

NZCASS 2014

Advanced Statistical Methods

Different types of analysis have been used to help us better understand different aspects of crime and victimisation. The three advanced statistical methods of modelling used as part of the NZCASS were:

- Multiple standardisation
- Logistic Regression
- The Gini coefficient

This document presents the aims, methods and results for these three methods.

Multiple Standardisation

The NZCASS main findings report and data tables show us that 32.9% of Māori had been victimised once or more in 2013, compared to 22.6% of the European population. This difference in victimisation is 10.3 percentage points.

Since Māori are over-represented in lower socio-economic groups and have a younger population, is victimisation really about being Māori or more to do with poverty or some other factor? This raises the question: If these factors (such as socio-economic status) were standardised across ethnic groups, are Māori still more at risk of being victimised?

Factor (variable) selection

There are a range of differences between the Māori and European populations. For example, a slightly higher proportion of Europeans live in major urban areas than Māori, but whether this is a strong difference relevant to victimisation needs to be tested to determine what factors (variables) to standardise by.

While many different factors could be considered, only 2 or 3 should be used due to sample size restrictions. This is supported by the Australian Institute of Health and Welfare (2011) principle that recommends a sample size of approximately 20–30 in each category for standardisation to be appropriate. This means that for the NZCASS sample size, standardisation should be conducted on approximately 2 or 3 variables (depending on category numbers).

Age and NZDep2013 quintiles were selected as the factors to standardise by. These 2 factors were selected because:

- User feedback from previous iterations of NZCASS analysis raised that socio-economic status is a 'confounding' reason for the higher rates of Māori victimisation. This means that when

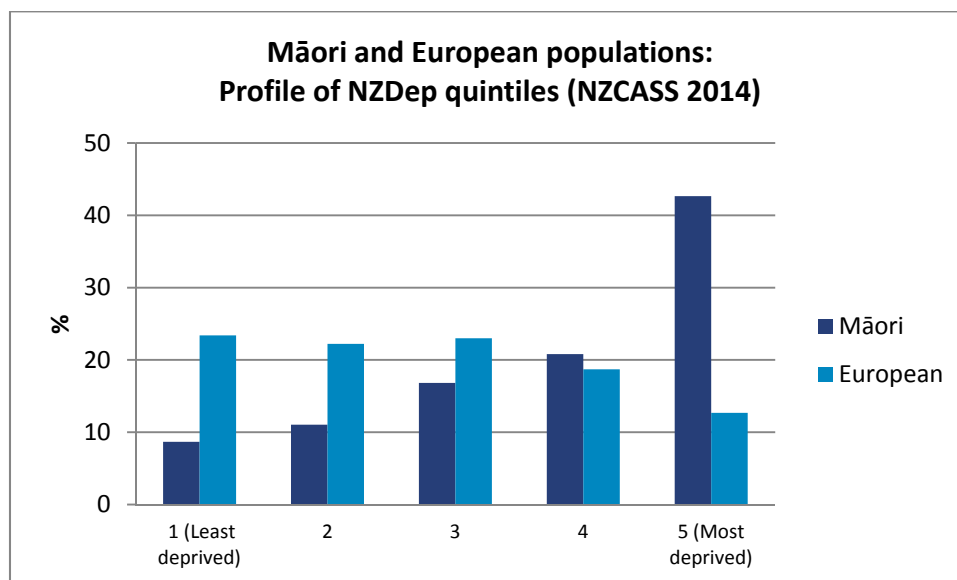
analysing the victimisation differences between Māori and Europeans, the question remains whether the difference is due to socio-economic differences rather than ethnicity, since lower socio-economic groups are more likely to be victimised and proportionately more Māori have lower socio-economic status. NZDep2013 was selected as the socio-economic measure since it is a multi-dimensional measure.

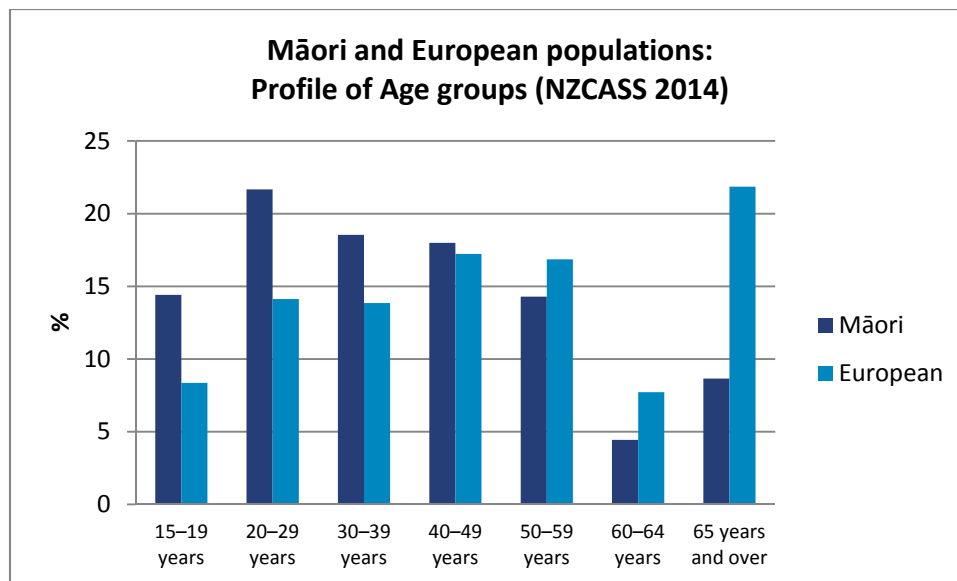
- In addition to socio-economic status, age was selected due to the age profile differences in the Māori and European populations. The 2013 Census shows the large age differences in the Māori and European populations; for example, Māori have a median age of 24 years, and Europeans have a median age of 41 years (Statistics NZ 2013).
- This was further supported by the data through using a decision tree analysis to determine the prominent drivers of victimisation (similar approach to the Statistics NZ (1998) paper). Once this analysis was performed, the 2 variables of age and deprivation emerged as important.

