731 such codes. Each code was assigned to one of 19 generic offence groups, and each group was given a ranking from 0 to 10 where 10 represented the most serious of safety related offences and 0 those offences considered to be irrelevant from a safety or risk point of view.

Every precedent code member within each generic group was also assigned a ranking from 0 to 10, denoting severity, and an offence score for each offence was calculated as the product of both group ranking and precedent code ranking.

Individual offence scores were then summed by driver license identifier, to provide an overall score per driver within the base dataset. Those drivers within the data set with a score of 0 had no offence charges during the year 2007.

The assignation of rankings to both offence groups and individual precedent codes was subjectively done. Whilst ranking would seem intuitive, it is likely that different individuals might assign different rankings.

Of note is that of the 3,475,986 driver licenses within the dataset, 17.11% (594,823) drivers were charged with one or more traffic offences during 2007.

Table 3: Offence Groups	
Offence	**Rank**
Drink/Drug Driving	10
Reckless Driving	10
Speed	8
Failure to Give way	7
Excess Load	6
Non Traffic Offence	5
General Behaviour	5
Dangerous Goods	4
No Driver License	3
Driver License Fraud	3
Driving Hours	3
False Driver License	3
Parking	2
Compliance	2
Nil WOF/COF	2
No MVL	0
Vehicle Issues	0
False Details	0
Nil RUC	0

3. Logit model results

Several runs to derive estimates were made. Every run achieved quick convergence with always between 7 and 10 iterations. As is usual nearly every run produced different estimates at convergence. The final results utilised were those that had been reproduced more than twice.

Model: Logistic regression (logit) Max likelihood (MS-err. scaled to 1)
Final loss: 4460.6956385 $Chi^2(4)=234.76$ p=0.0000

	Const.	**Gender**	**Age**	**Experience**	**Offences**
Estimate	-4.4256	-0.3677	-0.0134	-0.0091	0.0366
Standard Error	0.1650	0.0792	0.0042	0.0050	0.0031
t(183E3)	-26.8299	-4.6434	-3.1825	-1.8308	11.6291
p-level	**0.0000**	**0.0000**	**0.0015**	**0.0671**	**0.0000**
-95%CL	-4.7506	-0.5237	-0.0217	-0.0189	0.0304
+95%CL	-4.1006	-0.2117	-0.0051	0.0007	0.0428
Wald's Chi-square	719.8431	21.5609	10.1280	3.3517	135.2366
p-level	0.0000	0.0000	0.0015	0.0671	0.0000
Odds ratio (unit ch)	0.0120	0.6923	0.9867	0.9909	1.0373
-95%CL	0.0086	0.5923	0.9785	0.9813	1.0309
+95%CL	0.0166	0.8092	0.9949	1.0007	1.0438
Odds ratio (range)		0.6923	0.3556	0.5009	6327.0250
-95%CL		0.5923	0.1875	0.2380	1436.1200
+95%CL		0.8092	0.6745	1.0541	27874.5900

