Modelling crime victimisation and the propensity to report crime to police

Robert Tanton, René Jones and Godfrey Lubulwa

Paper prepared for the Methodology Advisory Committee Meeting
22 June 2001
Australian Bureau of Statistics

The team would like to thank the following ABS staff who have provided input and/or useful comments at various stages: Ken Tallis, Phil Bell, Carol Soloff and Bernard Wong. Staff from the following external agencies have also been generous with their time and valuable comments: The Australian Institute of Criminology, NSW Bureau of Crime Statistics and Research, NSW Police Department, WA Crime Research Centre, WA Justice Dept., Queensland Treasury, SA Office of Crime Statistics, the Australian Centre for Policing Research.

For comments or questions please contact Robert Tanton on (02) 62525506 or
robert.tanton@abs.gov.au

Executive Summary

The ABS is currently undertaking two projects using data from the 1998 National Crime and Safety Survey. These projects are being done in collaboration with the Australian Institute of Criminology and the ABS National Centre for Crime and Justice Statistics in Melbourne.

The two projects currently under way are:

- deriving small area estimates of crime prevalence and incidence rates; and
- estimating the propensity to report crime.

Direct survey estimation of variables at small area level yields unacceptably high standard errors due to the small sample sizes. Increasing the sample size can overcome this problem, but this is an expensive option. The small area estimates of crime project is an attempt to derive small area estimates of crime rates using a combination of the 1998 National Crime and Safety Survey and 1996 Census data. Results from this project will lead to a greater understanding of the crime in local areas by practitioners in the field.

It is proposed that a Logit model is used to model the propensity to report and to test the significance of socio-economic factors that affect an individual's reporting decision. Results from this model will give practitioners in the field a better understanding of which groups have a higher or lower propensity to report crime.

Analysis Branch is involving criminologists and practitioners in this work through the project board, and through a peer review group reviewing the methods. The project board has a Police representative and a criminologist from the Australian Iinstitute of Criminology on it. The peer review group consists of criminologists and statsiticians in the criminology field.

Approach

Modelling the Crime and Safety Survey
A major challenge in terms of methodology is how to accommodate and adjust for the survey design which has a combination of stratification and clustering aspects. A number of techniques to take into account survey design have been considered. It is proposed that survey design is taken into account through a weighting system in the proposed estimation procedure (either using replicate weights or a weighted least squares technique).

Small area estimation of crime rates
Small area estimation applies to a range of techniques that allow the calculation of estimates for small areas from larger areas, using survey and other data. The techniques essentially "borrow strength" from related areas to find more accurate estimates for a given area.

The small area technique proposed in this study is based on a regression approach. Under this method regression coefficients are estimated from the survey for larger areas. Then these regression coefficients are applied to census data from smaller areas to derive estimates for these smaller areas.

Estimation of the propensity to report crime to police

It is proposed that a Logit model is used to model the propensity to report crime. The model will be estimated using data from the 1998 National Crime and Safety Survey. The dependent (binary) variable is the decision to report the crime. Independent variables include age, sex, whether a weapon was used, injury and whether anything was stolen. For this model, use is made of unit record survey data.

Data sources and variables

The main source of data for the two projects is the 1998 National Crime and Safety Survey. This survey covered the major crimes of break and enter, attempted break and enter, motor vehicle theft, robbery, assault, and sexual assault. These crime types can be joined into two broad crime groups, being personal crime (assault, sexual assault and robbery) and Property Crime (break in, attempted break in, and motor vehicle theft).

The dependent variables for the small area estimates model are the victimisation prevalence and incidence rates. Different models are estimated for each crime type, and broad crime group (Personal and Property crimes).

The independent variables for the small area estimates model came from the 1996 Census. The Census has full coverage of Australia, and provides reliable population figures at a small area level. Some independent variables (for assault only) also came from the 1998 National Crime and Safety Survey.

The dependent variable for the propensity to report model is the 'reported to police' variable from the 1998 National Crime and Safety Survey. The independent variables for this model are from the 1998 National Crime and Safety Survey.

Robustness

The following tests for robustness are proposed. For the small area estimates model these tests include:

- Simple aggregation of the small area estimates of crime and comparison of these aggregates with reliable survey estimates;
- Sensitivity analysis of the model to investigate how small area estimates vary as selected key independent variables change; and
- Use of the coefficients from the 1998 National Crime and Safety Survey against the 1991 Census data, and compare the rates estimated using these data with the victimisation rates in the 1993 National Crime and Safety Survey.

For the propensity to report crime project, they are:

- Sensitivity analysis of the model; and

- Run the same model against the 1993 Crime and Safety Survey and compare the results.

Results

This paper is exploring methodology to be used in the study. While some draft results may be available by the meeting, we do not present any results int his paper.

Proposed Discussion Points for MAC

We would welcome MAC's comments on any aspect of this paper. Particularly valuable would be comments on:

(i) the **methodology, techniques and research strategies** we have used; and

(ii) what conceptual and statistical frameworks might we use for our longer term research on Crime.

- We are particularly interested in views on the small area estimation technique used, since this is a generic method and could be used for other surveys. Have we adopted the best techniques for small area estimation? Have we got the trade off between simplicity and statistical rigour about right?

- Is the method we have used to take survey design into account appropriate? What improvements can we make to the proposed method, and what are the costs (in terms of complexity) and benefits (in terms of better estimates) associated with the improvements?

- Have we adopted the right solution to multicollinearity?

- Are our tests for robustness appropriate? Are there any other tests we could do?

Table of Contents

Executive summary

1. Introduction

2. The 1998 National Crime and Safety Survey
   2.1 Description of the survey
   2.2 Sample Design
   2.3 'Not Stated's

3. Small Area Estimates of Victimisation
   3.1 Small area estimation techniques
   3.2 Description of the approach to the estimation of small area crime victimisation rates
   3.3 Preliminary results (if available)
   3.4 Other relevant issues excluded from the study due to unavailability of data

4. Propensity to Report
   4.1 Method
   4.2 Data
   4.3 Preliminary results (if available)
   4.4 Validating the model

5. Conclusion

Appendix 1. Taylor Expansion Method

Appendix 2. List of variables with 'not stated' classification, and the proportion of 'not stated's

Appendix 3. Graph of survey weight and 1/Standard Error

Appendix 4. Full description of the data for the small area estimates model

Appendix 5. Full description of the data for the propensity to report model

Bibliography

1. Introduction

1. This paper discusses the methods proposed for use in two projects currently under way:

- deriving small area estimates of crime prevalence and incidence rates; and
- estimating the propensity to report crime.

2. These projects use the 1998 National Crime and Safety Survey (NCSS) to model crime rates and the propensity to report crime. They were started in December 2000, and the planned completion date is late 2001.

The aim of this paper is to:

- describe the methods we are using to estimate the models; and
- discuss the issues we face with the estimation.

3. The paper has four sections. Section 2 discusses aspects of methodology which apply both to the small area estimation project and the propensity to report crime to police project. Section 3 discusses methods which are specific to small area estimation, while section 4 deals with methods specific to the estimation of the propensity to report. Section 5 makes some concluding remarks.

2. The 1998 National Crime and Safety Survey

2.1 Description of the Survey

4. This survey was conducted in April 1998, with the reference period April 1997 to April 1998. The survey covered the more serious crimes that affect the largest number of people, being:

- **break-in** - an incident where the respondents home had been broken into. The home included their garage or shed, but excluded their car and garden.
- **attempted break-in** - an incident where there were signs that an attempt was made to break into the respondents home. The home included their garage or shed, but excluded their car and garden.
- **motor vehicle theft** - an incident where a motor vehicle was stolen from any member of the household.
- **assault** - an incident other than a robbery involving the use, attempted use, or threat of force or violence against the victim.

- **sexual assault** - an incident which was of a sexual nature involving physical contact, including rape, indecent assault, and assault with intent to sexually assault. Sexual harassment was excluded. Only females aged 18 years and over were asked sexual assault questions.

