ANALYSIS OF COMPLEX SURVEYS

A thesis presented in partial fulfillment
of the requirements for the degree of
Masterate in Science
in Statistics
at Massey University

JANE YOUNG

May 1997

ACKNOWLEDGEMENTS

I can't believe that it is finally coming to an end!! Of course, the first person I must thank is Associate Professor Stephen Haslett. Thank you Steve for that endless supply of time, effort, supervision, guidance, advice, wisdom and support over the year (oh, and patience!!). I have learnt so much from Steve, maybe because he never wanted to tell me the answers. '...I could tell you the answer but you would not learn anything...' was one of his favourite sayings I seem to recall. His endless supply of knowledge never ceased to amaze me.

Thanks also goes out to Dr Siva Ganesh for the use of his most beloved PC. Without it I fear that my computer analyses would have taken me another year to do (when they finally let me have a bigger and faster machine!). Also thanks for his expertise in SAS and multivariate statistics, which came in handy at almost the right times.

I also must thank Mr Alasdair Noble for the time and effort which he put into reading what I thought was my final draft. Even though it meant more work and sleepless nights, his comments and suggestions were invaluable.

The staff in the Department of Statistics have been a wonderful support throughout my studies at Massey University. Even though I am absolutely sick and tired of studying, I know I will miss this place when I finally leave.

Lastly, I would like to thank all of those friends and family of mine that have supported me and kept asking me 'when are you going to finish?'. I'm not sure how many of them actually knew what I was studying though.

WHEW!! Well, all I can say is that the light at the end of the tunnel is no longer an oncoming train!

ABSTRACT

Complex surveys are surveys which involve a survey design other than simple random sampling. In practice sample surveys require a complex design due to many factors such as cost, time and the nature of the population.

Standard statistical methods such as linear regression, contingency tables and multivariate analyses are based on data which are independently and identically distributed (IID). That is, the data is assumed to have been selected by a simple random sampling design. The assumptions underlying standard statistical methods are generally not met when the data is from a complex design. A measure of the efficiency of a design was found by the ratio of the variance of the actual design over the variance of a simple random sample (of the same sample size). This is known as the design effect (deff). There are two forms of design effects; one proposed by Kish (1965) and another termed the misspecification effect (meff) by Skinner et al. (1989). Throughout the thesis, the design effect referred to is Skinner et al. (1989)'s misspecification effect. Cluster sampling generally yields a deff greater than one and stratified samples yields a deff less than one.

Some researchers have adopted a model based approach for parameter estimation rather than the traditional design based approach. The model based approach is one which each possible respondent has a distribution of possible values, often leading to the equivalent of an infinite background population, called the superpopulation. Both approaches are discussed throughout the thesis.

Most of the standard computing packages available have been developed for simple random sample data. Specialized packages are needed to analyse complex survey data correctly. PC CARP and SUDAAN are two such packages. Three examples of statistical analyses on complex sample surveys were explored using the specialized statistical packages. The output from these packages were compared to a standard statistical package, The SAS System. It was found that although SAS produced the correct estimates, the standard errors were much smaller than those from SUDAAN. This led, in regression for example, to a much higher number of variables appearing to be significant when they were not.

The examples illustrated the consequences of using a standard statistical package on complex data. Statisticians have long argued the need for appropriate statistics for complex surveys.

CONTENTS

CHAPTER THREE

CHAPTER FOUR

APPENDIX

LIST OF FIGURES AND TABLES

PREFACE

This thesis covers some standard methods for analysing complex surveys. Chapter one discusses some common sampling designs and provides general theoretical background to the problem of analysing complex survey data.

Data from two survey questionnaires involving some complex design are used throughout the thesis to illustrate statistical methods and to provide some actual survey data for analyses. The questionnaires used are presented in chapter two.

Chapter three discusses the effect of a complex design in regression analysis. A brief overview of the traditional regression methods is given and this leads to the effect of a complex design on the regression parameters. To adjust for the survey design, alternatives to the ordinary least squares estimator are considered.

Another common statistical technique in sample surveys is the use of contingency tables for categorical data. Analysis of contingency tables in chapter four includes various chi-square test statistics, the effect of complex designs on the standard chi-square statistics and the development of appropriate adjustments.

In chapter five, the focus is on multivariate data analysis. The effect of complex designs on the covariance matrix and different estimates of the covariance matrix is considered under the design and model based approaches. In particular, principal components is discussed as the main multivariate technique.

Some computing examples based on ‘real life’ sample surveys are in chapter six. The computing programs used for the analysis of a complex survey are PC CARP and SUDAAN. The outputs from these packages are compared with a package that does not adjust for complex surveys; this package will be SAS.

The final chapter includes a summary and conclusions.

CHAPTER ONE

INTRODUCTION

1.1 SURVEY SAMPLING

Survey sampling is a form of selecting a part of some finite population so that one may obtain statistically based estimates. These estimates are used to infer something about the finite population. A sample survey is often less time consuming and costs less than observing every unit in the population. For example, a nutritionist may be interested in finding out the number of primary school children who eat breakfast but has a time or cost constraint which does not enable her to ask every primary school child. By selecting a number of school children across a number of primary schools, this would take less time and would give the nutritionist an estimate for all primary school children. Sometimes it may not be practical to observe every unit, e.g., when observing the total number of a rare bird in a particular forest, parts of the forest may not be accessible due to heavy undergrowth, fallen trees etc. and hence a sample is observed that will give an estimate of the population total. Such estimates are used routinely in economic and social policy and in scientific research.

Around the 1950's several books were published on important developments in sampling theory (Binder, Kovar, Kumar, Paton and van

Baaren, 1987). Of these, Cochran (1963) has outlined the principal steps in a sample survey.

Some of these are:

- Defining the target population to be sampled, i.e., the population from which information is wanted.
- Constructing a sampling frame. The frame is a list that accounts for the whole of the survey population from which the sample is drawn, and should match the target population as closely as possible.
- Selection of the sample. The particular random sampling scheme used depends on many factors such as cost and time.
- Degree of precision: how accurate an estimate is required.
- Information for future surveys. After a sample has been collected and analysed, information gained can act as a guide to designing future surveys.

In a finite population, which we shall call U, it is possible to identify each and every unit. These units can be numbered, say 1 to N. Associated with each unit is a variable of interest which is often called the y variable. There are usually more than one y variable collected in any sample survey. This y variable is a fixed variable but not necessarily known. For example, let the finite population be all people of New Zealand (each person labelled starting from 1 through to N) and the variable of interest could be an individuals income for the year 1996. In the traditional design based inference, this y variable for a person, say y_k, (k being the kth unit of the population), is unknown but fixed.

Often, there exists auxiliary variables, denoted by z. These auxiliary variables are variables that contain information about the finite population and are available to researchers before sample selection. For example, an auxiliary variable, z, can be the age of all the people in New Zealand.

Generally one sample, s, of size n (out of N) units, is to be selected from a total of S possible samples from the finite population, U. Each sample has a non zero probability of being selected called the sample selection probability and denoted by p(s). Each unit in the population also has a non zero (positive) probability of being selected. The probability associated with each unit is called the inclusion probability, i.e., the probability of a unit being included in a sample, s, and is denoted by π_i (i = 1,...,N). Such sampling is called probability sampling. Given there is some random mechanism to select these units, then it is also known as probability random sampling.

By observing a sample, s, on the y variables, it is possible to estimate the population parameter of interest, θ. Say the population parameter of interest is the total. The sample s will give an estimate (some function of the y variables) of the population total. Let t(y) denote the estimate. Another sample from the same population, say s_a, will give a different estimate, $t_a(y)$ and yet another sample, say s_b, will give yet another estimate, $t_b(y)$. Since there can be many samples from the same finite population, there can be many estimates. When defined over the whole set of samples, S, the estimates are random variables (as it can take values t(y), $t_a(y)$, $t_b(y)$ and so on). Let T(y) denote the class of estimates. This T(y) is called an estimator. A 'good' estimator is one which varies little around the unknown parameter, θ. This is usually measured by the variance of the estimator provided T(y) is an unbiased estimate of θ, and by the estimate's mean square error if there is bias. For unbiasedness, the expected value of T(y) taken over all possible samples, is equal to θ.

The way in which the units are selected to draw a sample is referred to as a sampling scheme (Lehtonen and Pahkinen, 1995) or a sampling design (Thompson, 1992). Different probability based sampling designs use random selection from the population in different ways. The population structure usually dictates a particular sampling design. Choosing an inefficient random sampling design for a particular population structure can give less accurate results for a fixed cost. Following are some common sampling designs.

1.2 SAMPLING DESIGNS

1.2.1 SIMPLE RANDOM SAMPLING

This is the simplest of the random sampling designs. Simple random sampling is when a sample of size n is drawn from a finite population of N units. Each unit has an equal probability of being selected (inclusion probability), and each unit is randomly selected from the population, in such a way that each pair of units also have equal chances of both being in the sample (joint inclusion probability).

In some situations, the unit is replaced back into the population and the unit may be re-selected. This is sampling with replacement (WR). Usually, the n units are drawn out one at a time and are not replaced, for example, the game lotto where the first ball is chosen out of a total of 40 and the second is chosen out of 39 and so on until all six balls are chosen. This is referred to as sampling without replacement (WOR). However, a finite population correction (fpc) is needed as a correction factor for the variance of the estimator of a mean or total in the case of WOR. The fpc is given by (1-f) where f = n/N. Providing N is large and n is small, the fpc is approximately 1. Under these special conditions the fpc can be omitted from analyses.

The number of distinct samples of size n from a population of size N is:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}, \tag{1.1}$$

and the selection probability for a sample, s, is $1 \Big/ \binom{N}{n}$. This selection probability is the same for all samples of sample size equal to n.

The mean of the sample,

$$\begin{aligned} \bar{y}_{srs} &= \frac{y_1 + y_2 + \ldots + y_n}{n} \\ &= \sum_{i=1}^{n} \frac{y_i}{n} \end{aligned} \tag{1.2}$$

and the sample estimate of the variance is given by,

$$s^2 = \sum_{i=1}^{n} \frac{(y_i - \bar{y}_{srs})^2}{n-1}. \tag{1.3}$$

The sample estimate of the variance of sample mean is

$$V(\bar{y}_{srs}) = \frac{s^2}{n}. \tag{1.4}$$

For WOR sampling, summations over the n units in the sample now involve distinct units (i.e., no repeats), formulae (1.2) and (1.3) still apply and equation (1.4) becomes

$$V(\bar{y}_{srs}) = \frac{s^2}{n}(1-f)\,. \tag{1.5}$$

Advantages of simple random sampling is that it is mathematically simple and that most 'mainstream' statistical analysis techniques assume simple random sampling of the units (Kish, 1965).

1.2.2 STRATIFIED SAMPLING

In stratified sampling the population is divided up into subpopulations called strata. The strata are mutually exclusive (i.e., no unit belongs to more than one stratum), exhaustive (i.e., all units belong to a stratum) and not necessarily the same size. If samples are randomly selected from each strata by simple random sampling, then this is known as stratified random sampling.

Generally each stratum will consist of similar units. The finite population of interest will quite often have some natural stratification such as gender, soil type or city size. Selection of units from one stratum is independent from another. Hence the variances of the estimators for each stratum can be added to give the variance of the estimator for the whole population.

If a sample of size n is to be selected from H strata, and no prior information is available about the strata, then it is reasonable to select an equal number of units from each strata, i.e., $n_h = n/H$, where n_h is the sample size of stratum h. However, in cases where the strata differ in size, it may be more sensible to sample more units from the larger strata. This is done by proportional allocation of the n units to be selected across the H strata, i.e., $n_h = nN_h/N$, where N_h is the total number of units in stratum h.