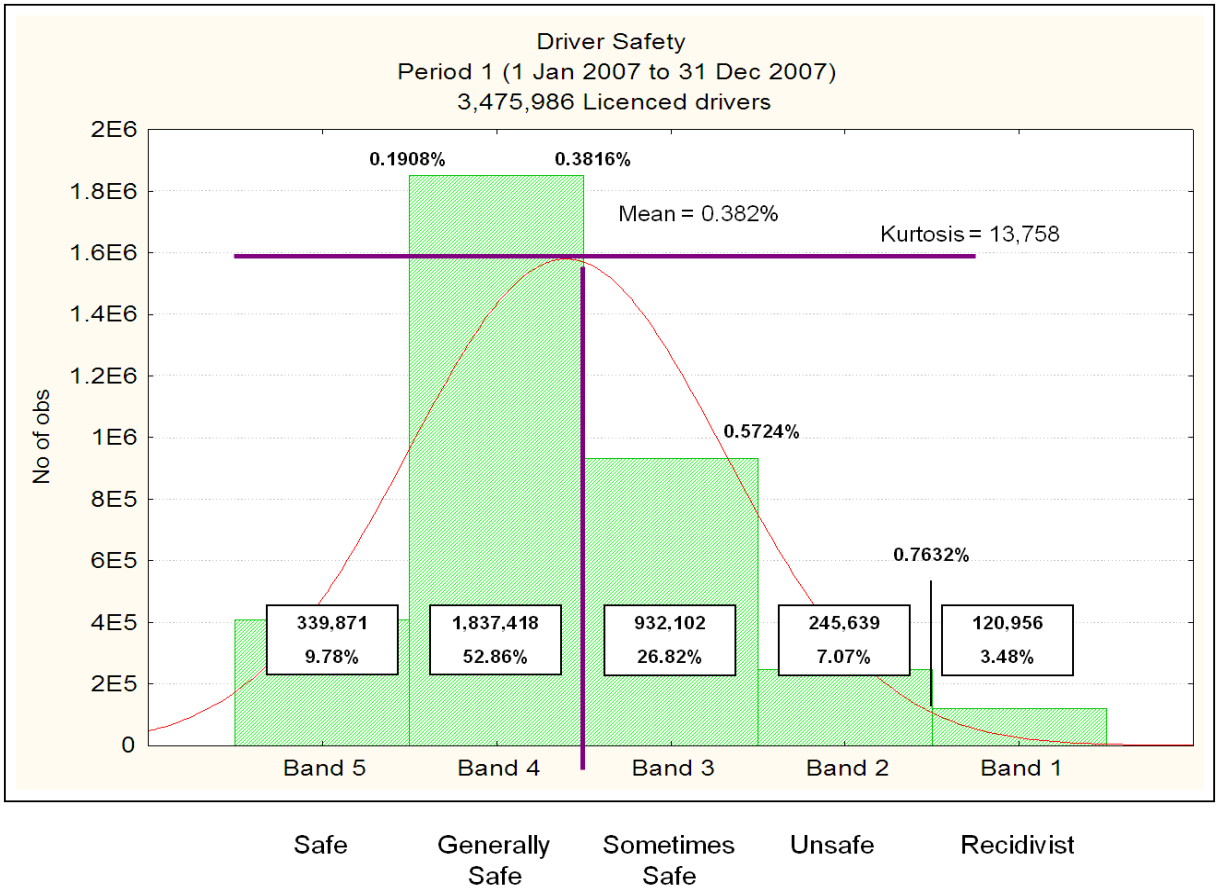
Table 4: Logit Model Results

All variables are highly significant with the exception of Driver Experience, which is significant at the 10% level. The confidence interval for "Experience" also includes 0 which is of some concern. These coefficients are the resulting population estimates, and they have been used within a fixed model context for application to all drivers within the register to produce probability scores.

4. Model utilisation

The logit model is applied to all New Zealand license holders within the register, and a histogram of the probability results produced. It was desirable to have no more than five classes, and band widths were derived by studying a frequency table of all probabilities. Band thresholds were selected in such a way that the resulting histogram has as close to a normal shape as was possible. Class widths are equidistant.

Figure 3: Histogram of Driver Probabilities



Keeping the logit model as fixed and keeping the band thresholds fixed, probability scores are run six monthly for the year ending June and December for each subsequent year. By monitoring Mean, Kurtosis and percent membership of the recidivist group over time, changes in these parameters provide a general indication of changes in the safety of New Zealand drivers. Naturally it is desirable to see decreases in the mean, increases in the kurtosis, and decreases in recidivist membership, or a "squeezing" of the histogram from right to left.