- **robbery** - an incident where someone had stolen (or tried to steal) property from a respondent by physically attacking them or threatening them with violence.

5. The 1998 National Crime and Safety Survey was a supplement to the ABS April Monthly Population Survey. Information was sought from 51,800 persons, of whom about 42,200 (81.4%) responded. Data pertaining to households was sought from 25,600 households, of which about 20,900 (81.6%) responded. The respondents were all persons aged 15 years and over who were the usual residents of private dwellings. Residents of non-private dwellings (hospitals, motels and prisons) were excluded from the sample, as were those living in remote or sparsely settled areas.

6. The sample size meant that reliable estimates can be achieved for some statistical regions (Tables 9.1 and 9.2 in ABS: 1999b show estimates for Statistical Regions), but geographic areas below this suffer from small sample size. Therefore, small area estimation techniques are needed to calculate estimates for small areas.

2.2 Sample Design

7. The NCSS uses a complicated sampling design, with clustering and different strata, to reduce the cost of running the survey. The effect of the clustering is that households are chosen in groups. The 1998 National Crime and Safety Survey sample is clustered for the practical reason of reducing the costs associated with forming a list of dwellings and of interviewer travel between selected dwellings. The more highly clustered the sample, the cheaper it is to enumerate, but the less reliable the estimates derived from the sample since the sample is confined to a small area and would be unrepresentative of the whole area.

8. Stratification improves the accuracy of sample estimates by ensuring that different groups in the population are correctly represented. It also makes estimation of results for various geographic areas both easier and more accurate.

9. However, stratification means not all observations have the same probability of being chosen, so OLS regression, which assumes random sampling, may not lead to Best Linear Unbiased (BLU) estimates.

10. Options for taking into account survey design are:

- Ignore the effect of clustering, and take into account stratification by including the survey weights in the model. These are available on the unit record file. These can be included in the model using a SAS WEIGHT statement (with the weights adjusted for survey size), or the variables that the weighting depended on can be included in the model.

- Estimate the model taking into account the sample design (both clustering and stratification). This can be done through the SAS SURVEYREG procedure, the WESVAR program, or SAS with a program to calculate variances from replicate weights. WESVAR uses replicate weights to adjust the variance for design effects, and this can also be done with a SAS program. SAS PROC SURVEYREG estimates a variance-covariance matrix using a Taylor series expansion.

11. The first option is fairly easy to implement, and we will run some models of this type to compare to models that take clustering into account.

12. For the second option, to calculate the coefficients, both WESVAR and SAS SURVEYREG use weighted least squares. Estimates for the regression coefficients are calculated as:

$$\hat{\beta} = (X'WX)^{-1}(X'WY)$$

where W is a diagonal matrix constructed from the survey weights (and replicate weights for WESVAR), X is the matrix of independent variables, and Y is the matrix of dependent variables. WESVAR then uses

$X'WX$ for the full sample and each replicate, and solves the equations using a SWEEP algorithm.

13. WESVAR uses replicate weights to estimate the variances of the estimates. This procedure uses 30 different samples from the data to estimate 30 different weights. Analyses is done separately using each of the replicate weights, and again using the overall weight, and the differences between the outcomes can be used to measure the standard error. The variance using WESVAR is estimated as:

$$\hat{V}_b = c\sum_{k=1}^{G}(b_{(k)} - b)(b_{(k)} - b)'$$

where G is the number of replicates, and c is a weight depending on which method for calculating the replicate weights has been used.

14. To estimate the variance, SAS PROC SURVEYREG uses a Taylor series expansion method to estimate the covariance-variance matrix, and therefore does not need the replicate weights. More on the Taylor Expansion can be found in Appendix 1.

15. PROC SURVEYREG does need the stratum sampling rates or a dataset with the stratification variables and the population totals to estimate the variances. These are not available for the small area model. They are available for the propensity to report model, but PROC SURVEYREG does not allow estimation of logistic regression models. This means PROC SURVEYREG cannot be used for either of our models.

16. SAS can be used to calculate the regression coefficients using WLS, with the estimates being increased to population estimates using the survey weights, and the weights for the WLS being 1/SE. The variances can then be estimated in SAS using the replicate weights. The variance would be calculated as:

$$\left(\sum_{n=1}^{NumReps} (Replicate_n - Est)^2 \right) * (Numreps - 1) / NumReps$$

This means all estimation is in SAS, rather than using another program (WESVAR).

17. We would like to estimate the coefficients and variances using a number of different methods, to give us some idea of the effect each method has on the final results. The methods proposed for each model are described in the sections on estimating the models.

Key Issue - Survey design

Analysis Branch is likely to confront the issue of how best to take into account survey design in other projects which record unit survey data.

A. In what ways can complex survey design affect modelling and other analyses?
B. Is the effect of sample design on econometric models significant?
C. Have we chosen the best methods to take it into account?
D. Has there been any work done on comparing models which take the different varieties of survey design into account and models that don't?

2.3 'Not stated's

18. The other issue common to both projects relates to the 'not stated's on the file. These are items on the survey form where no answers were provided by respondents. There is no knowledge as to what the actual values could be. They only affect some (not all) variables on the file, and most of the variables affected will not be used in the proposed models. The highest proportion was for the "Whether security measures added in the last 12 months" question, where the not stated response was 3.4% of the total response. For the rest, the 'not stated's constitute between 0.5% and 1% of the total response.

19. For a regression model, there are four ways to treat 'not stated's:

- Leave them in as a separate category and in the totals;
- Take them out, reducing the totals;
- Pro rata them out to the other categories, or separately within imputation groups; or
- Model the Stateds and 'not stated's separately.

20. Appendix 2 shows a list of variables affected by a 'not stated' classification; and the extent of the 'not stated's. Because (a) the effect of the 'not stated's is not likely to be large because they form a small proportion of total response, and (b) the 'not stated's only exist for variables that may not be used in the modelling, it is proposed that they are left in the data file as a separate category, and no attempt is made to pro-rata or remove them. Any totals are left with the 'not stated's in. Basing the model on proportions calculated using a total with 'not stated's out is the same as pro-rataing the 'not stated's, which assumes a distribution for the 'not stated's that may not be correct.

Key Issue - How to deal with 'not stated's

Have we used the most appropriate way of dealing with the 'not stated's?

3. Small Area Estimates of victimisation

21. The aim of this project is to derive estimates for crime rates for areas smaller than can be supported by the 1998 NCSS.

3.1 Small area estimation techniques

22. There are a number of techniques under the broad heading small area estimation. These include synthetic estimation, simple empirical and hierarchical Bayes; and Empirical Best Linear Unbiased Prediction (see Ghosh, M & Rao, J: 1994 for a summary of these techniques).

23. We are using a a synthetic estimation technique, because the broad technique (regression modelling) is well known, and is understood in the criminological field. Broadly synthetic estimation calculates relationships from aggregate survey data and applies them to small area Census data. The relationships can be simple ratios, or regression coefficients. Examples of synthetic estimation are in Purcell, N & Linacre, S: 1976; Ghosh, M & Rao, J: 1994; and Stasny, E, Goel, P & Rumsey, D: 1991.

24. The disadvantage with synthetic estimation is that it assumes that the victimisation rates for the chosen socio-demographic groups from the survey data are the same as the victimisation rate for that group at the local level. The regression modelling, by plotting a line through the independent variable observations, is averaging the independent variables across these regions. The slope of these lines (the coefficients) are then being used to predict the smaller areas. The estimates for the smaller areas are likely to break down if the small area does not have similar coefficients for the independent variables to the larger area. This could happen if there is some other characteristic of the small area which the model does not take into account. For instance, the synthetic estimate may be a 10% Motor Vehicle Theft rate on Cocos Island, given the socio-demographic composition of the population. However, the true rate is 0%, because there is no market for stolen cars; and there are very few cars on the Island. This could be taken into account by including the number of motor vehicles in the model.

Key Issue - Choice of small area estimation technique

Is synthetic estimation using a regression model the most appropriate small area estimation technique given the data available and the target audience for the small area estimates?