The sample mean for a general stratified design is given by

$$\bar{y}_{st} = \sum_{h=1}^{H} \frac{N_h}{N} \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \tag{1.6}$$

where $N = N_1 + ... + N_H$ and $n = n_1 + ... + n_h$ is the total number of units in the entire sample. Such an expansion estimator 'expands' the sample values.

The sample variance of the mean is

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{i=1}^{H} N_h (N_h - n_h) \frac{s_h^2}{n_h} \tag{1.7}$$

where $s_h^2 = \dfrac{1}{n_h - 1}\left[\sum_{i=1}^{n_h}\left(y_{hi} - \bar{y}_h\right)^2\right]$ is the estimated variance of the stratum mean in stratum h and $\bar{y}_h = \dfrac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$ is the sample mean in stratum h.

Each strata has a selection probability $p_h(s_h)$ where h = 1,...,H. Since the selection in one strata is independent of another, the total selection probability is

$$p(s) = p_1(s_1) \times p_2(s_2) \times ... \times p_h(s_h). \tag{1.8}$$

The sample selection probabilities differ in stratified sampling from simple random sampling, even for proportional allocation, since including strata gives the sample conditional structure. For example, suppose there was a population of 10 elements from which 2 are to be chosen. The selection probability is $1\Big/\binom{10}{2}$ = 1/45. If the population was to be stratified into two strata, 1 to 5 and 6 to 10, randomly selecting one element from each stratum would give a selection probability for each possible sample of $1\Big/\binom{5}{1} \times 1\Big/\binom{5}{1}$ = 1/25 (using equation (1.8)). Hence the sample selection probabilities will differ, because not all samples possible using simple random sampling can be drawn for a stratified sample, even with proportional allocation.

The main advantage of stratified random sampling is a gain in precision due to a decrease in the variance. Other advantages include sampling convenience (each strata can be treated as a population), and flexibility in forming strata.

1.2.3 CLUSTER SAMPLING

Cluster sampling involves samples from clusters of units from the population. Although the population is divided into groups each of which is similar to stratified sampling (exhaustive, mutually exclusive), it differs in that clusters are selected at random (rather than all being selected as for strata).

Each cluster is usually unequal in size. Kish (1965) lists some problems due to unequally sized clusters when all elements in c selected clusters are sampled. Firstly, the sample size becomes a random variable since it depends on cluster size of selected clusters. Sometimes it is useful to use ratio estimates as it makes use of auxiliary information (information about the population) and can give more reliable estimates of the population values (Sukhatme and Sukhatme, 1970). The sample ratio of the means for two variables, say y and x, is given by $R_n = {}^{\overline{y}_n}\!\Big/_{\overline{x}_n}$, where $\overline{y}_n$ and $\overline{x}_n$ are the sample means for y and x respectively for a sample of size n. In cluster sampling,

$R_n\theta_x$, where θ_x is the population mean of x, 'is not an unbiased estimate of the population mean (θ_y)' (Kish, 1965; pg 183). Another problem is that variance estimates are not unbiased estimates of the true variance and lastly, the variance formulas are usually complicated.

For simple cluster sampling with equal cluster sizes (and all elements selected within sampled clusters), the mean is

$$\bar{y}_{srs} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$= \frac{1}{aB}\sum_{\alpha=1}^{a}\sum_{b=1}^{B} y_{\alpha b} \tag{1.9}$$

where n = total number of elements in the sample,

a = total number of clusters randomly sampled,

A = total number of clusters in the population,

B = total number of elements in a cluster (assuming equal cluster sizes)

and n = a×B (Cochran, 1963).

The variance is given by

$$\mathrm{Var}(\bar{y}_{cl}) = \left(\frac{1}{a}\right)\left(\frac{AB-1}{B^2(A-1)}\right)s^2[1+(B-1)r]$$

$$\approx \frac{s^2}{n}[1+(B-1)r] \tag{1.10}$$

where $s^2 = \sum_{\alpha,b}\frac{(y_{\alpha b}-\bar{y}_{srs})^2}{n-1}$ (assuming the sampling fraction n/N is negligible)

and r = intracluster correlation (Cochran, 1963). The intracluster correlation measures the extent which the elements in a cluster are dependent. If $r > 0$, $V(\bar{y}_{cl}) > V(\bar{y}_{srs})$, if $r < 0$, $V(\bar{y}_{cl}) < V(\bar{y}_{srs})$ and if $r = 0$, this is as if the elements were completely independent of each other, $V(\bar{y}_{cl}) = V(\bar{y}_{srs})$.

The advantages for this method are reduced costs, time and convenience for a given accuracy, even although sample sizes may be larger than for the simple random sample possible with the same resources (sometimes it is more appropriate to cluster sample, e.g., to survey all members in a household).

1.2.4 SYSTEMATIC SAMPLING

The idea behind systematic sampling is really quite simple. Elements are generally listed in some sort of way where every kth unit is selected. The starting point is usually a number selected randomly between 1 and N/n which is also the first element selected into the sample. For example, one way to systematic sample 10 houses from a total of 100, a number is randomly generated between 1 and 10 and this is the starting house number. Suppose number 3 is chosen randomly. House number 3 is then the first element of the sample and every 10th house is selected thereafter.

Hansen et al. (1953) point out that it is very hard to estimate the precision of a systematic sample when the population is small. In a large population, the estimates of precision are often approximated by assuming that systematic and random sampling are equivalent; where the sample frame has been ordered by a variable correlated with the variable of interest, this assumption is conservative.

1.2.5 MULTISTAGE SAMPLING

Sampling can consist of many stages of selection. The units selected at the first stage are known as the primary sampling units (PSU). If another sample is selected from each PSU then these are known as the secondary sampling units (SSU). This is also referred to as a two stage sampling design. If more units are selected from within each SSU, these are tertiary sampling units (TSU). This is then a three stage sampling design. The number of sampling stages can be greater, depending on the population structure and what information is to be gathered.

At each stage the sampling method can be any sampling design. For example, suppose we sample cities and within these units we sample households. Furthermore, we could sample individuals within the households. This would then be a three stage sampling design. The sampling design at the first stage could have been selected by simple random sampling and the households could be selected by probability proportional to size of household, if this were known. Within each household, the sampling design may again be a simple random sample.

The total variance of an estimator T(y), depends on the specific sampling designs used at each stage. The total variance is made up of components corresponding to each stage of the sampling design.

The main advantages of multistage sampling over single stage sampling are that it is flexible, and often more accurate.

1.3 CONFIDENCE INTERVALS

A confidence interval gives a measure of how accurate the estimate is. The probability or confidence level that the true parameter, θ, lies within the confidence interval is denoted by $1 - \alpha$. Ideally the confidence level should be close to 1. For example, $\alpha = 0.05$ gives a $100(1-\alpha)\%$, 95% confidence level. That is, for 95% of all possible samples, the confidence interval contains the true parameter.

In practice, the parameter, θ, is unknown and the confidence interval is constructed based on the sample estimate of θ, $\hat{\theta}$. The $100(1-\alpha)\%$ confidence interval is calculated by

$$\hat{\theta} \pm Z_{1-\alpha/2}\left(\hat{V}(\hat{\theta})\right)^{1/2} \qquad (1.11)$$

where $Z_{1-\alpha/2}$ is the value from the standard normal tables, i.e., $\alpha = 0.05$, $1 - \alpha/2 = 0.975$ and $Z = 1.96$. If the sample size, n, is small, then generally the t value with n-1 degrees of freedom, $t_{1-\alpha/2,n-1}$ from the Students t distribution is used instead of the standard normal distribution.

A point to note, the confidence interval assumes that the distribution of the parameter, $\hat{\theta}$, is approximately normal. This assumption is usually met through an application of the central limit theorem (CLT). The CLT states that as $n \to \infty$, $\hat{\theta}$ is approximately normal. (See Thompson (1992) for details.)

1.4 SOME BASIC CONCEPTS

The estimators presented so far are for the population mean and variance. These estimators have certain properties which make them 'good' estimators of the true population parameter. One of these properties is unbiasedness. Unbiasedness refers to an estimator which, on average, gives the true population value. That is, the expected value of the estimator equals the true population value. In mathematical notation, $E(T(y)) = \theta$, where θ is the population parameter to be estimated, and $E\big(T(y)\big) = \sum_{s \in S} T(y)p(s)$ where p(s) is the probability of selecting the sample s from the set of all possible samples S. All of the expansion estimators given in the different sampling designs are unbiased. If an estimator is biased, the bias is measured by $b(\theta) = E(T(y)) - \theta$.

Unbiasedness is not sufficient to determine a good estimator. Although unbiasedness is a desirable property, it does not measure the variation of the estimator. The mean square error (MSE), defined by $E[(T(y) - \theta)^2]$ is a measure of how close the distribution of the estimator, T(y), is concentrated around θ. Ideally we would like a small mean square error so that a sample drawn from the population will have an estimate which is close to the true population value. The MSE can also be written as

$$MSE(T(y)) = V(T(y)) + (b(\theta))^2. \tag{1.12}$$

From equation (1.12), it is easy to see that an unbiased estimator will have a MSE equal to the variance.

Quite often, researchers will use the term efficient. This concept may determine which estimator to use in the case where there may be more than one estimator which is unbiased. Suppose we had two unbiased estimators, T_1 and T_2. T_1 is said to be more efficient than T_2 if its variance is less than the variance of T_2, i.e., $V(T_1) < V(T_2)$. The extension of the idea of efficiency corresponds to an estimator being more efficient if it has a smaller MSE, when bias is present in at least one of the estimators being compared.

1.5 DESIGN EFFECTS

Survey data, when collected, are not always independent or identically distributed (IID). The selection of survey data is more complicated than simple random sampling due to the complexity of the population structure on the sampling scheme used. Standard analytical procedures have still been applied on the assumption that the sample survey data is IID. This causes problems with analysis on complex designs as the assumptions do not hold, i.e., observations in a cluster sample are no longer independent.

This led to the development of the concept of a design effect (Kish, 1963). The design effect illustrated the need for alternative variance estimation when the data were clustered or in strata.

Kish (1965) outlined a technique for dealing with some of the difficulties caused by complex designs. The design effect (DEFF) is the ratio of the actual variance of the design with the variance of a simple random sample, i.e.,

$$\text{DEFF} = V(y)/V_{srs}(y) \tag{1.13}$$

where $V(y)$ is the variance from the complex design and $V_{srs}(y)$ is the variance of a hypothetical simple random sample design of the same sample size.

This also measures the efficiency of a design against the simple random sampling design. For a stratified design, the DEFF is usually less than one. However, in cluster sampling, the DEFF is usually greater than one, hence it is not as efficient as a simple random sample.

Skinner (1989) describes a misspecification effect (MEFF, also termed a design effect and described as such in later chapters) which is similar to Kish's design effect, except that it is used primarily at analysis stage, where $V_{srs}(y)$ cannot be estimated because there is not an additional simple random sample available. The MEFF is given by:

$$\text{MEFF} = V(y)/E(V_o) \qquad (1.14)$$

where $E(V_o)$ is the expected variance of a simple random sample, i.e., if the available sample data was treated as a simple random sample and the variance of y was calculated on this basis.

When the MEFF is less than one, V_o is overestimating V(y) and when MEFF is greater than one, V_o is underestimating V(y).

Often, the actual variances are unavailable in practice and an estimate of the DEFF and MEFF are calculated using the corresponding estimated variances from the sample data. An estimate of DEFF is

$$\text{deff} = \hat{v}(y)/\hat{v}_{srs}(y), \qquad (1.15)$$

and an estimate of MEFF is

$$\text{meff} = \hat{v}(y)/E(\hat{v}_o). \qquad (1.16)$$

In stratified multi-stage designs with moderate sample sizes, $E(\hat{v}_o)$ and $\hat{v}_{srs}(y)$ will be approximately equal (Skinner, 1989), in which case, deff will be approximately equal to meff.