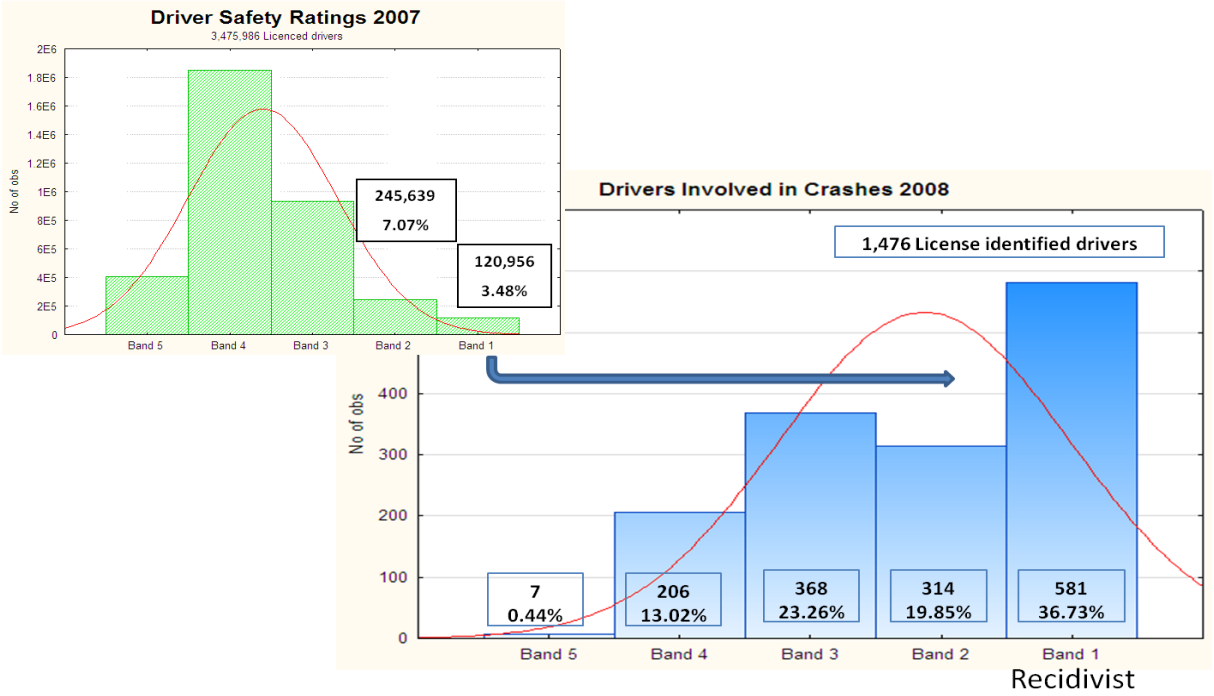
As time has progressed, a time series of probabilities for each and every license holder is developing. This presents more opportunity for future three dimensional analyses. Further, it might be possible to measure and value changes in safety in such a way that these values be correlated to the costs associated with probability mitigations, such that a cost benefit ratio of those mitigations results.

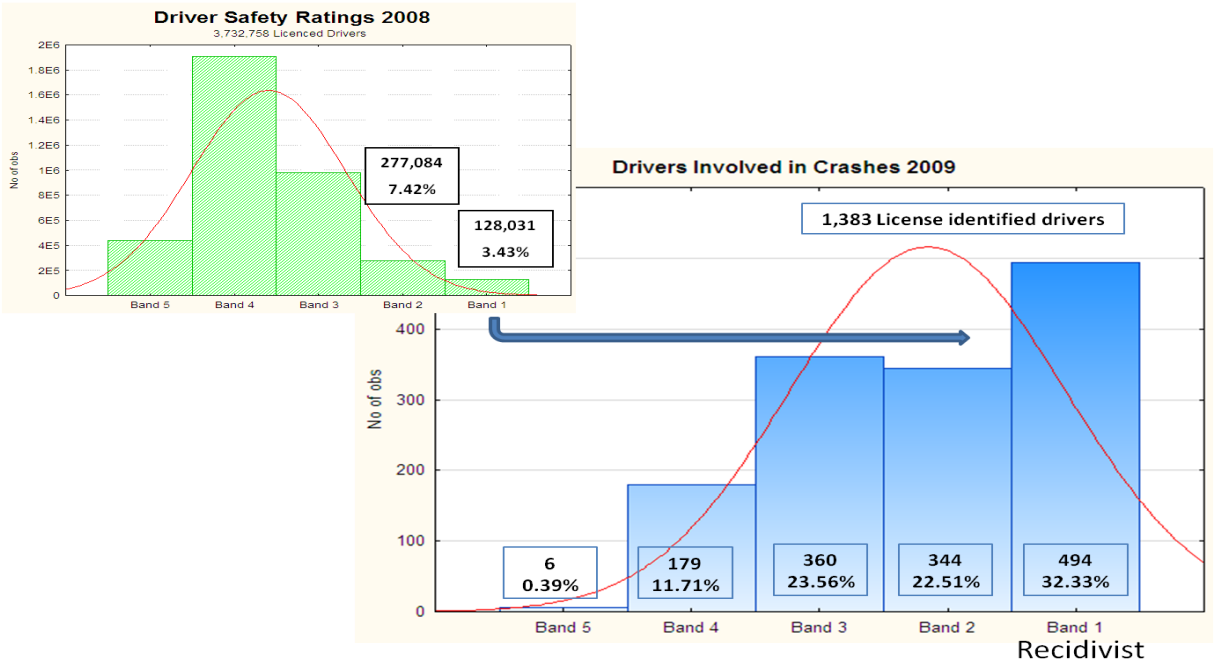
4.1. The predictive capability of the model

With the identification of the recidivist group as a specific target for risk mitigation strategies, then the question remains one of what possible impact might any range of strategies have on crash incidents should those strategies be implemented.