This analysis does not attempt to control for *all* differences between Māori and Europeans, but rather to consider some of the main differences to then assess the size of the victimisation risk gap once these factors are controlled for.

Structural differences of age and deprivation

Viewed graphically we can see the profile differences between Māori and European for age and deprivation. In comparison to Europeans, Māori are a younger population and proportionally more Māori live in areas of higher deprivation.





All things being equal

The effect of contributing factors can be removed by standardisation, which re-weights the factors of the Māori and European populations to give the same structure as the combined Māori and European population – that is, we are trying to answer the question: What would the victimisation risk gap be between Māori and Europeans if they had a similar age and deprivation structure? This means that any remaining differences in victimisation between Māori and Europeans would then not be due to the age and deprivation differences between the 2 populations.

This approach to control for multiple factors at once is termed *multiple standardisation* – a term and analysis technique used, for example, in Australian Bureau of Statistics (2014). For this analysis, age and deprivation are the factors to standardise by, which means we are analysing the effect of what if both Māori and Europeans had the same age structure *and* deprivation. The process to do this is to ‘weight up’ or ‘weight down’ the responses to give the same profile across the combined population. For example, since there are more Māori aged 15–19 years, these responses would be ‘weighted down’ and the European 15–19-year-old responses would be ‘weighted up’ so the age and deprivation profiles are the same. Then the victimisation rates are re-calculated on the re-weighted dataset.

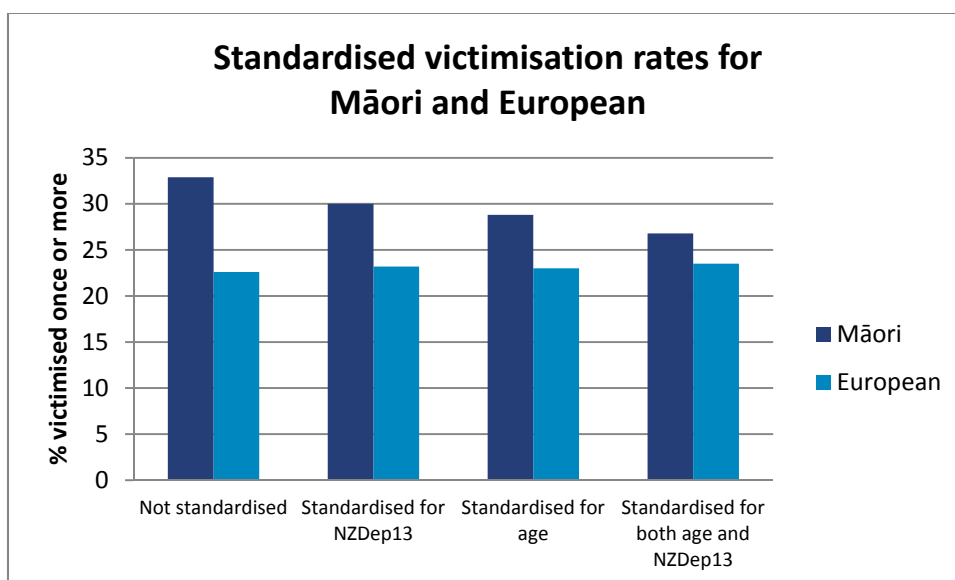
For people that identified as *both* Māori and European, this analysis halved their original weight so they would only contribute once in the analysis (half as Māori and half as European). If this wasn’t done, people who identify as both ethnicities would be analysed twice, which would in effect decrease the difference between Māori and Europeans as more weighted individuals would be alike.

Multiple standardisation results

After adjusting for a combination of age and deprivation, the gap in victimisation risk between Māori and Europeans fell from 10.3 percentage points to 3.3 percentage points. This difference remains statistically significant. These results are summarised as follows, along with the individual contribution to decreasing the victimisation risk gap for each of age and deprivation.