3.2 Description of the approach to the estimation of small area crime victimisation rates

25. The approach adopted for this study has six main steps:

Step 1 - Choice of levels of geography to use in the study

Step 2 - Choice and construction of measures of victimisation

Step 3 - Merge estimates of crime victimisation from NCSS 1998 with Census 1996

Step 4 - Estimate the regression equation (at ABS Statistical Subdivision) using NCSS 1998 and Census 1996

Step 5 - Apply the model estimated in Step 4 to derive small area estimates of crime prevalence and incidence rates

Step 6 - Validate the small area estimates

The rest of the section discusses each one of these steps in turn.

Step 1 - Choice of levels of geography to use in the study

26. Figure 1 shows the Australian Standard Geographical Classification Areas (ASGC). The geography we are using from the NCSS is the statistical region. Estimates of crime by statistical region are provided in ABS: 1999b. The Census data allows one to apply the models at the Census Collection District (CD) geographic area. However, this level of geography will provide unreliable estimates because local effects are not being taken into account in the model. In this study:

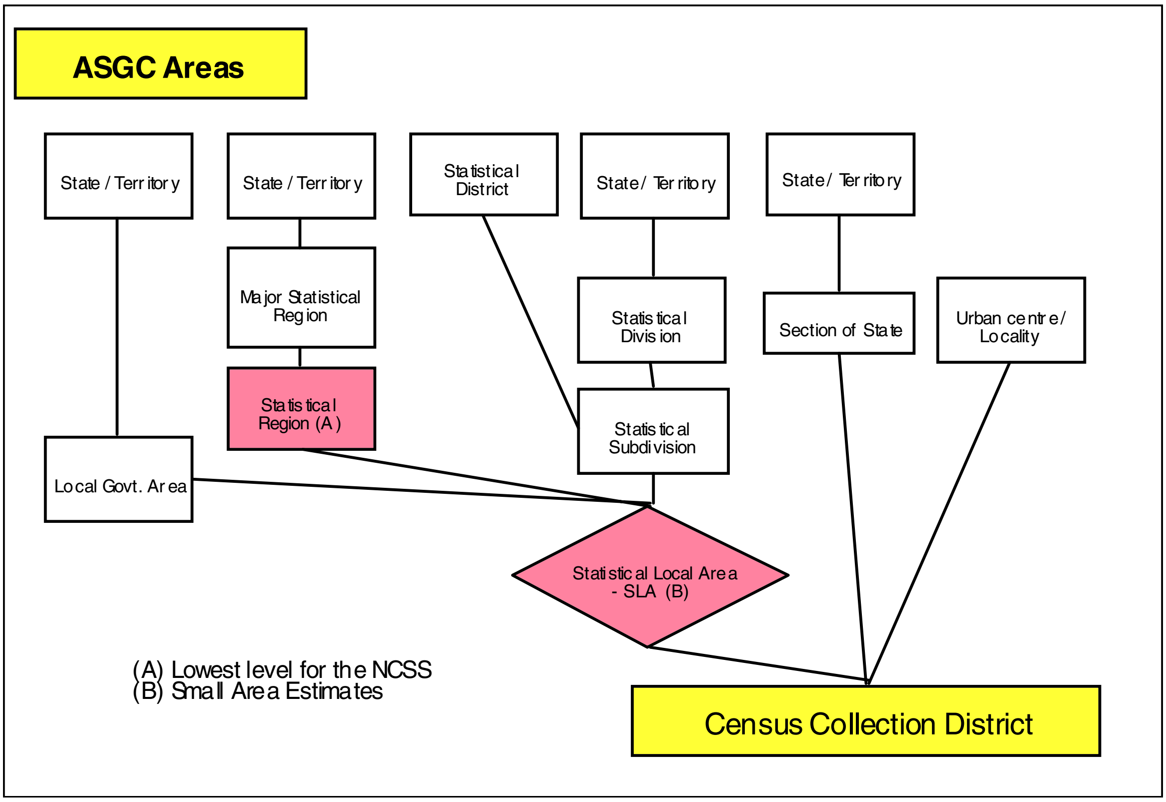
- the regression equation is estimated at the Statistical Region; but
- the small area estimates are derived at Statistical Local Area level.

27. Both these classifications cover the whole of Australia.

28. The statistical region level was chosen for the estimation of the model because most of the estimates at this level are reliable. Those which are not will be taken into account in the implicit averaging done by the weighted least squares regression model. Because of the stratification, weigthted least squares is already being used in the estimation of the model. To get the best estimates taking into account the standard errors, the weight should be 1/SE, so those areas with a high SE get a lower weight. The stratification uses the survey weight for the Weighted Leat Squares estimation. There is some similarity between these two, and we think they are similar enough to use the survey weight as a weight to take into account the standard errors of some areas. A graph of the survey weight and 1/SE is shown in Appendix 3.

29. The Statistical Local Area geography is small enough to cover suburbs in urban areas, but is not so small that the estimates of crime are localised. It is a geography that practitioners in the peer review group suggested would be useful to have estimates for.

Figure 1:
Australian Standard Geographical Classification Areas (ASGC)



Step 2 - Choice and construction of measures of victimisation

30. Let H_i = the total number of households in region i;

P_i = the total number of persons in region i who are at least 15 years old;

$$\delta_{hij} = 1 \text{ } if \text{ } household \text{ } h \text{ } in \text{ } region \text{ } i \text{ } is \text{ } (for \text{ } at \text{ } least \text{ } 1 \text{ } time) \text{ } a \text{ } victim \text{ } of \text{ } crime \text{ } j,$$
$$= 0, \text{ } otherwise.$$
$$j = \{Break \text{ } in, Attempted \text{ } break \text{ } in, Motor \text{ } vehicle \text{ } theft\}$$
$$v_{pik} = 1 \text{ } if \text{ } person \text{ } p \text{ } in \text{ } region \text{ } i \text{ } is \text{ } (for \text{ } at \text{ } least \text{ } 1 \text{ } time) \text{ } a \text{ } victim \text{ } of \text{ } crime \text{ } k,$$
$$= 0, \text{ } otherwise.$$
$$k = \{Robbery, Assault, \text{ } Sexual \text{ } assault\}$$

n_{rl} is the number of times crime l (which could be household or personal) has occurred to victim r

31. In the following equations, prevalence is the number of the relevant population that have been a victim of a given offence <u>at least once</u> in the reference period. Incidence is the total number of incidents of the offence that occurred in the reference period. As some victims experience repeated incidents of victimisation, incidence numbers are typically higher than prevalence numbers.

<u>Household crime prevalence rates</u>

32. The prevalence rate for household crime i is denoted by $\pi_{ij}^{(h)}$ where

$\pi_{ij}^{(h)}$ ={ the number of households which (over the last 12 months, for at least one time) have been victims of household crime j in region i)/(the total number of households in region i}* 100;

$$= (\sum_{h=1}^{H_i} \delta_{hij} / H_i) * 100 \quad ------(1a)$$

<u>Personal crime prevalence rates</u>

33. The prevalence rate for personal crime k is denoted by $\pi_{ik}^{(p)}$ where

$\pi_{ik}^{(p)}$ ={ the number of persons who (over the last 12 months, for at least one time) have been victims to personal crime k in region i)/(the total number of persons in region i}* 100;

$$= (\sum_{p=1}^{P_i} v_{pik} / p_i) * 100 \quad ------(1b)$$

Household crime incidence rates

34. The incidence rate for household crime i is denoted by $\lambda_{ij}^{(h)}$ where

$\lambda_{ij}^{(h)}$ ={ the total number of households which (over the last 12 months) have been victims to crime j in region i)/(the total number of households in region i}* 100;

$$=(\sum_{h=1}^{H_i}\delta_{hij} * n_{hj} / H_i)*100 \quad ------(2a)$$

Personal crime incidence rates

35. The incidence rate for personal crime k is denoted by $\lambda_{ik}^{(p)}$ where

$\lambda_{ik}^{(p)}$ ={ the total number of persons who (over the last 12 months) have been victims to personal crime k in region i)/(the total number of persons in region i}* 100;

$$=(\sum_{p=1}^{Pi}\nu_{pik} n_{pk} / P_i)*100 \quad ------(2b)$$

36. Equations (1a), (1b), (2a) and (2b) express the victimisation rates as percentages. These rates are derived from NCSS 1998. Percentages are used because they are easier to model than numbers of crime, which, because they are count data, follow a Poisson distribution rather than a normal distribution, so OLS regression is not appropriate.