By identifying the preceding year probability scores of those “at fault” drivers involved in fatal or serious injury crash incidents during 2008 and 2009, we find that on average 33% of these serious incidents involved drivers that had in fact already been identified in the year prior, amongst the 3.5% of drivers categorised as being recidivist.

Figure 4: Recidivist impact on crashes





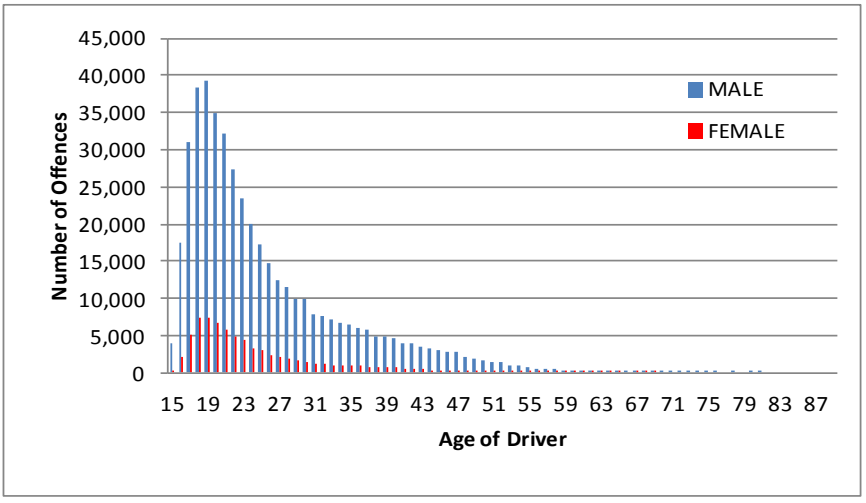
The implication is somewhat obvious. If we are able to address issues to do with the group of 120,000 recidivists, then we have a chance of reducing these serious crash incidents by a third. If we add in the band 2 drivers, deemed to be unsafe, then by specifically targeting some 360,000 of our high risk drivers, we could potentially eliminate half of our fatal or serious injury crash incidents.

Identifying this relatively small group of recidivist drivers gives us opportunity to develop and pursue proactive risk mitigation strategies that might not ordinarily be contemplated.

4.2. The typical recidivist driver

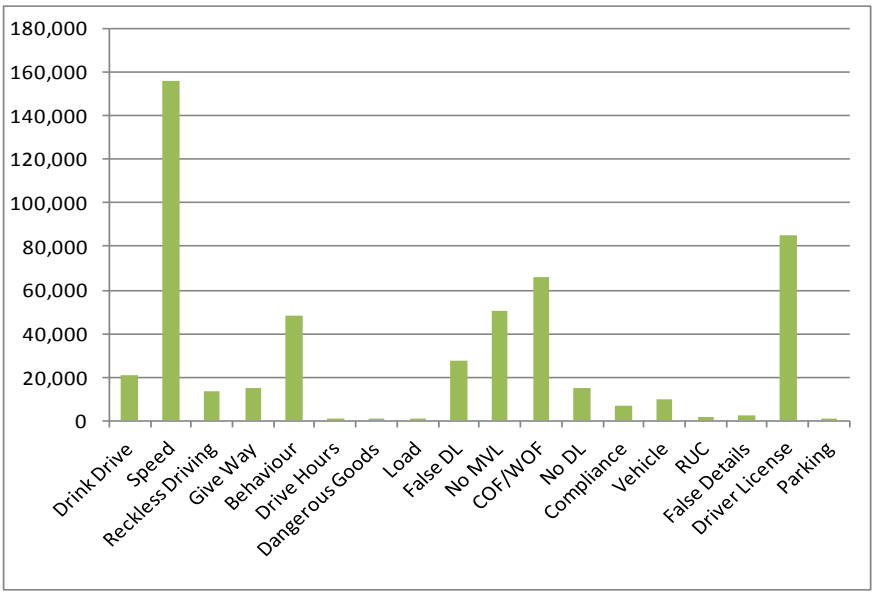
To be somewhat more current, recidivist drivers identified for the year ending 31 December 2009 show that by far the majority are Males aged mainly between 16 and 24. Whilst this age group is predominant, it also includes males ranging from 25 to 59 years of age and a few even older.

Figure 5: Age distribution of recidivist drivers (2009)



Typically the identified risky driving behaviour is speed, with also driver license issues and vehicle fitness issues also involved. There also exists an unhealthy behaviour problem, probably associated with their interaction with police.