Table 1: Multiple Standardisation Results

	Māori	European	Difference (percentage points)
Not Standardised	32.9%	22.6%	10.3
<i>Standardised for multiple factors:</i>			
Both age and NZDep13	26.8%	23.5%	3.3
<i>Standardised for individual factors</i>			
Age	28.8%	23.0%	5.8
NZDep13	30.0%	23.2%	6.8



This analysis tells us that once controlling for age and deprivation, Māori are still more likely to be victimised than Europeans; however, over two-thirds of the victimisation risk difference can be accounted for by age and deprivation alone.

This analysis only accounts for 2 of the structural differences between the Māori and European populations. There will be other factors that may or not be measured in the NZCASS that will impact the victimisation risk differences between the 2 populations.

Logistic regression

The NZCASS data tables present a wide range of cross-tabulations of 2 or more data items. However, this presentation does not take into consideration the relationship between different factors. For example, young people are more likely to be students and live with flatmates, while people aged 65 years or more will be more likely to be retired. So if a factor comes through as statistically significant compared to the New Zealand average, such as being a student, it may not be that because someone is a student they are at higher risk of victimisation, but rather because student status is correlated with being younger. Due to this, it is difficult from the factor tables to assess which factors are directly related to victimisation, and which factors are secondary factors related to victimisation only through correlation to another factor. Regression was used as a modelling technique to deal with this overlap, where the model can control for holding other variables constant. The modelling process selects the most predictive combinations of many overlapping factors, and estimates their effect.

Models

The regression models were built for the probabilities of the following response variables.

Victimisation of:

- all offences
- interpersonal violence
- interpersonal violence by an intimate partner
- burglary
- thefts and damage offences
- vehicle offences.

As each of these response variables are whether someone was victimised or not, the models were fitted as logistic regression models. This is a widely used method for explaining binary response variables.

These models model whether someone was victimised or not, rather than the number of times they were victimised. This was done since whether someone was victimised or not is more stable, whereas the number of times someone was victimised can be affected by a few highly victimised people. But note – slightly different results might have emerged based on predictors for the number of times someone was victimised.

Explanatory variables

Table 2 summarises the explanatory variables included in the initial models.

Table 2: Explanatory variables included in initial regression models

Household offences (Burglary, vehicle offences)	Personal offences (All offences, interpersonal violence, intimate partner violence, theft and damage offences)
<i>Personal factors</i>	
	Sex
Age ◊	Age ◊
Ethnicity	Ethnicity
	Partnership status
<i>Economic factors</i>	
	Employment status
	Financial stress: limited to buy item for \$300 ◊
Financial stress: can meet unexpected expense	Financial stress: can meet unexpected expense
	Personal income ◊
Household income ◊	Household income ◊
<i>Household factors</i>	
Household composition	Household composition
Tenure and landlord type	
<i>Geographic factors</i>	
Urbanisation	Urbanisation
Region	Region
<i>Other factors</i>	
Average rating of social disorder ◊	Average rating of social disorder ◊

◊ Included in model as a continuous variable. All other variables treated as categorical.

Note: For the burglary and vehicle offences model, it is somewhat artificial to analyse household offences against personal characteristics (such as age and ethnicity), since this depends on which respondent in the household was selected for the interview. For this reason characteristics such as sex and employment status were not included in the household models. Age, ethnicity and financial stress are also personal factors, but these are considered more homogenous amongst household members than other personal factors. However, caution is advised when interpreting these personal factors in the household models, and the interpretation is that the characteristic reflects the average profile of household members.

These explanatory variables were selected from the standard range of demographic and geographic data items included in the NZCASS data tables. From this standard set, there are 3 exceptions:

1. NZDep2013 quintiles were excluded because they are derived from multiple measures of deprivation, some of which were also included in the model. It is preferable for un-derived variables to be retained as they are easier to interpret.
2. Legally registered relationship status was excluded as partnership status is derived from this.

3. A social disorder rating was included. This was calculated from the 6 aspects of neighbourhood crime problems. Respondents were asked to rate each of the 6 aspects from '1 – A very big problem' through to '4 – Not a problem at all'. The disorder rating was calculated for each respondent as the average of the codes for all 6 aspects. The scale was reversed so that a higher rating indicated greater disorder.

Model specification

From this list of explanatory variables the model that was fitted was:

$$\text{victimised} \sim \text{personal} + \text{economic} + \text{household} + \text{geographic} + \text{other factors}$$

where victimised is a binary response where 1 = victimised, 0 = not victimised.

The categorical variables in **Error! Reference source not found.** were included in the model as dummy (indicator) variables. This means individual variables were included for each category within a factor, which takes the value of 0 or 1 to indicate the presence or absence of a category (eg has value 1 to represent 'employed' and has value 0 to represent everyone else). The alternative was to include the factor categories as 1 variable (eg employed, unemployed, etc), but dummy coding was done to assist interpretation of the victimisation odds ratio for that category compared to the rest of the population, rather than that category compared to a specified baseline.