Step 3 - Merge estimates of crime victimisation from NCSS 1998 with Census 1996

37. In this step, for each crime j, the study proposes to create matrices of the following form:

$$\begin{array}{l}
\pi_{j1}\ \ \lambda_{j1};\ Z_{11}\ \ Z_{21}\ \ldots\ Z_{T1}\ ;Y_{11}\ldots Y_{T1};X_{11}\ \ldots\ X_{T1} \\
\pi_{j2}\ \ \lambda_{j2};\ Z_{12}\ \ Z_{22}\ \ldots\ Z_{T2}\ ;Y_{12}\ldots Y_{T2};X_{12}\ \ldots\ X_{T2} \\
. \\
. \\
\pi_{jJ}\ \ \lambda_{jJ};\ Z_{1J}\ \ Z_{2J}\ \ldots\ Z_{TJ};\ Y_{1T}\ldots Y_{TJ};X_{1T}\ \ldots\ X_{TJ}
\end{array} \quad -------(3)$$

38. The data matrix in equation (3) could be for either a personal crime or a property crime. The matrix comprises four main parts.

Estimates of crime rates

39. The data in first two columns would be generated from step 2 above and gives estimates of crime rates from NCSS 1998 at the Statistical Division level. These variables are taken from the 1998 National Crime and Safety Survey. The base is therefore persons over the age of 15 living in private dwellings. The denominator for the rates is the relevant population for the survey (persons aged 15 and over), as described in Section 2 above.

Socio-demographic characteristics of residents from Census 1996

40. The Y matrix in (3) introduces the person-related attributes corresponding to the different regions as extracted from 1996 census. These attributes include income, unemployment, aboriginality, etc. The regression model is measuring the association between this population and victimisation. It does not follow that a significant association indicates that group is more likely to be victims of crime; they may be the offenders; or they may just always live in certain high or low crime areas.

41. Appendix 4 shows the list of variables currently considered for inclusion in the Y matrix in (3). All variables are expressed as the proportion of people in the sub-group, except income. This is expressed in three ways:

1. Median Income;
2. Lowest quartile of income, for a measure of poverty; and
3. Skewness, as a measure of the relative differences in income. The skewness measure used is $Skewness = \dfrac{\dfrac{N^2}{(N-1)/(N-2)} * \sum((Xi-\bar{X})^3)/N}{\sigma^3}$. This is a standard measure given by the program we are using to extract the Census data.

Geographic Region Characteristics from Census 1996

42. The data in the columns of the Z matrix would be derived from 1996 Census and relate to variables associated with the different regions. They measure region characteristics like population density, dwelling tenure, and dwelling types. A full description of the variables is given in Appendix 4. Again, the denominator for any rates is the population aged 15 and over. The reason for this is that the NCSS does not include victims under 15 years; and offenders aged under 15 in most of the categories collected by the Crime and Safety Survey are rare.

Socio-demographic characteristics of victims from NCSS 1998

43. It could be theorised that victim characteristics are an important determinant of victimisation for some crimes. For example, in the case of assault, it may be that higher victimisation is associated with certain characteristics of the victim. In (3) this data set is represented by the matrix X.

44. All victim characteristics are expressed as rates, with the denominator for the victimisation data being the survey weighted population, which was persons who were at least 15 years of age living in private dwellings in Australia in April 1998.

45. The victimisation characteristics are from the National Crime and Safety Survey. Including victimisation characteristics will mean having to extract crosstabs from the NCSS, which will affect the reliability of the survey estimates. Thus a model that uses the X matrix is likely to be less reliable than a model using the socio-demographic characteristics Y which are extracted from Census 1996. Because of this, the first model only includes the Y characteristics, with the X characteristics will be added in later versions of the model. A full description of the variables is given in Appendix 4.

Step 4 - Estimate the regression equation (at ABS Statistical Subdivision) using NCSS 1998 and Census 1996

46. In this step the study estimates 16 regression equations using data from NCSS at the Statistical Subdivision level and the corresponding 1996 Census data for selected variables. The 16 regression equations estimated are:

(i) All household crime - prevalence rate
(ii) All household crime - incidence rate
(iii) Household crime - Break in - prevalence rate
(iv) Household crime - Break in - incidence rate
(v) Household crime - Attempted Break in - prevalence rate
(vi) Household crime - Attempted Break in - incidence rate
(vii) Household crime - Motor vehicle theft - prevalence rate
(viii) Household crime - Motor vehicle theft - incidence rate

(ix) All personal crime - prevalence rate
(x) All personal crime - incidence rate
(xi) Personal crime - Robbery - prevalence rate
(xii) Personal crime - Robbery - incidence rate
(xiii) Personal crime - Assault- prevalence rate
(xiv) Personal crime - Assault- incidence rate
(xv) Personal crime - Sexual Assault- prevalence rate
(xvi) Personal crime - Sexual Assault- incidence rate

47. The estimated equations are of the following form:

Household crime j prevalence rates

$$\pi_j^{(h)} = a_j + d_{j1}\ Z_1 + d_{j2}\, Z_2 + \ ...\ + d_{jT}\ \ Z_T + e_{\ k1} Y_1\ + e_{\ k2}\, Y_2\ + ... + e_{\ kT}\ Y_T + f_{\ k1} X_1\ + f_{\ k2}\, X_2\ + ... + f_{\ kT}\ X_T + \varepsilon_{ij}\ \ -----(4a)$$

Household crime j incidence rates

$$\lambda_{j2}^{(s)} = a_j^* + d_{j1}^*\ Z_1 + d_{j2}^* Z_2\ +\ ...\ + d_{jT}^* Z_T + e^*_{\ k1} Y_1\ + e^*_{\ k2}\, Y_2\ + ... + e^*_{\ kT}\ Y_T + f^*_{\ k1} X_1\ + f^*_{\ k2}\, X_2\ + ... + f^*_{\ kT}\ X_T + \varepsilon_{ij} -----(4b)$$

Personal crime k prevalence rates

$$\pi_k^{(s)} = \alpha_k + \beta_{k1}\ Z_1 + \beta_{k2}\, Z_2\ +\ ...\ + \beta_{kT}\ \ Z_T + \gamma_{k1} Y_1\ + \gamma_{k2}\, Y_2\ + ... + \gamma_{kT}\ Y_T + \delta_{k1} X_1\ + \delta_{k2}\, X_2\ + ... + \delta_{kT}\ X_T + \varepsilon_{ij}\ -----(4c)$$

Household crime j incidence rates

$$\pi_k^{(s)} = \alpha_k^* + \beta^*_{k1} Z_1 + \beta^*_{k2}\, Z_2\ +\ ...\ + \beta^*_{kT}\, Z_T + \gamma^*_{k1} Y_1\ + \gamma^*_{k2}\, Y_2 +\ ... + \gamma^*_{kT}\ Y_T + \delta^*_{\ k1} X_1\ + \delta^*_{\ k2}\, X_2\ + ... + \delta^*_{\ kT}\ X_T + \varepsilon_{ij}\ -----(4d)$$

48. This regression model will be estimated with both SAS and WESVAR, so the results can be compared. Unfortunately, the model cannot be estimated using PROC SURVEYREG in SAS because the stratification variables are not available in the aggregated dataset (the aggregation is larger than the stratification variables). The models we intend to estimate are:

1. Base model with no adjustment for stratification or clustering. The NCSS data will be weighted before estimating the model.
2. Base model estimated using WLS in SAS, with the weight being the adjusted survey weight.
3. Model estimated in WESVAR using the replicate weights.
4. Model estimated in SAS using WLS, with the weight being 1/SE (calculated from the replicate weights) and the survey estimates increased to populaiton estimates using the survey weights.