Derivatives from the DEFF include the design factor (DEFT) and an effective sample size. The DEFT is the square root of the DEFF, DEFT = DEFF$^{1/2}$. A measure of the 'amount of information' in the sample is given by the effective sample size, $n_e = n/\text{DEFF}$. It gives the sample size (n_e) of a simple random sample design which is needed to obtain the same level of precision as a particular complex design of size n.

1.5.1 MULTIVARIATE DESIGN EFFECTS

In more realistic situations there are a number of variables and a number of parameters. In these cases, it is easier to use matrix-vector notation. For example, if we had n observations on y, this could be written as y_i (i = 1,...,n) or as an n×1 vector, $\mathbf{y}_i$, if for each i, data were instead observed from p different variables, the data could be summarized in an n×p matrix. Note that in matrix-vector form, the vector or matrix is usually denoted in bold. The variance of the p×1 vector **y** which is made up of the estimators of the p parameters of interest will be a matrix with diagonal elements σ_{ii} (i = 1,...,p), and the off diagonal elements σ_{ij} (i,j = 1,...,p; i ≠ j). This is denoted by **V**(**y**). Hence the multivariate design effect is given by

$$\mathbf{D} = \mathrm{E}(\mathbf{V}_o)^{-1}\,\mathbf{V}(\mathbf{y}). \tag{1.17}$$

The corresponding estimate of **D** is given by

$$\hat{\mathbf{D}} = \mathrm{E}\left(\hat{\mathbf{V}}_o\right)^{-1}\hat{\mathbf{V}}(\mathbf{y}). \tag{1.18}$$

Matrices are very important, especially with multivariate data. However, the matrices can get very large and unmanageable. Here we introduce the terms eigenvalues and eigenvectors. The eigenvalues are scalars which indicate the spread of the data in p dimensions and the eigenvectors are vectors indicating direction associated with the corresponding eigenvalue. They play an important role in reducing the matrices into a more comprehensible and interpretable manner. Details of finding the eigenvalues and eigenvectors can be found in chapter four.

The eigenvalues of the design matrix are called the generalized design effects, denoted by δ.

1.6 THE PROBLEM IN THE ANALYSIS OF SURVEY DATA

A complex survey is one which involves a sampling scheme other than simple random sampling. Standard statistical analyses are based on the assumption of IID observations, viz., simple random samples. In most surveys, the sampling scheme is usually of a complex design involving some natural clustering or stratification in the population structure. Applying the standard statistical methods to a complex survey sample violates the assumptions underlying these classical methods and can result in misleading conclusions (Kish and Frankel, 1974; Holt, Smith and Winter, 1980; Skinner et al., 1989).

Sample surveys were developed primarily for descriptive statistics (Kish and Frankel, 1974). Analytic statistics developed as the need grew for understanding and explaining rather than merely describing the population. Descriptive uses of survey data are those such as means, standard deviations, ratios, proportions and totals. They are descriptive as they only concern estimates for a particular population at a particular time.

For prediction or inference purposes, analytic statistics are required. To be able to generalize or predict future observations, some underlying explanation of the descriptive findings is sought. The objective of analytic studies is to find conclusions that can be generalized.

Even though sample surveys were used before the development of sampling schemes, the population of interest usually had some population structure. Samples collected from the population would not have been a simple random sample. Concern grew over how to select samples that were representative of the population and how to generalize the results from a sample back to the population. Due to Neyman (1934; cited in Skinner et al., 1989), randomized methods of selection were introduced into the sampling process. Around the 1950s, sampling techniques surfaced and later the DEFF (Kish, 1965). As the statistical methods used to analyse survey data were based on simple random samples, more attention was now given to complex designs (Binder et al., 1987). Complex statistics were needed for complex surveys as the standard statistics on clustered data gave smaller standard errors and narrower confidence intervals (Skinner et al., 1989). Hence in analytic analyses, the results would appear more significant than they really were.

1.6.1 DESIGN AND MODEL BASED APPROACHES

The design based approach is one which has been traditionally used in sample surveys (Cassel, Särndal and Wretman, 1977). All formulae discussed previously are design based. It is called a design based approach because the sampling design is considered important. Also known as the fixed population approach (Cassel et al., 1977), inferences from samples refer only to the finite population chosen.

Contrary to a design based approach, is a model based approach. This approach (also known as the superpopulation approach) assumes an infinite, hypothetical population, often referred to as the superpopulation. The strict superpopulation approach does not require the knowledge of the sampling design, since statistical inference from a sample is conditional on the sample.

1.6.2 THE SUPERPOPULATION

The concept of a superpopulation has been around for over half a century (Cassel et al., 1977). Traditionally researchers have relied on design based approaches, that is, given a finite population, a sample is selected and an estimate of the parameter is found. Because the values in the finite population are fixed, the only randomness comes from the sampling design. A design based approach does not depend on any assumptions about the finite population.

The superpopulation, or model based approach, relies on an assumed model for the population. The **y** values for individual sample elements, $\mathbf{y}_i$, i = 1,...,N are no longer fixed but are random variables, $\mathbf{Y}_i$. The joint distribution of these $\mathbf{Y}_i$ values is denoted by ξ. It is this ξ distribution which is modelled.

The sampling design is considered to be unimportant in a model based approach. This is because the sample, s, is assumed to be given and inferences are conditional on the given sample, s.

A model based approach has certain advantages over a purely design based approach. Although Särndal, Swensson and Wretman (1992) point out that it is an alternative approach to the traditional design based approach, it has proved to be practical and advantageous in sample surveys with some provisos. The superpopulation offers flexibility in finding estimators by placing different assumptions on the population. However, the superpopulation is good only when the assumed model holds. By combining the features of the design approach and the model approach, this would suggest an approach which would give desirable properties with respect to both the design and the superpopulation. This has been termed model assisted sampling (Särndal et al., 1992).

1.6.3 DESIGN AND MODEL EXPECTATIONS

For a design based approach, the expectation of an estimator is referred to as the design expectation, or p expectation, denoted by $E_p(\cdot)$. Define for an estimator T = T(s), for some parameter θ, the p expectation is $E_p(T(s)) = \sum_{s \in S} T(s)p(s)$. This expectation is with respect to the randomisation procedure used in the sample design. Given a particular sample s, define for an estimator T = T(s) for some parameter θ, the ξ or model expectation as $E_\xi(T|s) = \int T(s)dF_\xi(\{y_i \in s\})$ where F_ξ is the distribution function of the $\{y_i\}$ and $F_\xi(\{y_i \in s\})$ denotes the multivariate distribution function for the set of y_i belonging to a particular sample, s. For each y_i, the limits of integration are $-\infty$ to $+\infty$ in general.

For a joint design and model based approach, there is an expectation for T with respect to the combination of the design and the model. This can be denoted as $E_{p\xi}(T)$ or $E_{\xi p}(T)$ depending on the order. When the sampling design, p(s), does not depend on the variables of interest, it is said to be noninformative. Cassel et al. (1977) notes that when a sample design is noninformative then the $p\xi$ expectation is equivalent to the ξp expectation.

The corresponding variances will be denoted by $V_p(\cdot)$, $V_\xi(\cdot), V_{p\xi}(\cdot)$ for the design based, model based and joint $p\xi$ based approaches respectively. We can define for an estimator T, the p variance is $V_p(T|p(s)) = E_p[(T - \theta)^2]$ for all $s \in S$, the ξ variance of T is $V_\xi(T|s) = E_\xi[(T - E_\xi(T)]^2$ for a given sample s and the $p\xi$ variance of T is $V_{p\xi}(T) = E_{p\xi}[T - E_{p\xi}(T)]^2$ (Cassel et al., 1977)

The p or design bias is defined for T by $b_p(T) = E_p(T) - \theta$. The ξ or model bias is defined for T by $b_\xi(T) = E_\xi(T) - \theta$. The $p\xi$ bias is defined for T by $b_{p\xi}(T) = E_{p\xi}(T) - \theta$. The estimator T is said to be p, ξ or $p\xi$ unbiased if the corresponding bias equals zero (Haslett, 1985).

CHAPTER TWO

SURVEY DATA

There are two different survey questionnaires used in this thesis. The first is the 1986 Wellington Community Questionnaire on victimisation. The second questionnaire is the 1996 New Zealand National Survey of Crime Victims. Both of these questionnaires are surveys on crime and the effects of crime on the community. The main difference between these two questionnaires is that the 1986 survey was carried out in the Porirua and Lower Hutt regions of Wellington, New Zealand whereas the 1996 survey was carried out over the whole of New Zealand.

The houses were selected by probability sampling, i.e., all houses in the sampling frame have a non zero probability of being selected. Generally the responses collected from these surveys were from house to house interviews. The surveyors went to the selected houses (or dwellings) and conducted face to face interviews with the respondents. Ineligible houses (houses no longer at the given address etc.) were culled by the surveyors. Non respondents (no one home at the selected house) were noted as a non respondent and were not able to be used in any data analyses. For individual respondents who refused to answer some questions or the questions did not apply to them, these question responses could also not be included in the data analyses.

2.1 1986 WELLINGTON COMMUNITY QUESTIONNAIRE

This questionnaire was carried out in the Porirua and Lower Hutt regions of Wellington only by the Institute of Criminology and Institute of Statistics and Operations Research, Victoria University of Wellington. Individuals who participated in this survey were 16 years of age or older and who were residents in the Porirua or Lower Hutt regions at the time of the survey.

There were several objectives of the study but mainly to investigate people's concerns and to provide a data base about crime in the community. There are basically two parts to the questionnaire, one on personal offences (i.e., where a person has been a victim to a crime) and the other on household offences (i.e., burglary, theft, vandalism, etc., to household property).

The sampling design used was a three stage design. Both regions were divided into meshblocks which are contiguous areas each containing around 200 people using the Department of Lands and Survey maps. The primary sampling units consisted of an aggregation of meshblocks of which there were no more than 200 houses except in cases where a single meshblock exceeded 200 houses. Forty meshblock aggregates were selected systematically out of a total of 104 in the Porirua area and forty out of 177 were selected systematically in the Lower Hutt area. Every fourth house was selected from the selected PSU's and these were the secondary sampling units. The tertiary sampling units were eligible individuals at each household. Only one individual was selected at each household and this was done by selecting the individual who has a birthday first following the date of the interview.

2.2 1996 NEW ZEALAND NATIONAL SURVEY OF CRIME VICTIMS

The second questionnaire is the 1996 New Zealand National Survey of Crime Victims (Young, Morris, Cameron and Haslett, 1997) which was carried out by AGB McNair for the NZ Police, Ministry of Justice, Crime Prevention Unit, Department of Social Welfare, Te Puni Kokiri, Ministry of Youth Affairs, Ministry of Women's Affairs and Victoria University. The questionnaire is similar to the 1986 Wellington Community Questionnaire on victimisation, but this survey used New Zealand as the population of study, not just the Porirua and Lower Hutt regions. The main aim of this questionnaire was to get a measure of how much crime there is and what effects it may have on the victims.

The sampling design used was also a multistage design. The meshblocks from the 1991 Census developed by Statistics NZ were insufficient as they were considered too small (each meshblock contained about 30 to 70 houses). McNair Area Units (MAU) were developed which combined about 7 meshblocks together. These MAU's formed the primary sampling units. Selection of clusters i.e., MAU's was done by probability proportional to size. The target sample size is 4500 individuals, with 5 individuals per cluster which gives 900 clusters (i.e., MAU's) to be selected. The MAU's are systematically listed by geographical location and for each MAU the number of eligible individuals (15 years of age and over) was known, but not the actual location of each person. A total of 2,590,284 individuals were eligible but only 900 sample clusters are needed. Dividing 2,590,284 by 900 gives every 2878 person needs to be selected. By taking a starting point, a random number between 1 and

2878, and from there the MAU containing every 2878 person is selected into the sample. This is sampling with probability proportional to size as the proportion of eligible individuals in the sample from the selected MAU's is directly proportional to the total population. The 900 MAU's selected are the starting point for the cluster sample. From each of the identified MAU's, 5 households are selected, and the one individual selected in each of these households is interviewed.