When categorical variables are converted into multiple dummy variables, they can exhibit redundancy. For example, the dummy variable of 'Rest of South Island' can be identified as none of the other regions (Auckland, Wellington, Rest of North Island, Canterbury). This fails an assumption of logistic regression, and hence when this occurred, 1 less category than the number of categories was excluded from the initial model. The decision on which category to exclude was a conceptual one based on which category was most similar to the New Zealand average or most met the project team's research needs. For example, 'other multi-person household' was the household composition category excluded since it is a smaller category that wasn't significantly different from the New Zealand average. Similarly, 'Rest of South Island' was the region category excluded, since the NZCASS research needs were primarily to focus on the other regions. This is not to say these excluded categories are not important – these categories *are* still being represented in the model except more indirectly as they are represented as the '0' group of every dummy within that factor.

A number of variables were treated as continuous in the model (denoted with a \diamond in table 2) in order to keep the natural ordering of these variables in the model. The categories used for these variables varied from that presented in the data tables to ensure that similar increments were used. For example, age is presented in the data tables as a mix of 5-year and 10-year age groups, and with the upper group of 65 years and over. However, for the regression, age was included in evenly sized 5-year age groups, and the upper group was 75 years and over.

Interaction terms were included for selected variables (including age/ethnicity, employment status/ethnicity and household composition/financial stress); however, it was not appropriate to fit all possible interactions due to the number of explanatory variables and sample size restrictions. Quadratic terms of personal and household income were included for selected offence models where there was assessed to be a non-linear relationship with victimisation (eg when lower income has higher rates of victimisation, middle income has lower rates of victimisation, and then higher income has higher rates of victimisation). The assumption of linearity was valid for the remaining continuous variables.

The regression models were fitted unweighted. All other NZCASS analysis was conducted using weights to compensate for imbalances in the survey profile relative to the target population. However, with this regression analysis, the intention differs in regard to prediction. With other NZCASS analysis we aim to describe victimisation rates in 2013 (ie the 'how many'), whereas this regression analysis aims to provide an understanding of the predictors of victimisation (ie the 'who'). Hence it was decided an unweighted model was more appropriate, and this is consistent with the modelling approach in previous NZCASS iterations. Furthermore, there were practical considerations in that SAS does not implement a step-wise backwards weighted regression using proc surveylogistic. There were other, less important, reasons for preferring unweighted estimates such as robustness with respect to extreme weights. The differences between the weighted and unweighted results are briefly discussed under 'Weighted regression'.

Missing values in the explanatory variables (such as refusal or 'don't know' responses) were imputed to prevent the entire record being dropped during the modelling process. The following imputation method was used:

1. First, missing records were matched with potential donors using known characteristics of age, deprivation and ethnicity.
2. The donor's value of the missing cell were then assigned to the respondent. In cases where there were multiple donors (most of the time), 1 potential donor was randomly selected from the pool. In the very small number of cases where there were no donor matches on age, deprivation and ethnicity, the donors were matched on age and deprivation alone (every respondent was matched just on the 2 characteristics).

Influential observations were assessed, and a small number of observations (up to 6) were removed from each model.

Model selection

The regression analysis was a step-wise backwards elimination method, which starts with the full model and drops 1 variable at a time until all remaining variables are statistically significant. This is a widely used variable selection technique, but does have the known limitation that constructed confidence intervals are too narrow. This approach was selected since there are a large number of candidate explanatory variables, coded into 35 main effects plus interactions for the personal offences models. This works out to be billions of different model combinations, even without considering interactions. Hence the step-wise elimination method is a relatively quick method that can be automated to evaluate a range of explanatory variables.

As the NZCASS dataset has 100 imputations, the regression models were built on the 100 imputation datasets separately. The final model was specified by combining the results from the 100 individual models using the standard Rubin combining rules (see '**Error! Reference source not found.**' for further information). Only explanatory variables that were retained in at least 40 of the 100 models were included in the final model. This threshold of 40 was used because from an analysis of frequencies, this was a natural separation point where there was a cluster of variables only included in a few or dozen models, while the remaining were included in most or all of the models.

The significance level was set at 95%. Other significance levels were considered, such as 90%, but the 95% level was used for 2 main reasons:

1. 2014 NZCASS reporting was done at the 95% level

2. the modelling aimed to determine the ‘best’ predictors of victimisation (ie not including too many variables that may be included with a lower threshold).

While considering the need to not exclude potentially important predictors, this second consideration was balanced with the decision to include the explanatory variables retained in 40 of the 100 imputation final models. The threshold was slightly lower with the aim to be conservative, and to ensure the final pooled model had a similar number of explanatory variables to the final individual models.

Interpreting results

The logistic regression model expresses the logarithm of the odds ratio of being a victim as a linear combination of the explanatory variables. However, for ease of interpretation, the odds ratios have been presented by taking the exponential of the log odds ratios.

The odds ratio represents higher or lower odds of victimisation, while controlling for the other factors. An odds ratio greater than 1 indicates *higher* odds of victimisation, where a number less than 1 indicates *lower* odds of victimisation when compared to the reference group.