Multicollinearity

49. A key issue affecting the analysis under this step is multicollinearity. Because of the nature of the independent variables used in this model, there may be some multicollinearity. In particular, there are three measures for the three different types of income. The effect of multicollinearity is that the variance of the regression estimators may be over-estimated, and the regression coefficients may be very sensitive to small data changes (Johnston, J: 1991, p. 248).

50. This study proposes to test for multicollinearity in the following ways:

- look at a simple correlation matrix to identify variables that are collinear; and

- use will be made of the Variance Inflation Factor which measures the extent multicollinearity inflates the variance of the regression estimators. The VIF shows how much higher the variance of the estimator is compared to what the variance would have been if the variable being tested was not collinear with any other variables in the regression. It is calculated as 1/(1-r), where r is the correlation coefficient of the variable in question regressed against all other independent variables. Variables with a VIF of more than 5 (ie, r = 0.8) need to be carefully reconsidered.

51. For an estimation model, the effect of multicollinearity on the estimation side is not likely to be great. Because of the effect on the variances, it is not possible to rely on the t statistics for significance tests. It is also makes problematic sensitivity analysis of the models (because of the effect of multicollinearity on the standard errors and the regression coefficients).

52. To drop the variables causing the multicollinearity may mean some important explanatory variables (from a criminology theory perspective) may need to be removed. We will need to be careful when removing variables to ensure that there is still some proxy for the underlying theory in the model.

53. This study proposes to minimise the amount of multicollinearity in the model by removing variables causing some multicollinearity. However the final decision whether to exclude or include a variable will take into account (a) the VIF, (b) the variable's explanatory power in the model, and (c) the variable's theoretical significance. In particular, we expect to remove some of the measures of income from the model. We also expect some collinearity between qualifications and income.

Key Issue - Multicollinearity

Is the approach we have taken to the issue of multicollinearity appropriate?

Step 5 - Apply the model in Step 4 to derive small area estimates of crime prevalence and incidence rates

54. This step takes the coefficients from the estimated model in step 4 and Census data at a small geographic area to calculate small area victimisation estimates. The constant, which applies to the larger area, will be pro-rated to the smaller areas based on the estimates of crime without the constant. So the overall procedure will involve two steps. Equations (5a) - (5c) show the two steps in the case of household crime j.

Household crime j prevalence rates in small area 'r'

$$\pi_{jr}^{(h)} = Constant\,(\mathrm{r}) + \{\pi_{jr}^{(h)} \big| a_j = 0 \ in \ Equation(4)\} \ ------(5a)$$

$$\{\pi_{jr}^{(h)} \Big| (a_j = 0)\} = \hat{d}_{j1}\, Z_1 + \hat{d}_{j2}\, Z_2 + \ ... \ + \hat{d}_{jR}\, Z_R + \hat{e}_{j1} Y_1 + ... + \hat{e}_{jR} Y_R + \hat{f}_{j1} X_1 + ... + \hat{f}_{jR} X_R \ -----(5b)$$

$$Constant\,(\mathrm{r}) = a_j - (\sum_{r=1}^{R} \pi_{jr}^{(h)} \big| (a_j = 0)) * (H_r / \sum_{r=1}^{R} H_r) \ ----(5c)$$

'R' is the total number of regions from the NCSS survey.

55. Equation (5a) is implemented first for all small areas R in Australia. Equation (5c) is then used to estimate a constant for small area 'r'.

56. Because the regression model coefficients provide estimates of the actual crime rate in the small area, the small area estimates will not add to the survey data. At this point, an adjustment could be made to the small area estimates to force the sum of the smaller areas to equal the larger area survey estimate. Equation (5c) is an example of such an adjustment. Other adjustments that could be made are shown in Stasny, E, Prem, G and Rumsey, D: 1991. These include forcing the sum of small areas to equal the large area, but also minimizing the sum of squared differences between the original and revised small area estimates; or minimizing the squared relative differences between the original and revised small area estimates. Stasny, E, Prem, G and Rumsey, D: 1991 found that a constant scaling factor provided the best results.

57. In the peer review group , there was a split between the methodologists, who didn't want this adjustment made; and the policy analysts and advisers, who did want it made to make it easier to explain the estimates and justify them to critical audiences including Ministers, newspapers, and similar groups.

58. The other option that has been considered is to include into the estimation model the restriction that the sum of the small areas must equal the larger area totals. This would be similar to an income constraint commonly introduced in the estimation of systems of consumer demand equations (ie a consumer's expenditure must not exceed his or her budget). At this stage it is not clear how such an option can be implemented in the case of small area estimates of crime. It is harder to implement, short of estimating using a Bayesian method and using this restriction as a prior.

59. This relationship can also be used to validate the model, before it is used to adjust the model.

60. This method of deriving small area estimates is called synthetic estimation. We are also considering extending this to a composite synthetic approach. This method uses the synthetic estimates with direct estimates, giving more weight to the direct estimates when they are more reliable. The new estimate is calculated as:

$$\hat{Y}_i^C = w_i \hat{Y}_{1i} + (1 - w_i)\hat{Y}_{2i}$$

where $\hat{Y}_{1i}$ is a direct estimator, $\hat{Y}_{2i}$ is a synthetic estimate and w_i is a weight, which can be estimated in a number of ways, but usually use the standard errors in some way, so that direct estimates with a higher standard error get less weight. Ghosh, M & Rao, J (1994) have a discussion of the choice of weights for this model.

Key Issue - Introducing restrictions into the model

Is there an easy way to include a restriction that the sum of the smaller areas equals the larger survey area in the WLS model we are using?

Step 6 - Validate the small area estimates

61. Once estimates for small areas are derived three broad strategies are used to check or validate their accuracy:

i) Internal consistency checks are undertaken

62. One internal consistency check for the model is to use the survey estimates as a check of the model. In theory, the small area estimates should sum to the survey data. In practice, the OLS method is an approximation, so the small areas estimates of victimisation will not necessarily sum to the survey weighted estimates. The difference will give us some idea of how good the small area estimates are.

63. If the small area totals differ from the survey totals (see Step 5 above), the survey estimates can be used to scale the small area estimates so that they add up to the survey estimate. The problem with this is that the model is then being 'forced' to fit the survey data. The estimates are no longer efficient, and the variance of the estimates increases.

ii) Test to see how sensitive the results are to a change in the inputs;

64. This method tests the sensitivity of the model to changes in the data. If the change to the results is large, then the model is sensitive to any changes to the inputs, and may give widely varying results depending on the inputs. This would signal that care should be used when reporting the results of the model. To use this method, one would need to have adjusted for survey design in the calculation of the model, since the standard errors are used for the sensitivity analysis. These will be incorrect if the survey design has not been taken into account in the modelling.

iii) Compare the results to estimates derived from another method, source of data, or model.

65. This method tests the model against some external benchmark. For instance, the estimates from the model could be run against the 1991 Census and the results could be compared to the 1993 Crime and Safety Survey data (which was equivalent to the 1998 Survey for some crimes). If the model estimates are significantly different from the survey estimates, and there is no explanation for the difference, then the model needs to be questioned. The problem with this approach is the reporting rate for the Indigenous indicator on the Census seems to have changed between the 1991 and 1996 Census. Between the 1991 and 1996 Census, there was a 33% increase in indigenous persons. The ABS has calculated that 1/2 of this can be accounted for by births and deaths. The other half is an increased propensity by parents to identify children as indigenous. This means that if indigenous status is used as an estimator for small area crime rates, the coefficient applied in 1996 would be different to that applied in 1991.

66. Another external benchmark is the New South Wales series of Crime Victimisation Surveys. A model could be run against these surveys, and then small area estimates could be derived for the 10 year series. To do this, one would also need the independent variables available on an annual basis. Much of the data may be available annually from other sources, although probably not for small areas. Therefore, this validation technique may be difficult.