Note that in this questionnaire, there are two types of weighting variables used. One is an individual weight and the other is a household weight. These weights are equivalent to the inverse of the selection probabilities for individuals and households, with some poststratification by gender, age and ethnicity. I.e., the variables age (grouped), ethnicity and gender have been adjusted so that the total weights from the sample equal the respective population numbers for these variables.

2.2.1 CODING

Data collected from surveys are generally numeric. This data needs to be entered into a computer to be able to perform some (statistical) analyses. Coding is often used by researchers to help identify the question number with the correct response in the computer. Table 2.1 lists the codes of the variables (in bold) and the categories from the 1996 New Zealand National Survey of Crime Victims which have been used in examples throughout the thesis.

The questionnaires used are in A.1 and A.2 of the appendix.

Table 2.1. Codes used.

Code	Explanation
hhoffi	**Total number of household offences**
hhsize	**Household size (number of people living at the house)**
d4	**Type of household**
d4_1	One person living alone
d4_2	Solo parent with child/children
d4_3	Couple without children
d4_4	Couple with children
d4_5	Extended family/whanau
d4_6	Flatmates
d4_7	Family - other combination
d8d9a	Rented house from private owner
d8d9b	Rented house from Local authority/council
d8d9c	Rented house from Housing New Zealand
d8d9d	Owned house (including mortgage)
d8d9e	Other
d11	**Occupation**
d11_1	Level 1
d11_2	Level 2
d11_3	Level 3
d11_4	Level 4
d11_5	Level 5
d11_6	Level 6
d11_7	Level 7
d11_8	No main income earner
w1	**What kind of neighbourhood it is**
w1_1	Help each other
w1_2	Go own way
w1_3	Mixture
w6a1	**Rubbish and litter lying about on the streets/empty sections**
w6a1_1	Very big problem
w6a1_2	Fairly big problem
w6a1_3	Not a very big problem
w6a1_4	Not a problem at all
w6a2	**Broken windows in shops and public buildings etc.**
w6a2_1	Very big problem
w6a2_2	Fairly big problem
w6a2_3	Not a very big problem
w6a2_4	Not a problem at all
w6a3	**Graffiti on walls, schools, shops, churches etc.**
w6a3_1	Very big problem
w6a3_2	Fairly big problem
w6a3_3	Not a very big problem
w6a3_4	Not a problem at all

Table 2.1. Codes used (continued).

Code	Explanation
w6a4	**Uncontrolled dogs roaming the neighbourhood**
w6a4_1	Very big problem
w6a4_2	Fairly big problem
w6a4_3	Not a very big problem
w6a4_4	Not a problem at all
w6a5	**Teenagers hanging around on the streets**
w6a5_1	Very big problem
w6a5_2	Fairly big problem
w6a5_3	Not a very big problem
w6a5_4	Not a problem at all
w6a6	**Gang members living or hanging around the neighbourhood**
w6a6_1	Very big problem
w6a6_2	Fairly big problem
w6a6_3	Not a very big problem
w6a6_4	Not a problem at all
w6a7	**Drunks, glue sniffers or people high on drugs on the streets**
w6a7_1	Very big problem
w6a7_2	Fairly big problem
w6a7_3	Not a very big problem
w6a7_4	Not a problem at all
w7	**How often you go out at night**
w7_1	Never
w7_2	At least once a week
w7_3	At least once a fortnight
w7_4	At least once a month
w7_5	Less often than once a month
hhwgtpos	**Poststratified household weight**
inwgtpos	**Poststratified individual weight**
burgp	**Prevalence of burglary (0 = no, 1 = yes)**
d5	**Gender (1 = male, 2 = female)**
age_1	age between 15 and 24
age_2	age between 25 and 39
age_3	age between 40 and 59
age_4	age greater than 60
age_5	not specified
age_6	refused to give age
ethnic_1	European
ethnic_2	NZ European
ethnic_3	NZ Maori
ethnic_4	Pacific Islander
ethnic_5	Other

Table 2.1. Codes used (continued).

Code	Explanation
d10	**Personal work situation**
d10_1	Working in paid employment
d10_2	Home duties
d10_3	Retired/superannuitant
d10_4	Benefit/unemployed
d10_5	Sick/disabled and unable to work
d10_6	Unpaid work outside the home
d10_7	Student
d10_8	Other
d13	**Current marital status**
d13_1	Legally married
d13_2	Defacto relationship
d13_3	Single/never married
d13_4	Widowed
d13_5	Divorced/separated

CHAPTER THREE

REGRESSION

In sampling situations, it is common to find one or more variables that are linearly related to the variable of interest. This linear relationship can be described by a technique called regression analysis. The aim of regression is to predict the variable of interest, commonly known as the dependent variable, from these independent or regressor variables. Another important factor of regression is control. That is the ability to control some variables while manipulating others to see how the relationship alters. For example, suppose we wanted to measure what variables predict the total number of individual offences (from the 1996 questionnaire). Data is collected on the variables which may have an effect. These variables may be occupation, age and so on. Some of these variables can be controlled, e.g. controlling the occupation (only use the data of those employed).

Usually in a regression model (see section 3.1), there is one dependent variable which is measured in the presence of the independent or regressor variables. In the example above, the dependent variable is the total number of individual offences (on a victim) and the other variables are the regressor variables. Using the regressor variables, the dependent variable can be modelled and it is possible to use the regression model for prediction. Other uses of regression analysis are variable screening, model specification and parameter estimation. Variable screening is really to identify those variables that are not influencing the dependent variable and to exclude them from the model. Model specification is very important to regression analysis. This tries to select a better model for the data. Finally, in a regression model, the coefficients are usually unknown constants and have to be estimated. These uses are not necessarily mutually exclusive.

3.1 THE REGRESSION MODEL

A regression model is at the core of regression. All analyses in regression are related to a model. For one dependent and one independent variable which are known to be linearly related, this can be modelled in the form of a straight line equation, $y = bx$, and hence it is known as an example of a linear model. For example, individual offences (on a victim) might be linearly related with age and hence, this relationship may be modelled by a straight line equation. However not all regression models are linear. Non-linear models will be discussed further on in the chapter.

3.1.1 THE SIMPLE LINEAR MODEL

Without doubt, the simplest regression model (that also has an intercept) is one which has only one independent variable and the dependent variable. By taking i = 1,...,n values of the independent variable and observing the dependent variable, we can record the pairs of observations as (x_i, y_i). Assuming that the model is linear, it can be represented as

$$y_i = B_0 + B_1 x_i + \varepsilon_i \qquad \text{for } i = 1,\ldots,n \quad (3.1)$$

which is equivalent to the straight line equation, where B_0 and B_1 are the intercept and slope respectively. The ε_i is the error term.

To illustrate using the 1996 survey data, the mean of the dependent variable individual offences and the independent variable age can be plotted. The plot is given in figure 3.1. Note that similar patterns occur for other individual variables also.

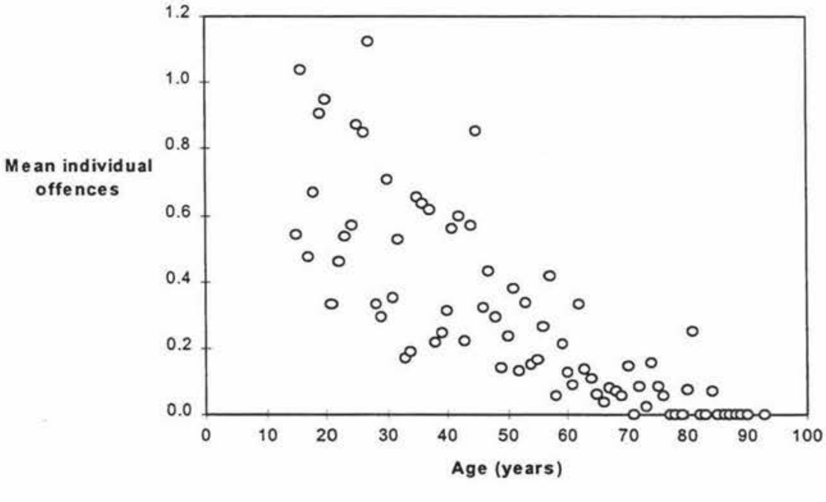

Figure 3.1. Scatter plot of the mean number of individual offences and age (in years).

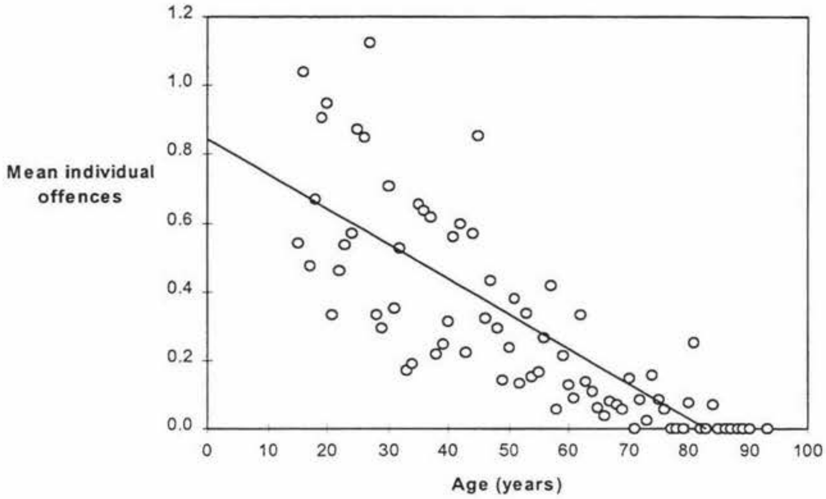

Figure 3.2. Regression line fitted onto the scatter plot of the mean number of individual offences and age (in years).

Recall that regression analyses seeks to find a relationship between variables. It is possible to see that the two variables in the example above may be related (see figure 3.1) and that the relationship can be modelled by a straight line. However, for a straight line, we need to estimate the parameters B_0 and B_1 (see section 3.1.2). Once B_0 and B_1 are estimated, a straight line can then be drawn on the plot to illustrate that a relationship exists. This is shown on figure 3.2. An example of prediction is, given that a person is 70 years old, then that person is predicted to have a mean of about 0.15 individual offences.

Clearly a linear regression is not completely adequate here as, if someone is over 83 years of age, the regression line gives a negative value for the mean number of individual offences. Figures 3.1 and 3.2 are for illustrative purposes only as there is also possible evidence for a non-linear regression line to be fitted.

3.1.2 PARAMETER ESTIMATION

Ordinary Least Squares (OLS)

The idea behind ordinary least squares (OLS) is to estimate B_0 and B_1 so that the residual sum of squares is minimised, i.e., $\sum_{i=1}^{n} \varepsilon_i^2$ is minimised. This is when

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \qquad (3.2)$$

is minimised, where

$$y_i = b_0 + b_1 x_i + \varepsilon_i \qquad (3.3)$$

and b_0, b_1 are the estimated B_0 and B_1 respectively.

Myers (1986) points out that it is useful to rewrite the linear model by centring the simple linear model. By centring the x values will be distributed around the value zero. This makes it computationally easier yet the regression equation is equivalent to that of a non centred model. Let $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, then

$$y_i = b_0' + b_1\left(x_i - \bar{x}\right) + \varepsilon_i\,. \qquad (3.4)$$

Note that $b_0' = b_0 + b_1 \bar{x}$ which equates (3.3) with (3.4).