Odds are not the same as probabilities – both are numerical measures of how likely an event is to occur, but have different interpretation.

PROBABILITIES

Probabilities are the chance, or risk, that something will occur. For example, if we use 2013 statistics to estimate future risk, there is a 9.5% probability that a household in Auckland will be burgled in the next year. Similarly, there is a 7.4% probability that a household outside Auckland will be burgled.

ODDS

Odds can be calculated from probabilities where $\text{odds} = \text{probability} / (1 - \text{probability})$. For example, the odds of a household experiencing a burglary in Auckland in the next year are 0.105 ($0.095 / 1 - 0.095$). Similarly, the odds of a household not in Auckland experiencing a burglary in the next year are 0.080 ($0.074 / 1 - 0.074$).

ODDS RATIO

The odds ratio is the odds relative to another group. For example, the odds ratio of a household experiencing a burglary in Auckland compared to the rest of the country is 1.31 ($0.105 / 0.080$).

INTERPRETATION

This means the odds of experiencing a burglary in Auckland is 31% higher than the odds for experiencing a burglary in the rest of the country.

The interpretation of odds ratios for categorical and continuous explanatory variables differs. For continuous variables, the odds ratio represents the victimisation odds change *with 1 unit increase in that characteristic* holding other variables constant. For example, the unit increase in age is in 5-year bands. Table 3 shows the odds ratio for age is 0.9 for violent interpersonal offences. This means the odds of victimisation are 10% lower for every 5-year age increase, holding other factors constant.

For categorical variables the odds ratio represents the victimisation odds *compared to the reference group*, holding the other variables constant. The reference group is as follows:

- For cases where only 1 category of that variable is retained in the final model or for the multiple response category of ethnicity, the reference group is the rest of the population. For example, **Error! Reference source not found.** shows that '1 parent with children' households have an odds ratio of 1.31 for violent interpersonal offences. This means the odds of being a victim of violent interpersonal offences are 31% higher than that of the rest of the population when other factors are held constant.
- For cases where there are 2 or more categories of that variable retained in the final model, the reference group is everything outside those categories. For example, **Error! Reference source not found.** shows the odds ratio for the 'all offences' model for Auckland is 1.18, and that Auckland and Canterbury were the 2 regions retained in the final model. This is therefore interpreted as the odds of being victimised in Auckland are 18% higher than compared to odds of someone not in Auckland nor Canterbury being victimised.

Final model results

Table 3 summarises the odds ratio and the corresponding confidence intervals for the best (final) models.

Table 3: Final Logistic Regression Model Results

Model	All offences	Burglary	Thefts and damage offences	Vehicle offences	Violent interpersonal offences	Intimate partner violence
Odds ratio						
[95% CIs]						
<i>Intercept</i>	0.09 [0.06 – 0.13]	0.02 [0.01 – 0.03]	0.02 [0.00 – 0.05]	0.03 [0.02 – 0.06]	0.07 [0.04 – 0.12]	0.11 [0.05 – 0.27]
Categorical variables						
<i>Sex: Female</i>		-		-		1.36 [1.05 – 1.75]
<i>Ethnicity: European</i>			1.31 [1.03 – 1.66]		1.28 [1.06 – 1.54]	
<i>Ethnicity: Māori</i>	1.24 [1.09 – 1.42]	1.41 [1.17 – 1.69]	1.64 [1.04 – 2.60]		1.56 [1.30 – 1.88]	1.56 [1.15 – 2.12]
<i>Partnership status: Partnered – legally registered</i>		-		-	0.67 [0.55 – 0.81]	0.51 [0.36 – 0.72]
<i>Partnership status: Partnered – not legally registered</i>	1.20 [1.04 – 1.40]	-		-	1.40 [1.14 – 1.72]	1.64 [1.20 – 2.24]
<i>Employment status: Retired</i>	0.74 [0.58 – 0.94]	-		-		
<i>Financial stress: Can meet unexpected expense</i>			0.90 (t) [0.68 – 1.19]		0.73 [0.59 – 0.90]	0.57 [0.42 – 0.77]
<i>Financial stress: Can meet unexpected expense *</i>			3.33 [1.13 – 9.81]			
<i>Household comp: 1 parent with child(ren) and other person(s)</i>						
<i>Household comp: 1 parent with child(ren)</i>	1.31 [1.09 – 1.58]	1.40 [1.06 – 1.85]			1.32 [1.03 – 1.68]	
<i>Household comp: 1 parent with child(ren) and other person(s)</i>			0.48 (t) [0.19 – 1.23]			
<i>Household comp: Couple only</i>					0.70 [0.55 – 0.89]	
<i>Household comp: Couple with child(ren)</i>		1.27 [1.04 – 1.56]			0.73 [0.59 – 0.90]	
<i>Household comp: Couple with child(ren) and other person(s)</i>		1.65 [1.07 – 2.55]				