67. Again, any large and unexplained movements over the 10 years would suggest a problem with the model.

68. Victoria also conducts a crime victimisation survey, the last one being in 1999. They have published results at a Local Government Area (LGA) level, so one could use the proposed model to estimate victimisation rates by LGA in Victoria, and compare these with the published Victorian estimates from their survey.

69. A further source of data could be Police estimates of crimes reported at a small area level. This method would test both the small area estimates model and the reporting model.

Key Issue - Validating the model

Have we covered all the options for validating the model? Is there anything else we could do?

3.3 Preliminary Results

70. Some results may be available before the June meeting, and if anything useful is available, we will send copies to MAC members.

3.4. Other relevant issues excluded from the study due to unavailability of data

3.4.1 Police as an independent variable

71. The number of police in an area is likely to have an effect on victimisation rates. However the magnitude and direction of this effect is not agreed by experts (see Marvell & Moody: 1996)

72. Unfortunately, for small areas, police presence is very difficult to quantify. States may have police that regularly move between areas. The only information readily available is police expenditure and this is only available at a State level (SCRCSSP: 2000). Even if the data were available obtaining it would be difficult due to its sensitive nature (ABS: 2000). Therefore the proposal is that police is not included as an independent variable in the model.

3.4.2 Drug related offences

73. Many break-ins may be drug related, so areas where there are more drug-dependents may have higher break and enter victimisation rates.

74. The NCSS is a 'victimisation' survey - The fundamental question being asked is 'have you been a victim of crime?'. To study the issue of whether break and enter victimisations are drug related, offender data, and in particular, motivation for the crime, would be required. These data are unavailable in the NCSS and data from Justice System may be difficult to obtain and may not be appropriate for use in this study (ABS: 2000).

75. The NCSS has a neighbourhood perception question:

> 'What are the problems from crime or people creating a public nuisance in your neighbourhood?'

76. This question was asked of all respondents and one possible response specified illegal drugs, and another was break-ins. Because break-ins may be associated with illegal drugs, it is difficult to associate these two perceptions from the survey. Further, the respondent could choose as many problems as they perceived as being a nuisance. This makes it difficult to interpret the priority of the 'perceptions'. The question is also a subjective question - what one person interprets as 'drug problems' may be different from what another perceives as drug problems.

77. Illegal drugs are not perceived as a major problem, ranking as sixth most perceived neighbourhood problem. This places them almost equal with drunkenness and louts/gangs. Because of the problems associated with getting data on drug offenders, this study does not include drug use in the model.

3.4.3 Offenders and victims travelling between areas

78. If a person from a given area is victimised in another area (say, inner city), the NCSS enumerates the victimisation where the person resides. This means that in some cases the small area model which assumes that place of residence is the same as place of victimisation may be erroneous.

79. For better results one would need data on where the victimisation took place and assign it to that given area. From the survey, there is no question that allows us to identify where the crime took place. It is also difficult to measure the extent of the problem of mismatch between the place of residence and the place of victimisation.

80. The NCSS has a question relating to where the incident take place. Possible responses include At home, At place of work or study, In a place of entertainment, etc. Analysis of these responses reveals that 35% of all personal crime occurs at home. It can be assumed that for 35% of cases (total personal crime) this issue is not a problem. For the remaining 65% it may be. Unfortunately, without data on the location of the offence, there is nothing that can be done about this problem, except to recognise it and document it in our results.

4. Propensity to report crime

81. The aim of this project is to identify victim characteristics that affect the decision to report a crime.

4.1 Method

82. The decision to report crime is a binary response variable (0/1). It is therefore proposed that a logit regression model be used to model the relationship between the decision to report and other victim characteristics. This is similar to the method used in other studies of reporting behaviour (see Skogan, W: 1994; Kury, H, Teske, R & Wurger, M: 1999; and Carcach, C: 1997). A bivariate probit model has also been used in the literature (Macdonald, Z: 2001), which also models the Yes/No victimisation variable. The logit regression takes victimisation from the survey.

83. The estimates would be calculated by crime type (personal and property); and by crime, if the data allows it. The dependent variable would be the propensity to report the latest victimisation. There is not enough information on victimisations before the latest one (any victimisation besides the latest has a smaller set of data attached to it in the survey), so initially only the latest victimisation will be considered.

84. The SAS procedure SURVEYREG will not estimate logistic regressions. However, the program WESVAR does, and with replicate weights, we can calculate the standard errors for a WLS estimation.

85. The models we intend to estimate are:

1. Model with no adjustment, and no weighting on the unit record file;

2. Model including the survey weight as a weight in WLS;

3. Model including a survey weight adjusted to the survey population as the weight for WLS;

Model including 1/SE (calculated in SAS using the replicate weights) as the weight for WLS, and the survey estimates increased to populaiton estimates using the survey weights.

Key Issue - Propensity to report method

Is the logit regression model the best model for this problem? Are there any better models that could be considered?

4.2 Data

86. Appendix 5 shows a 'wish list' of data. Problems due to the quality of the data and the relative standard errors may mean that some of the data will be aggregated, or will be included in one model but not in another.

87. Most of the independent variables are binary variables, with the exception being the number of previous victimisations. All the data come from the 1998 National Crime and Safety Survey.

4.3 Preliminary Results

88. Some results may be available before the June meeting, and if anything useful is available, we will send copies to MAC members.

4.4 Validating the model

89. Once estimates for the propensity to report for different groups are derived, how does one check, or validate, them? There are two broad methods:

i) Test to see how sensitive the results are to a change in the inputs; and
ii) Compare the results to estimates derived from another method, source of data, or model.

90. The first method tests the sensitivity of the model to selected data changes. If the change to the results is large, then the model is sensitive to changes in the selected variables, and may give widely varying results depending on the inputs. Note that this validation method requires that the effect of the sample design has been taken into account in the modelling stage, since it uses the standard errors to calculate the sensitivity of the input variable.

91. A test of the reporting behaviour could be to check the reporting propensities against Police estimates of crimes reported. However, there are differences between the NCSS reporting data and Police data. The reporting question on the 1998 National Crime and Safety Survey asked "Did you tell the Police about the most recent incident?". In some cases, someone else (friend or family) will report the offence to police. In other cases, the police will discover the offence themselves; or may not record the offence in their system because it is a minor offence. They also may classify the offence differently from what the victim classifies it as on the NCSS. Therefore, the reported number will not match Police reported statistics (see ABS : 1999a, p. 82 for a comparison of the NCSS and Police data).

92. One could also run the model against the 1993 National Crime and Safety Survey for those crimes that did not have definitional changes between 1993 and 1998, and compare the results.

Key Issue - Validating the propensity to report model
Are there any other ways the model could be validated?

5. Conclusion

93. The 1998 National Crime and Safety Survey has been used in two modelling exercises. The first was to calculate small area estimates of victimisation using victimisation data from the survey and socio-demographic data from the 1996 Census. Because the Census was conducted in August 1996; and the 1998 Survey asked about victimisation over the past year; the data are temporally fairly close.

94. Applying regression analysis to this survey raised the issue of how to adjust for sample design. We propose to deal with this in a number of different ways, and compare the results.

95. The small area estimation model has raised some interesting questions about the accuracy of survey estimates, and how they can be enhanced for small areas.

96. The propensity to report crime model used only the survey data; and used a logit model to estimate reporting behaviour for different groups.

97. While we have not presented any results or conclusions from our models in this paper, if reasonable results are available before the MAC meeting, we will certainly send these out to the members.

Appendix 1
Taylor Expansion Method

The Taylor series expansion can be used to estimate the covariance-variance matrix of the estimated regression coefficients. For a full description of this method, see Fuller: 1975.

For a stratified, clustered sample, observations are represented by a matrix

$$(w, y, X) = (w_{hij}, y_{hij}, x_{hij})$$

where w denotes the sampling weights, y denotes the dependent variable, X denotes the design matrix, h is the stratum number with a total of H strata, i is the cluster number with a total of n_h clusters, j is the unit number within cluster i of stratum h, with a total of m_{hi} units.