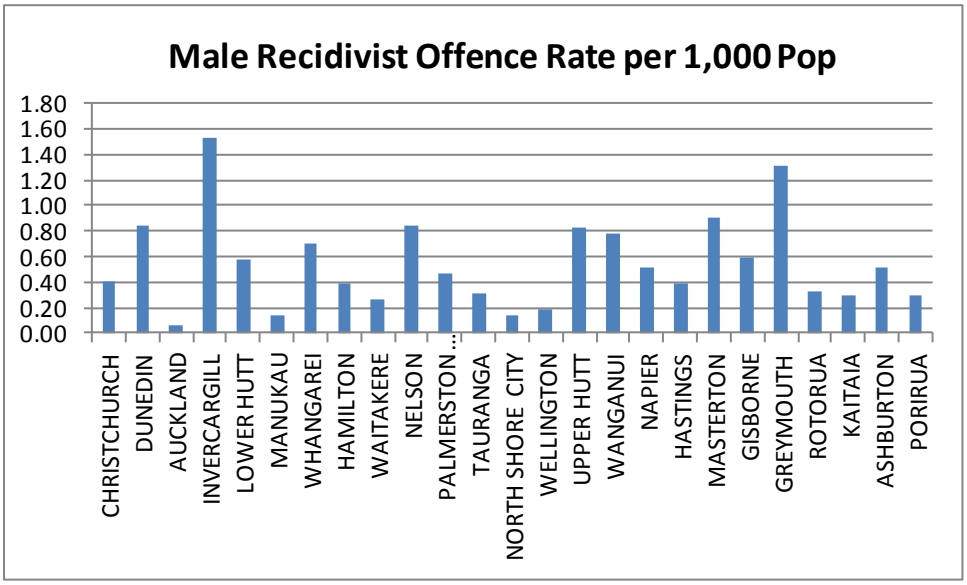
Figure 6: Recidivist Offence types (2009)



The combination of young males, speed, vehicle issues, license issues and attitude toward police, all might point to a "boy racer" problem, but clearly there is also a high number of what might be described as "those not involved in boy racing".

Of interest is where the recidivist driver is located. Whilst one would assume it is a main city problem, on a per capita basis, towns such as Invercargil, Greymouth, Masterton, Wanganui and Nelson have a disproportionate number of offending recidivist drivers.

Figure 6: Predominant Recidivist location (2009)



Conclusions

Whilst the logit model may lack a desired degree of academic robustness, it is a very helpful tool in at least ranking New Zealand drivers objectively. The simplicity of logit modelling, its ease in terms of interpretation and the fact that it enables consolidation of our entire driver related data into one system, has enabled the achievement of its initial objective; that being the identification of recidivist, or unsafe drivers.

By identifying at risk drivers specifically, the New Zealand Transport Agency now has available an opportunity to better target and pursue, in a proactive way, those drivers that are most likely to be involved in future crash incidents.

As a next step, a true understanding of the phycology underlying recidivist behaviour, in terms of their total lack of cognisance of risk, or their complete lack of risk averseness, will enable the design of several risk mitigations and controls that have a potential to half our crash incidents.

References

1. Fan Ye and Dominique Lord (2010) *Investigating the Effects of Underreporting of Crash Data on Three Commonly Used Traffic Crash Severity Models: Multinomial Logit, Ordered Probit and Mixed Logit Models* https://ceprofs.civil.tamu.edu/d**lord**/.../**Ye**_and_**Lord**_Cr... - United States

2. Tim Martin and Stephen Rutherford (2003) *Rail Demand forecasting in Hong Kong and Shenzhen*, *www.inrosoftware.com/en/pres_pap/asian/asi99/paper14.doc*

3. .Yannis and C.Antoniou, *A mixed logit model for the sensitivity analysis of Greek drivers' behaviour towards enforcement for road safety,* www.openstarts.units.it/dspace/bitstream/.../Yannis_Antoniu_ET37.p...

4. Abisai Polus, Yoram Shiftan and Sitvanit Shmueli-Lazar, *Evaluation of the waiting-time effect on critical gaps at roundabouts by a logit model,* www.ejtir.tbm.tudelft.nl/issues/2005_01/pdf/2005_01_01.pdf

5. JC Milton, VN shankar and FL Mannering (2008) *Highway accident severities and the mixed logit model: An exploratory empirical analysis,* http://www.scopus.com/record/display.url?eid=2-s2.0-38149090972&origin=inward&txGid=ahCr1uEnmxgzRcXflmQ-4r7%3a2

6. Linda Watson and Stuart Newstead (2009), *Vehicle Safety and Young Drivers*, Monash University Accident Research Centre.