Minimising the sum of squares of the residual is where

$$\frac{\partial}{\partial b_j} \sum_{i=1}^{n} \varepsilon_i^2 = 0 \quad \text{is satisfied,} \qquad \text{for } j = 0,1. \quad (3.5)$$

Hence

$$\frac{\partial}{\partial b_0{}'}\sum_{i=1}^{n}\left(y_i - b_0{}' - b_1\left(x_i - \bar{x}\right)\right)^2 = 0,$$

$$\frac{\partial}{\partial b_0{}'} = -2\sum y_i + 2nb_0{}' + 2\sum b_1\left(x_i - \bar{x}\right) = 0,$$

$$b_0{}' = \frac{\sum y_i}{n} = \bar{y} \qquad (3.6)$$

and

$$b_0 = b_0{}' - b_1\bar{x}$$

$$\Leftrightarrow b_0 = \bar{y} - b_1\bar{x}. \qquad (3.7)$$

Similarly

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}\left(y_i - b_0{}' - b_1\left(x_i - \bar{x}\right)\right)^2 = 0,$$

$$\frac{\partial}{\partial b_1} = -\sum y_i\left(x_i - \bar{x}\right) + \sum b_0{}'\left(x_i - \bar{x}\right) + b_1\sum\left(x_i - \bar{x}\right)^2 = 0,$$

$$b_1 = \frac{\sum y_i\left(x_i - \bar{x}\right)}{\sum\left(x_i - \bar{x}\right)^2} = \frac{S_{xy}}{S_{xx}}. \qquad (3.8)$$

Thus the estimators for intercept and slope respectively, found by the OLS method are equations (3.7) and (3.8). The corresponding variances are given by

$$\mathrm{Var}\left(b_0\right) = \mathrm{Var}\left(\bar{y} - b_1\bar{x}\right)$$

$$= \mathrm{Var}\left(\sum_{i=1}^{n}\frac{y_i}{n}\right) + \mathrm{Var}\left(b_1\bar{x}\right)$$

$$= \frac{\sigma^2}{n^2} + \overline{x}^2\left(\frac{\sigma^2}{S_{xx}}\right)$$

$$= \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right) \tag{3.9}$$

for the intercept and for the slope,

$$\mathrm{Var}(b_1) = \mathrm{Var}\left(\frac{\sum y_i(x_i - \overline{x})}{\sum(x_i - \overline{x})^2}\right)$$

$$= \frac{1}{\sum(x_i - \overline{x})^2}\mathrm{Var}\left(\sum y_i\right)$$

$$= \frac{\sigma^2}{S_{xx}}. \tag{3.10}$$

Maximum Likelihood Estimation (MLE)

Another method of estimating the regression parameters is the maximum likelihood estimation method. Assuming that the errors, ε_i, are IID and follow the normal distribution (for simplicity let $\sigma = 1$), $\varepsilon \sim N(0,1)$. The likelihood function is given by

$$L(\varepsilon_i) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}\varepsilon_i^2}. \tag{3.11}$$

Taking logs from both sides, we get

$$\ln L(\varepsilon_i) = \frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\varepsilon_i^2 \tag{3.12}$$

and maximising the likelihood function means taking partial derivatives and setting to zero,

$$\text{i.e., } \frac{\partial}{\partial B_j}\left[-\frac{1}{2}\sum_{i=1}^{n}\varepsilon_i^2\right]=0,$$

$$\frac{\partial}{\partial B_j}\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i - B_0 - B_1x_i)^2\right]=0 \qquad \text{for } j = 0,1 \quad (3.13)$$

which gives exactly the same results as for the OLS method. However, this is only true under the assumption that ε_i follow a normal distribution. If this assumption is not met, then the OLS is not always the same as the MLE method.

3.2 MULTIPLE LINEAR REGRESSION

In most real life applications of regression there are usually more than one regressor variable. From the previous example of the mean individual offences as the dependent variable and the age as the regressor variable, one may feel that the mean individual offences is not only influenced by age but also other variables such as ethnicity, length of time in the neighbourhood and so on.

The model for a multiple linear regression, is just an extension of the simple linear regression

$$y_i = B_0 + B_1x_{i1} + B_2x_{i2} + \ldots + B_px_{ip} + \varepsilon_i \qquad \text{for } i = 1,\ldots,n \quad (3.14)$$

where y_i is the dependent variable (the one which we are trying to predict and draw conclusions from), the x_{ij}'s (j = 1,...,p) are the regressor variables, for example x_{i1} = age, x_{i2} = length of time in the neighbourhood and so on. B_0 is the intercept and the B_j's (j = 1,...,p) are the regression coefficients and ε_i is the random error associated with the model. This model is still in the linear form and the parameters to be estimated are the intercept and the unknown coefficients, i.e., B_j for $j = 0,...,p$.

3.2.1 PARAMETER ESTIMATION

Since we are dealing with multiple parameters, it is easiest to represent these in matrix notation. Several observations are made on the p number of x variables that lead to several observations on the y. In matrix notation,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_{p-1} \\ B_p \end{bmatrix} \text{ and } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{bmatrix}, \quad (3.15)$$

hence (3.14) can be written as

$$\mathbf{y} = \mathbf{XB} + \varepsilon \quad (3.16)$$

The residual sum of squares is given by $(\mathbf{y} - \mathbf{XB})'(\mathbf{y} - \mathbf{XB})$ and to minimise, solve

$$\frac{\partial}{\partial \mathbf{B}}(\mathbf{y} - \mathbf{XB})'(\mathbf{y} - \mathbf{XB}) = \mathbf{0},$$

$$\text{i.e., } \frac{\partial}{\partial \mathbf{B}}(\mathbf{y'y} - \mathbf{X'B'y} - \mathbf{y'XB} + \mathbf{B'X'XB}) = \mathbf{0},$$

$$\text{i.e., } \frac{\partial}{\partial \mathbf{B}}(\mathbf{y'y} - 2\mathbf{X'B'y} - \mathbf{B'X'XB}) = \mathbf{0},$$

$$\text{i.e., } -2\mathbf{X'y} + 2\mathbf{X'XB} = \mathbf{0},$$

$$\text{i.e., } \mathbf{X'XB} = \mathbf{X'y},$$

$$\text{hence } \mathbf{B} = (\mathbf{X'X})^{-1}\mathbf{X'y}. \quad (3.17)$$

The variance for **B** is given by

$$\begin{aligned}\operatorname{Var}(\mathbf{B}) &= \operatorname{Var}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right)\\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\operatorname{Var}(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\\ &= \left(\sigma^2\mathbf{I}\right)(\mathbf{X}'\mathbf{X})^{-1} \qquad (3.18)\end{aligned}$$

3.3 WEIGHTED LEAST SQUARES

In ordinary least squares, the assumption is that the errors are IID and the variance of the errors are homogeneous. This assumption is quite often violated when working with real data, either due to the nature of the population or to the sampling. Alternative methods are usually needed to estimate the B's as OLS will not give accurate estimates. However, OLS will typically give consistent estimators.

Suppose the data collected has an increase in variance. To compensate for the different variances, a weighting proportional to each observation's variance is used. These weights can be represented in matrix notation,

$$\mathbf{W} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^2 \end{bmatrix}. \qquad (3.19)$$

The diagonal elements of the matrix **W** are the variances of each observation (1,2,...,n) and the off diagonal elements are assumed zero. The weighted least squares (WLS) estimate of **B** is given by

$$\hat{\mathbf{B}}_{\mathrm{w}} = \left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y}. \tag{3.20}$$

The variance of $\hat{\mathbf{B}}_{\mathrm{w}}$ is

$$\mathrm{Var}\left(\hat{\mathbf{B}}_{\mathrm{w}}\right) = \left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)\,\sigma^2. \tag{3.21}$$

An extension to the case where the weighting matrix is not diagonal is possible.

3.4 SUBSET SELECTION

Often where there are a lot of regressor variables, there exists a major problem to decide which regressor variables to go into the regression model. Usually only a subset of all of the regressor variables are selected into the model. There are various reasons why only some of the variables are used.

One of the main reasons is to cut down the costs as a subset of variables would obviously require less time and money to collect than for the full set of variables. Care has to be taken however, not to omit relevant variables. Some of the variables do not necessarily affect the dependent variable and thus sometimes a reduced model is all that is required to give a model which is statistically as good as the full model. For example, if two regressor variables are highly correlated, then you may only need one of those variables in the model (Seber, 1977). The percentage of variance explained also increases with the number of regressor variables which is why a subset may be more preferable. Which regressor variables to include into the model depends on which are easier to measure and which cost less.

There is no 'correct' model, only good ones. To decide which models are better than others, there are a few measures that are referred to.

The first is the coefficient of determination, R^2,

$$R^2 = \frac{SS_{\text{regression}}}{SS_{\text{total}}} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}. \tag{3.22}$$

It is also the square of the correlation coefficient. This R^2 measures the total variability explained by the regressors in the model. If the model is a good fit then it should explain most of the variability, hence a large R^2 value. However, R^2 should be treated with caution as more variables will increase the value of R^2. This can lead to an overfit of the model (including too many regressor variables) which is not necessarily a better model than a simpler one.

The second measure is the estimate of the mean square of error (MSE), s^2. This is the variation in the error term,

$$\begin{aligned} s^2 &= \text{MSE} \\ s^2 &= \frac{\sum_{i=1}^{n}\left[y_i - \left(\hat{b}_0 + \hat{b}_1 x_{i1} + \ldots + \hat{b}_p x_{ip}\right)\right]^2}{n-k} \end{aligned} \tag{3.23}$$

Note that $k = p+1$ = number of parameters. A better model will have a smaller MSE so a simple 'rule of thumb' would be to choose the model with the smallest value of s^2, except that MSE decreases with the number of variables added.

One could also look at the adjusted R^2 as the third measure. The adjusted R^2 takes into account the additional terms in the model. This is calculated as

$$R^2(\text{adj}) = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-p}\right). \tag{3.24}$$

3.4.1 MALLOWS C_p STATISTIC

This statistic tries to select a subset of regressors by balancing the effect of underfitting (i.e., excluding terms which are important) and overfitting (i.e., including terms which are unimportant). A brief outline will be given here only, for further details consult Myers (1986).

The effect of underfitting will result in biased estimates of **B**, whereas overfitting will lead to large variances of the coefficients. Mallows C_p Statistic is based upon the MSE of the fitted values of the y_i's and is given by

$$C_p = \sum_{i=1}^{n} \frac{\mathrm{MSE}(\hat{y}_i)}{\sigma^2} = \sum_{i=1}^{n} \frac{\mathrm{Var}(\hat{y}_i) + (\mathrm{Bias}(\hat{y}_i))^2}{\sigma^2}. \qquad (3.25)$$

It is shown in Myers (1986) that $\sum_{i=1}^{n} \frac{\mathrm{Var}(\hat{y}_i)}{\sigma^2}$ = k (number of parameters) and $\sum_{i=1}^{n} (\mathrm{Bias}(\hat{y}_i))^2$ = $(s^2 - \sigma^2)(n - k)$, where s is the estimate of the residual variance. Hence

$$C_p = k + \frac{(s^2 - \sigma^2)(n-k)}{\sigma^2} \qquad (3.26)$$

if σ^2 is known.

The desirable model would be one which has a C_p Statistic equal to k.

3.5 NON-LINEAR REGRESSION

Not all regression analyses are in the linear form. To illustrate a non-linear regression, consider this example (from Myers, 1986),

$$y_i = B_0 e^{B_1 x_i} + \varepsilon_i \tag{3.27}$$

the parameters B_0 and B_1 are estimated by the method of least squares, i.e., minimising the sum of squares of the residuals (SS_{res}), where SS_{res} is given by

$$\sum_{i=1}^{n} \left(y_i - B_0 e^{B_1 x_i}\right)^2 . \tag{3.28}$$

As with solving the OLS, (3.28) is differentiated with respect to B_0 and B_1 and the resultant equations set to zero. These equations are found to be

$$\sum_{i=1}^{n} \left(y_i - B_0 e^{B_1 x_i}\right)\left(-e^{B_1 x_i}\right) = 0 , \tag{3.29}$$

$$\sum_{i=1}^{n} \left(y_i - B_0 e^{B_1 x_i}\right)\left(-B_0 e^{B_1 x_i} x_i\right) = 0 \tag{3.30}$$

for B_0 and B_1 respectively.