The estimate of $\hat{\beta}$ is calculated as

$$\hat{\beta} = (X'WX)^{-1} X'WY$$

Let

$$r = y - X\hat{\beta}$$

The following vectors are calculated:

$$e_{hij} = w_{hij} r_{hij} x_{hij}$$

$$e_{hi.} = \sum_{j=1}^{m_{hi}} e_{hij}$$

$$\overline{e}_{h..} = \frac{1}{n_h} \sum_{i=1}^{n_h} e_{hi}$$

and the p*p matrix

$$G = \frac{n-1}{n-p} \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h - 1} \sum_{i=1}^{n_h} (e_{hi.} - \overline{e}_{h..})'(e_{hi.} - \overline{e}_{h..})$$

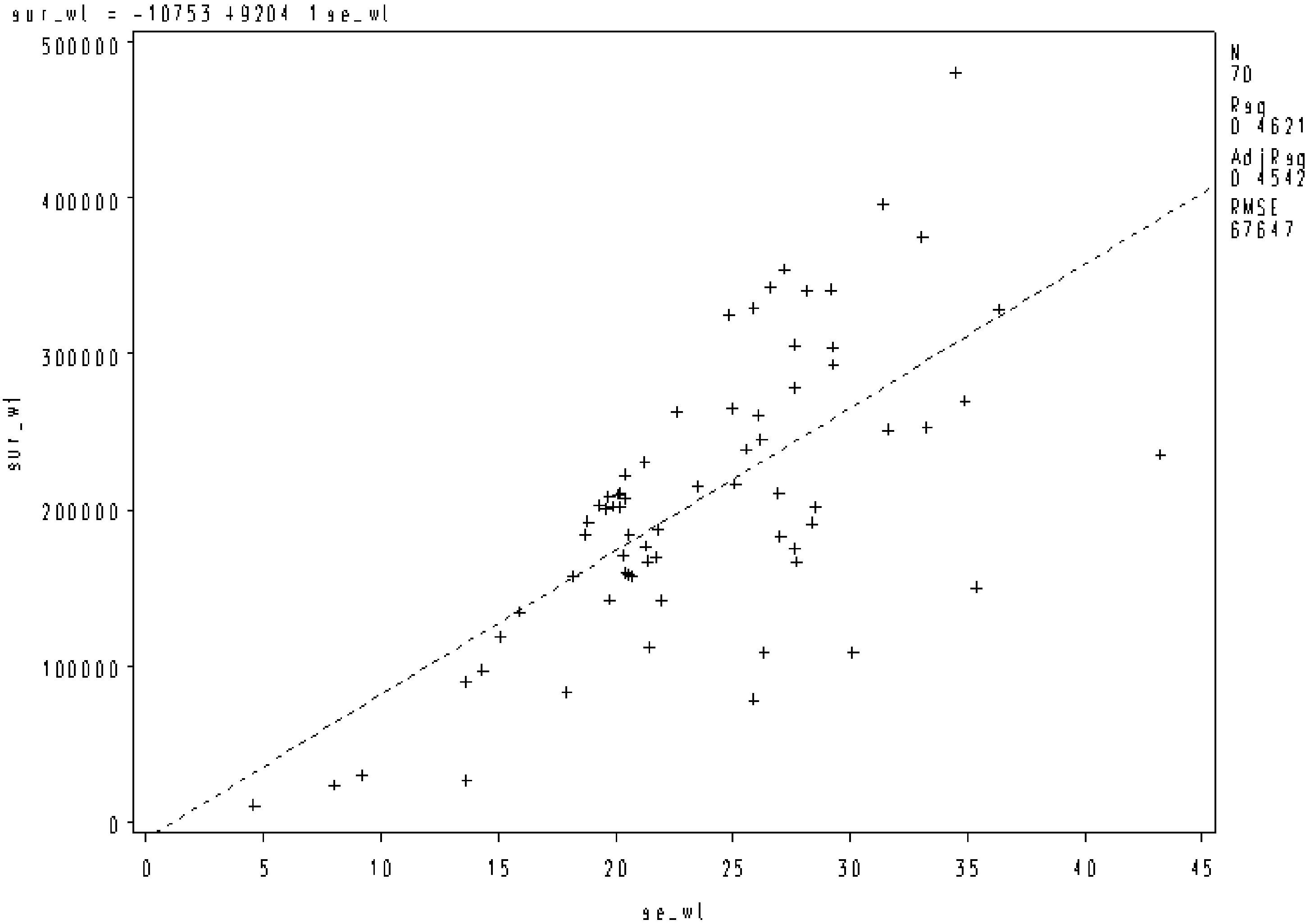
where f_h is the sampling rate for stratum h, and all the other variables are defined above. The covariance matrix of β is then calculated as:

$$\hat{V} = (X'WX)^{-1} G (X'WX)^{-1}$$

Appendix 2
List of variables with a Not Stated classification

Variable	Description	% Not Stated
HOMETYPE	Type of dwelling (separate house, flat, etc)	0.99
VISIBLE	Dwelling visibility (completely seen, partially screened, not visible, etc)	0.39
LOCATED	Dwelling location (near a laneway, on a corner, in a dead end St., etc)	1.20
CARTRAFF	Amount of vehicle traffic (constant, peak hour only, etc)	0.46
PEDESTRF	Pedestrian traffic (constant, peak hours only, etc)	0.73
CLOSETO	Live close to park, school, pub, shops, etc.	3.40
RESIDED	Length of residence at current address (less than 1 yr., 1 to 3 yr., etc)	0.36
OWNDRENT	Dwelling tenure (renting, own, buying, etc)	0.71
HAVEDOG	Whether household has a dog	1.61
DRIVEWAY	Whether there is a car in the driveway (all the time, sometimes, rarely, etc)	1.11
LIGHT	Whether there is outside lighting (street light, sensor light, etc)	1.00
WINDOW	Window security (bar, grills, etc)	1.71
DEADLOK	Whether doors have deadlocks	1.70
SCREEN	Whether have security screen doors	1.19
ALARM	Whether have an alarm system	2.22
SECMSAD	Whether security measures have been added in the last 12 mths	3.40
CRMHRE	Reason not told the police (break and enter)	1.31
ACTIONB	Action of break and enter offender (stole property, damaged property, etc)	0.01

Appendix 3
Graph of survey weight and 1/Standard Error



Appendix 4
Full description of Census data

Variables	Description	Theoretical basis for inclusion in the model
Victim variables (personal crime) from 1998 NCSS (X Matrix)		
Male aged 15 - 25	Male victims aged 15 - 25/all persons aged 15 and over, used for Assault only	Lifestyle theory: Males aged 15 to 25 are more likely to place themselves in situations where they are more likely to become victims of crime.
Female	Female victims/all persons aged 15 and over, used for Assault only	Females are more likely to be victims of personal crime, particularly single mothers.
Divorced/separated	Divorced or separated victims/all persons aged 15 and over, used for Assault and Sexual Assault only	Social disorganisation theory - Divorced or separated persons are more likely to be victims of crime.
Socio-Demographic characteristics from 1996 Census (Y Matrix)		
Household income	Median weekly household income. This is the income for a household, which may not contain a family (eg, group households). Lowest quartile of household income. Skewness measure of the lack of symmetry around the mean.	Income, in particular low income, and relative differences in income are associated with the Economic, Strain and Social disorganisation theories of crime.
Individual income	Median weekly individual income. Lowest quartile of individual income. Skewness measure of the lack of symmetry around the mean.	As above
Family income	Median weekly family income. This is the income for a family unit in a house. Lowest quartile of family income. Skewness measure of the lack of symmetry around the mean.	As above
Poverty	Proportion of families in the area within the lowest quartile of total income from the survey.	Social disorganisation theory - Low income areas are generally associated with more crime.
Indigenous persons	Indigenous persons as a % of the population	Social disorganisation theory - Greater the heterogeneity of a given population, less social bonds formed. Sub-culture theory - Variable included to test the hypothesis that cultural minorities are associated with higher crime.
Australian	Australian born persons of Australian born parents as a % of the population	See note about indigenous persons.
Low fluency in	Persons with no or poor English skills	See note about indigenous persons.