Model	All offences	Burglary	Thefts and damage offences	Vehicle offences	Violent interpersonal offences	Intimate partner violence
<i>Tenure and landlord type: Rented – government (local and central)</i>	-	1.41 [1.05 – 1.90]	-	-	-	-
<i>Urbanisation: Main urban area</i>	1.25 [1.07 – 1.46]	1.41 [1.15 – 1.74]	-	1.34 [1.04 – 1.72]	-	-
<i>Urbanisation: Secondary urban area</i>	1.41 [1.09 – 1.83]	-	1.79 [1.14 – 2.81]	-	1.41 [1.05 – 1.91]	-
<i>Region: Auckland</i>	1.18 [1.02 – 1.35]	-	-	1.67 [1.32 – 2.11]	-	-
<i>Region: Rest of North Island</i>	-	-	0.70 [0.55 – 0.90]	-	-	-
<i>Region: Canterbury</i>	1.23 [1.02 – 1.49]	-	-	-	-	-
Continuous variables						
<i>Age (Increasing age)</i>	0.93 [0.91 – 0.96]	0.96 [0.93 – 0.99]	0.96 (‡) [0.92 – 1.01]	0.91 [0.87 – 0.94]	0.90 [0.87 – 0.92]	0.88 [0.83 – 0.93]
<i>Age * Ethnicity: Māori (Increasing age)</i>	-	-	0.93 [0.87 – 0.99]	-	-	-
<i>Financial stress: Limited to buy item for \$300 (Increasingly limited)</i>	1.08 [1.04 – 1.13]	-	1.12 [1.03 – 1.21]	-	1.09 [1.03 – 1.16]	-
<i>Personal income (Increasing income)</i>	1.06 [1.01 – 1.12]	-	-	-	-	0.87 [0.79 – 0.97]
<i>Personal income squared (See note below)</i>	-	-	1.02 [1.00 – 1.03]	-	-	-
<i>Average rating of social disorder (Increasing disorder)</i>	1.86 [1.69 – 2.04]	1.98 [1.74 – 2.25]	2.20 [1.89 – 2.57]	1.81 [1.54 – 2.12]	1.71 [1.51 – 1.93]	1.50 [1.25 – 1.80]


Notes:


Only variables that were retained in the final models are presented.

‡ Not significant but included in final model as interaction term included.

* Interaction term – that is, the simultaneous combination of the 2 variables.

- Not included in initial model (depending on whether a personal or household offence).

 Decreasing victimisation odds (but note the interpretation of interaction terms).

 Increasing victimisation odds (but note the interpretation of interaction terms).

The interpretation of interaction terms needs to be done alongside the main effect terms. For example, for theft and damage offences the interaction age*Māori is negative (odds ratio 0.93) and the age main effect is negative (odds ratio 0.96), whereas the Māori main effect is positive (odds ratio 1.64). When calculating the combined probability for Māori, the change in the negative contribution is greater than the change in the positive contribution. Hence for Māori, the effect of increasing age is still lower victimisation odds.

The interpretation of quadratic terms (in this case personal income squared) is that since the term is greater than 1, the relationship with victimisation is 'U' shaped, in that people with lower and higher personal income have higher odds of victimisation, whereas people with middle income have lower odds of victimisation.

Note that the reverse interpretation can be applied. For example, people in a not-legally registered partnership status have *higher* odds of victimisation for all offences than the rest of the population. Conversely, this can be interpreted that people who are in this 'rest of the population' group (ie legally registered partnerships and non-partnered people) have *lower* odds of victimisation. This is particularly important to consider when some categories were excluded due to avoidance of dummy variable redundancy (see Model Specification).

Model fit

The predictive power of the logistic regression models has been measured by a statistic called the area under the Receiver Operating Characteristic (ROC) curve. If the model is weak at distinguishing victims from non-victims, the ROC statistic will have a value around 0.5 (no better than a coin toss). Whereas, if the model is perfect at distinguishing victims from non-victims, the statistic will have a value of 1.0. **Error! Reference source not found.** shows the ROC statistics for each model.

Table 4: Model Fit statistics for each regression

Model	ROC statistic
All offences	0.685
Burglary	0.686
Thefts/damage offences	0.699
Vehicle offences	0.687
Violent interpersonal offences	0.726
Intimate partner violence	0.756

As most of the ROC statistics range between high 0.6 to mid 0.7, this shows the models are helpful but do not have perfect explanatory power. There still remain other unmeasured factors or perhaps random behaviour that puts some people/households more at risk of victimisation than others. The household offences models have lower predictive fit than the personal offences, indicating there are more unmeasured factors (such as presence of alarms/CCTV) or random behaviour for these offences.

Caveats

There are a number of caveats to be aware of when using regression results.