However these equations, (3.29) and (3.30), are still non-linear in the parameters. To solve these equations requires some iterative procedure, the most common being the Gauss-Newton procedure (Myers, 1986). This procedure involves a Taylor series expansion.

For the general non-linear model of p regressor variables (and the intercept term), the model is written as

$$y_i = f(\mathbf{x}_i, \mathbf{B}) + \varepsilon_i \qquad \text{for } i = 1,\ldots,n \tag{3.31}$$

where $f(\mathbf{x}_i, \mathbf{B})$ is the non-linear function.

We need to minimise

$$\mathrm{SS}_{\mathrm{res}} = \sum_{i=1}^{n}\left(y_i - f\left(\mathbf{x}_i, B\right)\right)^2 \tag{3.32}$$

by expanding f($\mathbf{x}_i$, **B**) using the Taylor series:

$$f(\mathbf{x}_i, \mathbf{B}) = f(\mathbf{x}_i, \mathbf{B}_0) + \left(B_1 - B_{0,1}\right)\left[\frac{\partial f(\mathbf{x}_i, \mathbf{B})}{\partial B_1}\right]_{\mathbf{B}=\mathbf{B}_0} + \left(B_2 - B_{0,2}\right)\left[\frac{\partial f(\mathbf{x}_i, \mathbf{B})}{\partial B_2}\right]_{\mathbf{B}=\mathbf{B}_0} + \ldots$$

$$+ \left(B_p - B_{0,p}\right)\left[\frac{\partial f(\mathbf{x}_i, \mathbf{B})}{\partial B_p}\right]_{\mathbf{B}=\mathbf{B}_0} + \text{ remainder terms of higher order} \tag{3.33}$$

and letting $w_{ij} = \left[\frac{\partial f(\mathbf{x}_i, \mathbf{B})}{\partial \mathbf{B}_j}\right]_{\mathbf{B}=\mathbf{B}_0}$ and $\alpha_j = B_j - B_{0,j}$ we can write (3.33) as

$$y_i - f(\mathbf{x}_i, \mathbf{B}_0) = \alpha_1 w_{i1} + \alpha_2 w_{i2} + \ldots + \alpha_p w_{ip} + \varepsilon_i \qquad \text{for } i = 1,\ldots,n. \tag{3.34}$$

Equation (3.34) is then in the form of a linear regression where the α_j's are the regression coefficients and the w_{ij}'s are the regressor variables.

Estimation of the α_j's provides estimates for B_j's which replaces the first estimate of B_0. Note that $B_{1,j} = B_{0,j} + \alpha_{1,j}$.

3.5.1 A SPECIAL CLASS OF NON-LINEAR MODELS

There are a special family of non-linear models, called the Generalized Linear Models. So called because they generalize the linear models to allow more general distributions to be modelled. In chapter four, section 4.4 provides a detailed discussion on these non-linear models, including logistic regression and log-linear models.

3.6 REGRESSION ANALYSIS ON COMPLEX SURVEYS

So far the regression analyses discussed have been for design based, simple random sampling procedures. The methods used such as OLS are assuming that the errors are IID and these methods do not take into account any clustering or stratification in the population structure.

Regression analyses on complex survey designs are becoming increasingly common and hence the effects of complex survey data on OLS have been discussed widely (for example, Fuller, 1975; Holt, Smith and Winter, 1980; Nathan and Holt, 1980; Pfeffermann and Nathan, 1981).

Smith (1981) has pointed out that "when the population has a complex structure, it is not obvious what regression relationships should be examined" (pg 268). That is, the population will usually have clusters or strata and the question may be whether to fit one line to the overall data or several lines to the several subpopulations. Usually an aggregate line is fitted to the data and the subpopulations are appropriately weighted. He also points out that there are two types of inference associated with the estimation of the regression coefficient. Under a finite population (such as that of the previous section), the $\hat{\mathbf{B}}$ is a descriptive inference, merely describing the data collected. Assuming that the finite population is now taken from an infinite population, say a superpopulation, the regression coefficient $\hat{\beta}$ now allows an analytic inference. This is because the $\hat{\beta}$ found can be generalized to another sampled population from the same superpopulation. In the notation used here, for a finite population, the regression parameters are denoted as **B** and in an infinite population with an assumed regression model, the parameters are denoted as β. Holt et al. (1980) points out that if the model in the superpopulation is true, then $\mathbf{B} = \beta + O_p(N^{-1/2})$ (where O_p denotes further terms of the stochastic variable, β, to order $N^{-1/2}$). This is also true if the finite population is large.

One of the assumptions of the regression model (3.1) assumes independence between observations. Structured designs, for example cluster sampling, tend to bring about intracluster correlation among the observations. By assuming that the finite population is a random sample from a superpopulation, the structure of the finite population is considered to be unimportant. However, that is not to say that the survey design does not matter (Smith, 1981), as the survey design is found to affect the efficiency of the estimators (Holt et al., 1981). Kish and Frankel (1974) have found that clustered data, which normally exhibit intracluster correlation, led to sampling variances being inflated.

3.6.1 ESTIMATION OF β.

In a purely design based approach, the OLS estimator is not an unbiased estimator, i.e., that is $E_p\left(\hat{\mathbf{B}}\right) \neq \mathbf{B}$ except in the case of simple random sampling. Under the assumed model of the form

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{3.35}$$

with $E_\xi(\varepsilon|\mathbf{s}) = \mathbf{0}$ and $V_\xi(\varepsilon|\mathbf{s}) = \sigma^2\mathbf{I}$, the OLS estimator is now model unbiased. A design based unbiased alternative to the OLS estimator is a weighted least squares. However, instead of weighting by the inverses of the respondent variances, (as in equation (3.20)), it is weighted by the inverses of the respondent inclusion probabilities,

$$\hat{\mathbf{B}}_{\pi} = \left(\mathbf{X}'\Pi^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Pi^{-1}\mathbf{y}, \tag{3.36}$$

where

$$\Pi = \begin{bmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \pi_n \end{bmatrix} \tag{3.37}$$

and π_i's are the sample inclusion probabilities (i = 1,...,n).

Equation (3.36) is unbiased under the superpopulation model but not generally design unbiased.

Statistical computing packages without a weighting option will use the design based OLS estimator with corresponding design based variance equation (3.18). Those packages that do have a weighting option (using the inverses of sample inclusion probabilities as weights) will give the estimated regression coefficients using equation (3.36) with corresponding variances under the assumed model (3.35). The variance estimate of (3.36) is

$$\mathrm{Var}_{p\xi}(\hat{\mathbf{B}}_{p}) = \left(\mathbf{X}'\Pi^{-1}\mathbf{X}\right)^{-1}\sigma^{2}, \tag{3.38}$$

which is different to the true variance estimate of $\hat{\mathbf{B}}_{\pi}$. The true variance of $\hat{\mathbf{B}}_{\pi}$ is given by

$$\mathrm{Var}_{p\xi}(\hat{\mathbf{B}}_{p}) = \left(\mathbf{X}'\Pi^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Pi^{-1}\mathbf{I}\,\Pi^{-1}\mathbf{X}\left(\mathbf{X}'\Pi^{-1}\mathbf{X}\right)^{-1}\sigma^{2} \tag{3.39}$$

where $V_{\xi}(\mathbf{y}) = \sigma^2\mathbf{I}$. An extension to the case where $V_{\xi}(\mathbf{y}) = \sigma^2\mathbf{V}$ is possible.

In a model based approach, the sample s is assumed fixed, hence the selection probabilities are no longer used in the inference. This differs from the design based approach which uses the information about the sample design in inference (Smith, 1981). However, even though the sample design is not included in the model based inferences it should not be ignored altogether. Nathan and Holt (1980) have shown that selection schemes based on the variable of interest or the auxiliary variable can lead to biased estimators if the selection scheme is ignored.

Suppose we introduced an auxiliary variable, say **Z**, which contains all the information about the clustering or stratification of the population. This variable, also known as the design variable, is often used in the design stage of the survey but mistakenly not used in the regression model (Nathan and Holt, 1980).

Let p(s) denote the probability of selecting a sample s. This is known as the sampling design. Given **Z**, that the design variable is known to the researcher prior to sampling, the probability of selecting s is now a conditional probability, p(s|**Z**). The observed values, **Y** (from the regression equation, 3.35), are also related to **Z** and have probability p(**Y**|**Z**). Now for a noninformative design, the sampling design is independent of the **Y** values, hence p(s,**Y**|**Z**) = p(s|**Z**) p(**Y**|**Z**). Since s contains the information through **Z**, no extra information is gained in s that is not in **Z**. Thus, s and **Y** are conditionally independent, given **Z**. For example, suppose the interest of the study lies in the relationship between the total number of violent attacks (on a victim) and age (from the 1996 National Survey of Crime Victims). And suppose that income is available, which may be used as the design variable.

In Nathan and Holt (1980)'s study, they considered the case where the design variable, **Z**, is known at the design stage (and assuming that **Z** is related to **X** and **Y** in some way) but not used in the regression model. In the simplest model (one dependent, one independent and one design variable),

$$Y = X\beta + \varepsilon, \tag{3.40}$$

where $E_\xi(\varepsilon) = 0$ and $V_\xi(\varepsilon) = \sigma^2$. Let Y, X, and Z follow a trivariate normal distribution with mean and covariance matrix respectively,

$$\mu = \begin{pmatrix} \mu_Y \\ \mu_X \\ \mu_Z \end{pmatrix}, \ \Sigma = \begin{pmatrix} \sigma_{YY} & \sigma_{YX} & \sigma_{YZ} \\ \sigma_{XY} & \sigma_{XX} & \sigma_{XZ} \\ \sigma_{ZY} & \sigma_{ZX} & \sigma_{ZZ} \end{pmatrix}. \tag{3.41}$$

Then the usual OLS estimator of β is generally biased, given Z and s, under the assumed model (3.40). The expected value of the OLS estimator with respect to the superpopulation is given by Nathan and Holt (1980),

$$E_\xi\left(\hat{\beta}|Z,s\right) = \frac{\beta_{YX} + \beta_{YZ}\beta_{ZX}\left(v_Z^2/\sigma_Z^2 - 1\right)}{1 + \rho_{XZ}^2\left(v_Z^2/\sigma_Z^2 - 1\right)} + O_p\left(n^{-1}\right) \tag{3.42}$$

where $\beta_{YX} = \beta$, $\beta_{YZ} = \sigma_{YZ}\sigma_{ZZ}{}^{-1}$, $\beta_{XZ} = \sigma_{XZ}\sigma_{ZZ}{}^{-1}$; $\sigma_Z{}^2$, $v_Z{}^2$, are the population and sample variances of the design variable respectively and $\rho_{XZ}{}^2$ is the correlation of X and Z. Note that the term $O_p(n^{-1})$ is only needed if the distribution is non-normal. It can be seen that the OLS estimator, $\hat{\beta}$ will be conditionally unbiased if the sample variance of Z, $v_Z{}^2$, equals the population variance, $\sigma_Z{}^2$. The unconditional expectation for $\hat{\beta}$ is also biased, i.e., the estimator over repeated samples with respect to the superpopulation and with respect to the design. This estimator is given by

$$E_{p\xi}\left(\hat{\beta}|Z\right) = \beta + \frac{\sigma_Y}{\sigma_X}\left\{\frac{\rho_{YZ\cdot X}\rho_{XZ}\left(1-\rho_{XY}^2\right)\left(1-\rho_{XZ}^2\right)(Q-1)}{1+\rho_{XZ}^2(Q-1)}\right\} + O_p\left(n^{-1}\right) \tag{3.43}$$

where $Q = E_{p\xi}(v_Z{}^2|Z)/\sigma_Z{}^2$ and $\rho_{YZ\cdot X}$ is the conditional correlation of Y and Z, given X. (3.43) will be unbiased if Q = 1, assuming $\rho_{YZ\cdot X}$, $\rho_{XZ} \neq 0$ and

ρ_{XZ}, $\rho_{XZ} \neq 1$. For further details the reader is referred to the study by Nathan and Holt (1980).