English	as a % of population	
Unemployment	Unemployed persons as a % of the total population aged 15 and over	Strain theory - Unemployment causes financial hardship, and this is a motivation for crime intended to relieve hardship. Economic hardship may also give rise to frustrations that increase the frequency of aggressive acts.
Male population aged 15-25	Male persons aged 15 - 25 as a % of the population	Lifestyle theory: males aged 15 to 25 are more likely to be offenders.
Single parent families	Single parent with one or more dependant children persons as % of all families	Social disorganisation theory - Family disruption is often a primary cause of economic hardship.
No qualifications	Persons with no post-secondary qualifications as a % of the population	This variable is highly linked to unemployment, income and poverty variables. It will be included initially, but may be dropped if multicollinearity is a problem.
University degrees	Persons with post secondary qualifications as a % of the population	As above
Geographic Region characteristics from 1996 Census (Z Matrix)		
Living alone	Households with Person living alone/All households	Theory of guardianship - Persons living alone are easier victims than groups. Household left 'unguarded' for longer periods.
Dwelling tenure	Rented households / All households	Social disorganisation theory - Rented households tend to be more mobile forming less bonds with the community. Rent to a much lesser extent taps into the concept of 'wealth'. Households who own their own home maybe seen as 'wealthier' than those who don't.
Population density	Number of persons per square kilometre	Economic and social disorganisation theories - High density areas have relatively more opportunities for crime, more anonymity and less community bonding.
Dwelling type 1	Households in separate house/All households	Similar to the 'density' variable - see above
Dwelling type 2	Households in flats, units or apartment/All households	See above
Residential mobility 1	Families who changed address during the last year as a % of all families	Social disorganisation theory - People who are more mobile form less bonds with the community.
Residential mobility 5	Families who changed address during the last 5 years as a % of all families	As above.

Appendix 5
Data for the propensity to report model

Variables	Variable type	Theory
Seriousness Variables		As an offence becomes more serious, reporting rates will increase.
Whether something was stolen	Binary	
Whether victim was physically attacked	Binary	
Whether victim was threatened with violence	Binary	
Whether a weapon was used	Binary	
Whether the victim was injured	Binary	
Whether there was more than 1 offender	Binary	
Whether victim lived alone	Binary	Lone victims are more likely to report (security issues)
Relationship Variables (only available for Assault)		The closer the relationship between the offender and victim, the less likely the victim is to report the offence.
Victim was married to offender	Binary	
Victim was boyfriend/girlfriend	Binary	
Victim was ex-partner	Binary	
Victim was ex boyfriend/girlfriend	Binary	
Victim was partner or ex-partner AND offence occurred at home	Binary	Incidents occurring at home involving current family members are less likely to be reported.
Crime Variables		
Number of times the offence occurred in the last 12 months.	Discrete	Repeat victims are less likely to report repeat victimisations.
Individual Socio-Demographic Variables		Some socioeconomic groups do not have a good relationship with the Police, and are less inclined to report victimisation.
Sex	Binary	Females are more likely to report offences to the police.
Sex * Victim knew offender	Binary	Females are less likely to report if offender known.
Age 15 - 24	Binary	Young people are less inclined to report
Age 65 and over	Binary	Older people are more inclined to report
Males aged 15 - 24	Binary	Young Males are less inclined to report.
Unemployed	Binary	Unemployed are less likely to report.
Household Socio-Demographic Variables		
Whether victim was sole parent with	Binary	Sole parents with young children more

children under 15.		likely to report (security issues).
Whether been at current address 5 years or more.	Binary	Attachment to current area is associated with greater propensity to report.
Whether own home or not	Binary	Attachment to current area is associated with greater propensity to report.
Whether have security devices installed	Binary	Security devices associated with more attachment to house which is associated with higher reporting.
Whether area has perceived problems with property crime, violent crime, drug crime, drunkenness, louts/youths/gangs, prowlers/loiterers, vandalism, dangerous noisy driving, etc.	Binary	Households in criminogenic areas have lower propensity to report crime. Question used is the 'Perceptions of area' question from the NCSS.
Whether household in high risk area (100 metres from a school, railway station, pub, hotel, etc.	Binary	Households in these areas have a lower propensity to report crime. These data come from questions in the NCSS.

Bibliography

Australian Bureau of Statistics (ABS), (1996). *Women's Safety Survey Australia 1996*, Publication 4128.0, ABS, December 1996.

Australian Bureau of Statistics (ABS), (1999a). *Crime and Safety Australia*, Publication 4509.0, ABS, August 1999.

Australian Bureau of Statistics (ABS), (1999b). *Crime and Safety : Supplementary National and Standard Tables*, Publication 4509.0.40.001, ABS, October 1999.

Australian Bureau of Statistics (ABS), (2000). *Analysis Branch, Discussion Paper: Experimental Output Measures for the Australian Justice Sector*, ABS, November 2000.

Becker, G. (1968). 'Crime and punishment: An Economic Approach', *Journal of Political Economy,* Vol.76, pp. 169-217

Carcach, C. (1997). *Reporting crime to the Police*, Trends and Issues in Criminal Justice No. 68, Australian Institute of Criminology

Dubin & Rivers (1989). 'Selection bias in linear regression, Logit and Probit models.' *Sociological Methods and Research*, V 18, Nos 2 & 3, pp. 360 - 390

Fuller, W. (1975). 'Regression analysis for sample survey', *Sankhya: The Indian Journal of Statistics*, Vol. 37, Pt. 3, Series C, pp. 117 - 132

Ghosh, M. & Rao, J. (1994). 'Small area estimation: an appraisal', *Statistical Science*, Vol. 9, No. 1, February 1994, p. 55 - 93

Heckman, J. J. (1979). 'Sample selection bias as a specification error.' *Econometrica* 47(1), pp. 153 - 162

Johnston, J (1991). *Econometric Methods*, McGraw-Hill, Singapore.

Kelly, M. (2000). 'Inequality and Crime', *The Review of Economics and Crime*, November 2000, 82(4), pp. 530 - 539

Kury, H., Teske, R. & Wurger, M. (1999). 'Reporting of crime to the police in the Federal Republic of Germany : A comparison of the old and new lands", *Justice Quarterly*, Vol 16, No. 1, March 1999, pp. 123 - 151

Marvell & Moody (1996). 'Specification problems, Police Levels and Crime Rates', *Criminology*, Vol. 34, No. 4

Macdonald, Z. (2001). 'Revisiting the dark figure : A Microeconometric Analysis of the Under-reporting of Property Crime and Its Implications', *British Journal of Criminology*, Vol. 41, No. 1, Winter 2001, pp. 127 - 149

Purcell, N. and Linacre, S., (1976). *Techniques for the Estimation of Small Area Characteristics*, Third Australian Statistical Conference, 18-20 August 1976

SAS (1990). *SAS/STAT User Guide*, SAS Institute, USA.

Scott, A. J. (1986). 'Logistic regression with survey data', *ASA Proceedings of the Survey Research Methods Section*, 1986.

SCRCSSP (Steering Committee for the Review of the Commonwealth/State Service Provision), (2000). *Report on Government Services 2000*, AusInfo, Canberra.

Skogan, W (1994). *Contacts between the Police and Public: findings from the 1992 British Crime Survey*, Home Office Research Studies, HMSO, 1994

Smith, D. & Jarjoura, R. (1989). Household characteristics, neighbourhood composition and victimisation risk', *Social Forces*, Vol. 68, No. 2, December 1989

Stasny, E, Goel, P & Rumsey, D (1991). 'County estimates of Wheat production', *Survey Methodology*, Vol. 17, No. 2, December 1991, pp. 211 - 225