- **Correlation does not prove causation:** Not all possible drivers of victimisation are included in the regression models, since the NZCASS does not collect all possible characteristics or predictors of victimisation. Variables may be retained in the final model only because they are related to an unmeasured variable. For example, age was retained in all the final models, but it may not be someone's actual age that puts them at risk, but rather the way they socialise, how they live and the places they go. For this reason, the statistical models do not provide a perfect explanation of what predicts victimisation.
- **Collinearity:** Several variables may be correlated, but just 1 may be retained in the final model that is most closely related to victimisation. This does not mean that the other variables have no importance at all, but rather, it is not the 'best' predictor of victimisation.
- **Sample size:** An explanatory variable may not be retained in the final model simply due to not having the sample size to support its inclusion. This is not to say that it is not important in predicting victimisation. For example, this may affect the findings for Pacific peoples since the sample size is smaller than that for Māori or European ethnicity.

Weighted regression

Unweighted logistic regression models were used for the regression analyses presented above. The same models were re-run (using proc surveylogistic) with weights, stratification variables of region and urbanisation, and cluster variable of meshblock specified. The results of the unweighted and weighted models were compared.

Most importantly from this comparison, the differences between the odds ratios from the unweighted and weighted models were minor. The estimates were very similar for most explanatory variables, or where they did differ, the direction (ie higher or lower odds ratio) was consistent. Only a small number of estimates differed largely in terms of magnitude, and when that occurred it could be explained by the variable not being statistically significant in either the weighted or the unweighted model. Hence it can be concluded that overall, weighting made immaterial difference to the odds ratios.

While the effect on odds ratios was minor, the use of weights would have changed the model variable *selection* results. **Error! Reference source not found.** summarises which explanatory variables that were retained in the final unweighted model did not meet the 95% level criteria (had *p*-values greater than 0.05) when run weighted.

Table 5: Explanatory variables that do not meet the 95% threshold with a weighted model

Model	Explanatory variables
All offences	<i>Personal income</i> <i>Household comp:</i> 1 parent with child(ren) <i>Partnership status:</i> Partnered – not legally registered <i>Urbanisation:</i> Secondary urban area
Burglary	<i>Household comp:</i> Couple with child(ren) <i>Household comp:</i> Couple with child(ren) and other person(s) <i>Tenure and landlord type:</i> Rented – government (local and central)
Thefts/damage offences	<i>Ethnicity:</i> European <i>Financial stress:</i> Can meet unexpected expense <i>Household comp:</i> 1 parent with child(ren) and other person(s) <i>Region:</i> Rest of North Island <i>Urbanisation:</i> Secondary urban area
Vehicle offences	<i>Region:</i> Auckland
Violent interpersonal offences	<i>Ethnicity:</i> European <i>Partnership status:</i> Partnered – legally registered <i>Partnership status:</i> Partnered – not legally registered <i>Financial stress:</i> Can meet unexpected expense <i>Financial stress:</i> Limited to buy item for \$300 <i>Household comp:</i> Couple with child(ren) <i>Household comp:</i> 1 parent with child(ren) <i>Urbanisation:</i> Secondary urban area
Intimate partner violence	<i>Financial stress:</i> Can meet unexpected expense <i>Personal income</i> <i>Partnership status:</i> Partnered – not legally registered <i>Sex:</i> Female

Note: This table does not show which alternative variables (if any) would have been retained in a weighted model using the same step-wise backwards elimination method.

Most of these were close to the threshold of 95% (*p*-values of 0.05), but some were larger indicating there is sample design effect and clustering of those variables. Particularly for violent interpersonal

offences, the potential variable identification is quite different for the weighted and unweighted models. This could be explained by a small number of respondents with large weights having a large effect on the retention of variables, and/or if violent interpersonal offences are highly concentrated, then considering strata and clustering would effectively reduce the sample size, and hence reduce model power to determine predictors of victimisation. But in general, this is provided for user information and for the reasons discussed under 'Model specification', the unweighted model has been assessed as the more preferred model to be used to provide an understanding of predictors of victimisation.

Gini coefficient

This part of the advanced statistical methods section uses a 'Gini coefficient' to provide a measure of the distribution of victimisation. It is appealing as a statistic since it summarises the distribution in a single summary statistic, enabling comparisons over time and between groups to be captured succinctly. This can be used to see whether victimisation is becoming more or less equal over time, and also to assess whether the distribution of some offences is more unequal than others.

The Gini coefficient is a summary measure of inequality, used widely in analysing the distribution of income or wealth. Extending this to victimisation, it is a measure where:

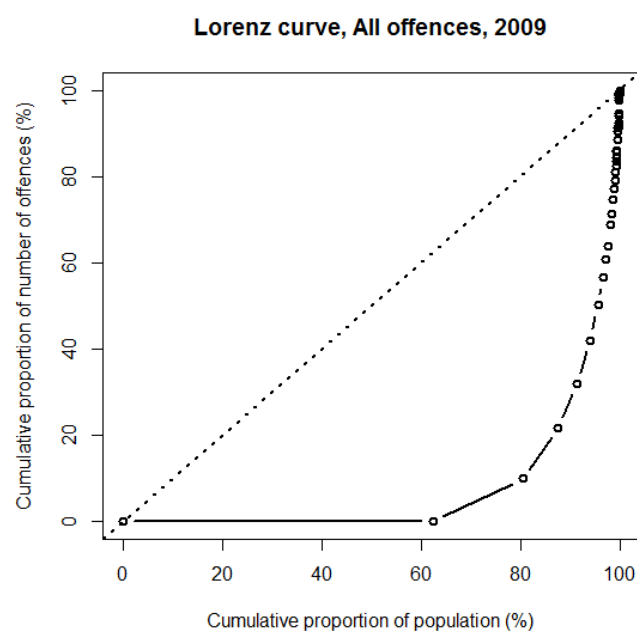
0 = perfect equality – where all members of the population experience the same amount of victimisation