An alternative estimator of a purely model based approach involves the maximum likelihood estimation method under the trivariate normal distribution of the superpopulation. This estimate is derived in Nathan and Holt (1980) as

$$\hat{\beta}^* = \frac{v_{YX} + \left(v_{YZ}v_{XZ}/v_Z^2\right)\left(\hat{\sigma}_Z^2/v_Z^2 - 1\right)}{v_X^2 + \left(v_{XZ}^2/v_Z^2\right)\left(\hat{\sigma}_Z^2/v_Z^2 - 1\right)} \tag{3.44}$$

where v.. are the sample covariances. A more general result for a multivariate case is given in Holt et al. (1980). Assuming a multivariate normal superpopulation in which the variables can be partitioned into three groups, say $\mathbf{Y}_i$ are the independent variables, $\mathbf{X}_i$ are the dependent variables and $\mathbf{Z}_i$ are the design variables. The mean vector and covariance matrix is now partitioned as

$$\mu = \begin{pmatrix} \mu_Y \\ \mu_X \\ \mu_Z \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} & \Sigma_{YZ} \\ \Sigma_{XY} & \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZY} & \Sigma_{ZX} & \Sigma_{ZZ} \end{pmatrix}. \tag{3.45}$$

The parameter to be estimated is the coefficient matrix,

$$\beta_{YX} = \Sigma_{YX}\Sigma_{XX}^{-1}. \tag{3.46}$$

The MLE of β_{YX} which Holt et al. (1980) calculates, is given by

$$\hat{\beta}^*_{YX} = \left\{\mathbf{V}_{YX} + \mathbf{V}_{YZ}\mathbf{V}_{ZZ}^{-1}(\hat{\Sigma}_{ZZ}\mathbf{V}_{ZZ}^{-1} - \mathrm{I})\mathbf{V}_{ZX}\right\}\left\{\mathbf{V}_{XX} + \mathbf{V}_{XZ}\mathbf{V}_{ZZ}^{-1}(\hat{\Sigma}_{ZZ}\mathbf{V}_{ZZ}^{-1} - \mathrm{I})\mathbf{V}_{ZX}\right\}^{-1} \tag{3.47}$$

where $\hat{\Sigma}_{ZZ}$ is the covariance matrix of the finite population and the **V**.. are the corresponding sample estimates from the given sample. This estimate (3.47) takes into account the design variables, which clearly the OLS estimate does not. Note also that (3.44) is a special case of (3.47).

When $\mathbf{V}_{ZX} = 0$ or $\hat{\Sigma}_{ZZ} = \mathbf{V}_{ZZ}$, (3.47) will reduce to the OLS estimator, $\hat{\beta} = \hat{\Sigma}_{YX}\hat{\Sigma}^{-1}_{XX}$. Holt et al. (1980) points out that the latter condition will occur for simple random sampling for large n. Also for equal probability designs, $\mathbf{V}_{ZZ}$ may be close to $\hat{\Sigma}_{ZZ}$ but for unequal probability designs, this is not usually the case.

Holt and Scott (1981) argue that it is unusual to find surveys that do not include some cluster sampling. This causes concern over intracluster correlations and its effect of inflating the variance of the estimators compared to the variances under simple random sampling.

Quite often the population will include some sort of clustering and sample surveys will usually include some cluster sampling. Holt and Scott (1981) considered the cluster effects in the assumed regression model under the superpopulation.

Suppose now that a sample of c clusters is selected from a total of C clusters, each cluster containing m elements. The total sample size, $n = \sum_{j=1}^{c} m_j$. Assuming the linear model (3.35), the covariance matrix of **Y** is $V_\xi(\mathbf{Y}) = \sigma^2\mathbf{V}$, where **V** is a block diagonal matrix with submatrices $\mathbf{V}_j$ an $m_j \times m_j$ matrices corresponding to observations from the jth cluster (j = 1,...,c) (Holt and Scott, 1981),

$$\mathbf{V}_j = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}. \tag{3.48}$$

If the intracluster correlations, ρ is known, the best linear unbiased estimator (BLUE) of β is

$$\hat{\beta}_{\text{BLUE}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}. \tag{3.49}$$

If ρ is unknown, iterative methods can be used to find β (Scott and Holt, 1982). In most survey studies, whether the correlations are ignored or unknown, the OLS estimator under the superpopulation model, i.e.,

$$\hat{\beta}_{\text{o}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{3.50}$$

is typically used instead. Both estimators are model unbiased but the variance of (3.50) is greater than the variance of (3.49), i.e., $V_{\xi}(\hat{\beta}_{\text{o}}) > V_{\xi}(\hat{\beta}_{\text{BLUE}})$. Scott and Holt (1982) have also shown that (3.50) can be rewritten as the sum of the between and within cluster components,

$$\hat{\beta}_{\text{o}} = \left(\mathbf{X}_{\text{b}}{}'\mathbf{X}_{\text{b}} + \mathbf{X}_{\text{w}}{}'\mathbf{X}_{\text{w}}\right)^{-1}\left(\mathbf{X}_{\text{b}}{}'\mathbf{X}_{\text{b}}\hat{\beta}_{\text{ob}} + \mathbf{X}_{\text{w}}{}'\mathbf{X}_{\text{w}}\hat{\beta}_{\text{ow}}\right) \tag{3.51}$$

where $\mathbf{X}_{\text{b}}{}'\mathbf{X}_{\text{b}}$ and $\mathbf{X}_{\text{w}}{}'\mathbf{X}_{\text{w}}$ are the between and within sums of squares and cross products respectively, $\hat{\beta}_{\text{ob}}$ and $\hat{\beta}_{\text{ow}}$ are the OLS estimates for the between and within parts respectively. Equation (3.51) assumes that the regression coefficient remains constant for each cluster.

The covariance matrix of $\hat{\beta}_{\text{o}}$ under the standard OLS procedures is given by

$$\mathbf{V}_{\xi}(\hat{\beta}_{\text{o}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{3.52}$$

which is different to the true covariance matrix of $\hat{\beta}_{\text{o}}$. This is given by

$$\mathbf{V}_{\xi}(\hat{\beta}_{\text{o}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D} \tag{3.53}$$

$$\text{where } \mathbf{D} = \mathrm{I} + \left\{ \left(\left(\sum_{j=1}^{c} \mathrm{m}_j \mathbf{X}'_{bj} \mathbf{X}_{bj} \right) (\mathbf{X}'\mathbf{X})^{-1} \right) - \mathrm{I} \right\} \rho. \tag{3.54}$$

From equations (3.53) and (3.54), it is clear that if $\rho = 0$, then $\mathbf{D} = \mathbf{I}$ and the variance will be equal to that of the OLS variance of a simple random sample. However, since intracluster correlation does exist in sample surveys, with cluster sampling, it is not likely that $\rho = 0$, hence $\mathbf{D}$ is known as the inflation factor (Scott and Holt, 1982). It is also termed as the misspecification effect (Holt and Scott, 1981) as it specifies how much error there is when the OLS variance is used when clustering is present. Details of the derivation of the inflation factor can be found in appendix 1 of Scott and Holt (1982).

3.7 THE AGE OF COMPUTERS

Times and technology have changed since Pearson first estimated β in 1902 (cited in Smith, 1981). With the aid of computers, software on statistical procedures such as regression coefficient estimation have been developed. However, most of these statistical packages involve only standard OLS estimates which are sufficient for simple random sample data. Similarly the variance is calculated as the standard variance from the standard OLS procedures. In some statistical packages such as SPSS and SAS, these give weighted estimates in the form (3.21) or using the inverses of the sample inclusion probabilities as the weights. Again these packages assume simple random sampling in the collection of data.

With complex surveys, the population is clustered or stratified. Standard OLS procedures do not take into account the population structure. However, there are now statistical packages that allow for complex surveys, for example, SUPERCARP, PC CARP and SUDAAN. Most analysts still tend to use the more popular, standard packages such as SPSS and SAS that assume simple random data. Although there exists a weighting option in these standard packages, and the estimated coefficients are the same, the conclusions may not be (For example, the variance estimate (3.38) used in statistical packages with weights is still not equivalent to the true variance (3.39).)

Smith (1981) points out that when using statistical packages which are based on OLS, data which do not have any population structure are the only appropriate data to be analysed. Any population suspected of clustering or stratification effects are best analysed with suitable computer packages that allow for structuring. Kish and Frankel (1974) has found that computing variances with standard procedures often leads to considerable underestimation and hence stresses the importance of using the right statistical package for estimation.

Section 6.2.2 gives an example of regression analysis.

3.8 SUMMARY

Regression analysis is widely used in the analysis of complex survey data. However, OLS estimators are less than ideal, especially when the population has clustering. The standard errors will appear much smaller than they really are and wrong conclusions could result.

Estimators have been used to adjust for the sample selection by incorporating the design variables into the regression model. OLS estimators for this regression model are found to be biased. Alternative estimators include a design based estimator, (3.36), which is design unbiased even under a superpopulation, a MLE which is approximately asymptotically unbiased and a BLUE taking into account the intracluster correlations.

When using computer packages, it is important to use packages that allow for complex designs. Although standard statistical packages that include weighting options give accurate estimates, they also give inaccurate standard errors which could lead to wrong conclusions.

CHAPTER FOUR

CONTINGENCY TABLES

Frequency tables are used widely in sample surveys. Otherwise known as contingency tables, categorical variables are cross-classified to form tables. Each cell usually contains count data. The count data are usually random variables of a binomial, multinomial, product-multinomial or Poisson distribution. A multinomial distribution is simply a generalization of a binomial distribution (Christensen, 1990). Similarly, product-multinomials are independent multinomials. Note that the Poisson distribution is related to the multinomial distribution by conditioning on the sum of the independent Poisson variates. Details of the proof can be found in Bishop, Fienberg and Holland (1975).

The assumptions for a binomial distribution are that:

- There are a fixed number of trials.
- Each trial is identical and independent.
- There are only two outcomes such as 'success' and 'failure'.
- The probability of success is constant.

Let the probability of success be denoted by p and let n denote the count i.e., the total number of times that the trial is performed.

In a multinomial situation, there are still a fixed number of trials that are independent and identical. There are now q outcomes and the probability is p_j where j = 1,...,q associated with the jth outcome and n_j is the count of the jth outcome occurring. Note that $\sum_{j=1}^{q} p_j = 1$.

Suppose we now have r = 1,...,t independent multinomials. The probability of each cell count, p_{ij}, is the product of the multinomial probabilities for each r. This is known as the product-multinomial distribution.

When the number of trials gets rather large (an arbitrary number) and the probabilities of occurrences are getting small, the Poisson distribution is much more useful. A Poisson process describes a number of occurrences or events over a specified time or space (Kohler, 1985). The assumptions for the Poisson distribution are that each event is independent and that the events occur randomly over time or are spread evenly over space. The probability of each event is assumed to be the same.

Statistical methods have been long employed to analyse these contingency tables, from simple one way tables to more complex multiway tables.