1 = total inequality – where only 1 person/household experiences all the victimisations, and everyone else experienced none

The smaller the Gini coefficient, the more even the distribution of victimisation. The larger a coefficient, the more concentrated victimisation is amongst a group of highly victimised people.

Mathematics of the Gini coefficient

The Gini coefficient is calculated in reference to the Lorenz curve – a graph with the horizontal axis showing the cumulative proportion of the population ranked according to the number of offences experienced, and the vertical axis showing the corresponding cumulative proportion of the number of offences they have experienced. For example:



If victimisation was evenly distributed amongst the population, then the Lorenz curve would be the diagonal (line of equality). The value of the Gini coefficient is the ratio of the area between the Lorenz curve and the diagonal to the area of the entire diagonal triangle.

Calculations for the NZCASS

This area ratio is equivalent to calculating the Gini coefficient through a direct formula. This was implemented in the NZCASS analysis as follows:

1. Calculate weighted frequencies and percentage of number of offences experienced for all number of offences i .
2. Calculate the cumulative number of offences experienced, and the cumulative percentage of population (x).
3. Calculate the cumulative percentage of offences experienced (y).
4. The Gini is then calculated as:

$$\sum [x_i \times y_{i+1}] - \sum [x_i \times y_{i+1}]$$

5. The final Gini is the last row (ie where $y = 100$).

Populations

The populations used for the Gini calculations varied for the comparisons across time and comparisons between offence groups:

- *Comparisons across time:* The population is the total *adult* population (ie those who experienced no offences were included in the calculations)
- *Comparisons between offence groups:* The population is the *victim* population (ie those who experienced no offences were excluded in the calculations).

For the comparisons across time, we want to see how the distribution of victimisation has changed for all adults, so the large percentage of adults who did not experience any victimisation are an important part of this story. However, for comparisons between offence groups, we are more interested in the distribution of victimisation for victims (ie excluding the non-victims), and since a large portion of the population would have experienced none of that offence group, it would skew the Gini statistic upwards.

Gini results

The Gini results are presented in the NZCASS Main Findings Report, and also included here along with the Lorenz curve graphs to assist interpretation.

Table 6: Gini Results – across time

	2006	2009	2014
Gini coefficient	0.810	0.814	0.886‡

‡ Statistically significant difference across time at the 95% confidence level from both 2005 to 2013 and 2008 to 2013.

The following Lorenz curve plot assists interpretation of why the Gini has increased in 2014 – that is, because there are now more people who have not been a victim of crime, but there is a small group of people who remain highly victimised, then the distribution of victimisation is now more unequal than in previous years.

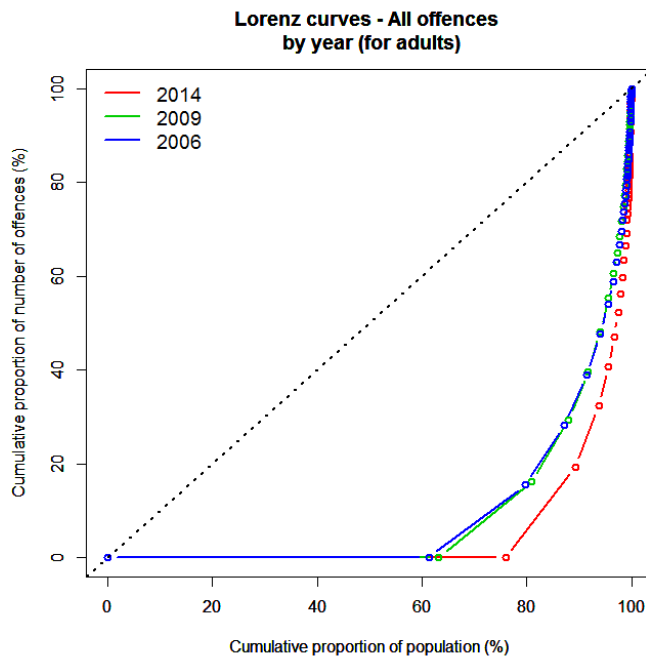


Table 7 and the following figure present the Gini results by offence group for 2014.

Table 7: Gini Results – between offence groups - 2014

	Gini coefficient
Interpersonal violence	0.566
Burglary	0.289*
Thefts and damage	0.262*
Vehicle offences	0.203*

* Statistically significant difference to the interpersonal violence Gini at the 95% confidence level.

**Lorenz curves - 2014
by offence (for victims)**