4.1 ONE WAY TABLES

A one way table is used commonly for tabulating a single categorical variable. The simplest table is a table with two cells and the data is binomial. For example, we can use a one way table to tabulate the gender of the respondents of the survey from the 1996 survey. This table will have two cells, one for males and one for females (disregarding all of the non-responses).

4.1.1 GOODNESS OF FIT TEST

A simple goodness of fit test tests the observed cell proportions, p_i, with some hypothesised cell proportions, p_{oi}.

Consider a table with 2 cells, i.e., I = 2, the null hypothesis for this test is $H_o : p_i = p_{oi}$, i = 1,2 against the alternative, H_1: $p_i \neq p_{oi}$ for at least one i. Karl Pearson proposed a test statistic for testing the null hypothesis as early as 1900 (Larsen and Marx, 1986). Named after him, the Pearson chi-squared test statistic for testing the more general goodness of fit (I>2), the hypothesis becomes $H_o : p_i = p_{oi}$ and the test statistic is given by

$$X_p^2 = \sum_{i=1}^{I} \frac{\left(n_i - np_{oi}\right)^2}{np_{oi}}$$

$$= n\sum_{i=1}^{I} \frac{\left(p_i - p_{oi}\right)^2}{p_{oi}} \qquad (4.1)$$

where n_i = observed cell frequencies ($n_i = np_i$),

p_i = observed cell proportions,

p_{oi} = hypothesised cell proportions

and $\sum_{i=1}^{I} n_i = n$. Note that (4.1) is asymptotically chi-squared with one degree of freedom (df).

Other alternative test statistics were developed such as the likelihood ratio test statistic, based on maximising the log likelihood function of the multinomial. The likelihood ratio statistic for testing H_o is (Lehtonen and Pahkinen, 1995)

$$G^2 = 2n\sum_{i=1}^{I} p_i \log\left(\frac{p_i}{p_{oi}}\right). \tag{4.2}$$

A similar test statistic to the Pearson is the Neyman test statistic,

$$X_n^2 = n\sum_{i=1}^{I} \frac{(p_i - p_{oi})^2}{p_i}. \tag{4.3}$$

This differs from the Pearson statistic only in the denominator. The Neyman statistic uses observed cell proportions rather than hypothesised cell proportions in the denominator. Both (4.2) and (4.3) are also asymptotically chi-squared with one df. Since the most commonly used test statistic is the Pearson statistic, and the Neyman is very similar to the Pearson, the Neyman test statistic will not be mentioned from now on.

On the side, the Pearson chi-square and the likelihood ratio chi-square are from the Power-divergence family of statistics (Read and Cressie, 1988). The Power-divergence Statistic is given by

$$\frac{2n}{\lambda(\lambda+1)}\sum_{i=1}^{I} p_i\left[\left(\frac{p_i}{p_{oi}}\right)^{\lambda} - 1\right] \tag{4.4}$$

and λ is some real-value parameter, $-\infty < \lambda < \infty$. When $\lambda = 1$, equation (4.4) reduces down to the Pearson chi-square statistic, (4.1), and as $\lambda \to 0$, (4.4) reduces to the likelihood ratio statistic, (4.2).

In most real situations, the data collected will not be simply categorised into two cells. Often, more than one categorical variable is required to analyse the data. Now we turn to two way tables in which there are two categorical variables.

4.2 TWO WAY TABLES

With two categorical variables, the first variable with I levels and the second with J levels, the observations on these variables can be displayed with an I×J contingency table. The Pearson and likelihood statistical tests will be discussed with the simplest two way table, the 2×2 table and then generalized to I, J ≥ 2 levels.

Suppose we are interested in finding out the ethnicity of the respondents as well as their gender. We can cross classify these variables and produce a two way table which gives information about, say, how many respondents were female Pacific Islanders.

4.2.1 INDEPENDENCE TEST

The independence test of a two way table with multinomial sampling is to test whether the categories have any association between them. The test hypothesis will be $H_o : p_{ij} = p_{i+}\, p_{+j}$ for i,j = 1,2, where the p_{i+} and p_{+j} are the ith row and the jth column totals respectively. The tables usually contain observed counts, n_{ij}. Expected counts are calculated by $m_{ij} = np_{i+}\, p_{+j}$ but the p_{ij}'s are not usually known (note that n is the count total). Substituting the p_{ij} with n_{ij}/n, we can calculate the estimated expected counts. These are given by $\hat{m}_{ij} = \frac{n_{i+}n_{+j}}{n}$. The Pearson chi-square statistic then equals to

$$X_p^2 = \sum_i^I \sum_j^J \frac{\left(n_{ij} - \hat{m}_{ij}\right)^2}{\hat{m}_{ij}} \qquad (4.5)$$

and df = (I-1)(J-1) = 1. In the case of the 2×2 table, this is asymptotically chi-square with 1 degree of freedom, i.e., χ_1^2. More generally, for i,j > 2 the Pearson test statistic still follows the chi-square distribution with (I-1)(J-1) df, $\chi_{(I-1)(J-1)}^2$.

The likelihood ratio chi-square is given by

$$G^2 = 2\sum_{i}^{I}\sum_{j}^{J} n_{ij}\log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right). \tag{4.6}$$

This is also asymptotically $\chi_{(I-1)(J-1)}^2$.

4.2.2 HOMOGENEITY TEST

This tests the homogeneity of one variable over the values of some other. Generally the data are distributed as product multinomials. That is, testing whether each level of variable I is distributed similarly within each level of variable J. In sample surveys, this may be testing whether independent samples of different regions are distributed the same. Usually the row numbers are fixed and the row proportions are equal to 1, i.e., $p_{i1} + p_{i2} + ... + p_{iJ} = 1$ for each i = 1,2,...,I.

An example of a homogeneity test hypothesis, the ethnicity distributions are the same for males and females.

The test hypothesis for the Pearson statistic is given by $H_o : p_{1j} = p_{2j}$, for j = 1,2 (for a 2×2 table), and the statistic is given by

$$X_p^2 = \frac{n_1 n_2 (p_{11} - p_{12})^2}{(n_1 + n_2)p_{+1}(1 - p_{+1})} \tag{4.7}$$

where n_1 is the sample size of row 1 and n_2 is the sample size of row 2.

For $j \geq 2$,

$$X_p^2 = \sum_i^I n_{i+} \sum_j^J \frac{\left(p_{ij} - p_{+j}\right)^2}{p_{+j}}. \qquad (4.8)$$

The test statistic for homogeneity and for independence are equivalent because, although the formulae are different, they are both essentially testing the same thing.

4.3 MULTIWAY TABLES

Let us consider now, an I×J×K table. Here there are I rows, J columns and K layers of the table which gives a three dimensional, or three way, table. The total cell count is denoted by n_{ijk}, with probability of each cell is p_{ijk}. The expected count of each cell is denoted by m_{ijk}. This is clearly more complicated to analyse than a one way or two way table. If the level of K is small, say 2 or 3, it is possible to treat each level of K as a separate table, i.e., K two way tables. However, the conclusions from these separate two way tables may not be accurate (Christensen, 1990). It is deemed necessary to keep the table as a three dimensional table so that accurate information is obtained.

The data set can get quite large in a three way table and that can cause difficulties also. For example, if each of the variables or categories only had two levels, a 2×2×2 table consisting of eight cells, it is still relatively easy to analyse. Given more levels at each category, say 5 levels, a 5×5×5 table produces 125 cells which will complicate analyses. For example, a three way table can consist of gender (two levels), ethnicity (five levels) and occupation (eight levels) which would give a total of $2\times5\times8 = 80$ cells.

Four way and higher dimensional tables further complicate the analyses and interpretation is more difficult. The usual simple chi-squared tests are no longer suitable and some model is needed to evaluate and describe the relationships between the variables (Agresti, 1990). A model that fits the data well, will evaluate the effects of the variables and other associations between the variables. The problem then lies in finding a suitable model. The next section introduces a family of models, the Generalized Linear Models, that consist of important models for categorical data.

4.4 GENERALIZED LINEAR MODELS

Generalized Linear Models (GLM's) can be thought of as an extension of simple linear models of the form such as (3.2), $y = B_0 + B_1x$ (McCullagh and Nelder, 1983). For a GLM, there are $x_i,...,x_n$ observations giving $y_i = B_0 + B_1x_i$, $i = 1,...,n$, or in the multiple linear case, $\mathbf{y}_i = \mathbf{B}^T\mathbf{x}$. where $\mathbf{B}^T$ is the vector of parameters transposed. There are three components to a GLM. The first component is that each of the y_i are assumed to be random, independent realisations of a random variable, Y_i. These Y_i's have the same distribution from the exponential family of distributions with constant variance, but with different means, μ_i i.e., $E(\mathbf{Y}) = \mu$. The second component of a GLM is a systematic component. It is a linear function of the explanatory variables, $\mathbf{B}^T\mathbf{x}_i$. The $\mathbf{B}^T\mathbf{x}_i$ are also sometimes denoted by η. Thirdly a link is needed to relate the random and systematic components. This link is some monotonic function of the explanatory variables, say $g(\mu) = \mathbf{B}^T\mathbf{x}_i$. An identity link is given when $g(\mu) = \mu$, i.e., when $\eta = \mu$.

Summarising the components in matrix notation,

$$E(\mathbf{Y}) = \eta \text{ and } \eta = G(\mu) = \mathbf{XB} \tag{4.9}$$

where $\mathbf{X}$ is an n×p matrix, known as the model matrix.

Simple linear models have a normal distribution as the random component whereas in a GLM, it can be of any distribution from the exponential family. Also the link function for simple linear models is the identity link. Two common models are outlined briefly to illustrate the components of the GLM.

4.4.1 LOG-LINEAR MODELS

Log-linear models are used mainly for modelling the expected cell frequencies in a contingency table (Payne, 1977). These frequencies are usually of a Poisson process (which belongs to the natural exponential family of distributions). The link for log-linear models is, not surprising, the log link, i.e., $\log(\mu) = \eta$.

Let $\hat{m}_{ij}$ be the expected cell count corresponding to the ith row and jth column of a 2×2 table.

The two way table would look like,

		Variable 2		
		1	2	Total
Variable 1	1	$n_{11}(\hat{m}_{11})$	$n_{12}(\hat{m}_{12})$	n_{1+}
	2	$n_{21}(\hat{m}_{21})$	$n_{22}(\hat{m}_{22})$	n_{2+}
	Total	n_{+1}	n_{+2}	n

where the bracketed $\hat{m}_{ij} = \dfrac{n_{i+}n_{+j}}{n}$ are the expected cell frequencies and the n_{ij} are the actual cell frequencies.

By taking the log of the expected values,

$$\log\left(\hat{m}_{ij}\right) = \log\left(\frac{n_{i+}n_{+j}}{n}\right)$$

$$= \log\left(n_{i+}\right) + \log\left(n_{+j}\right) - \log(n)\,, \qquad (4.10)$$

this gives a linear equation for the expected counts.

Equation (4.10) can be re-expressed as

$$\log\left(\hat{m}_{ij}\right) = u + u_{1(i)} + u_{2(j)} \qquad (4.11)$$

where u is the overall mean of the log of the expected counts, i.e.,

$$u = \frac{1}{IJ}\sum^{I}\sum^{J}\log\left(\hat{m}_{ij}\right);$$

$$u_{1(i)} = \left(\frac{1}{J}\sum^{J}\log\left(\hat{m}_{ij}\right)\right) - u\,,$$

and $$u_{2(j)} = \left(\frac{1}{I}\sum^{I}\log\left(\hat{m}_{ij}\right)\right) - u\,.$$

Note that $\sum_{i} u_{1(i)} = \sum_{j} u_{2(j)} = 0$.

If there is an interaction between variable 1 and variable 2, then the model would become

$$\log\left(\hat{m}_{ij}\right) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \tag{4.12}$$

where $u_{12(ij)}$ is the interaction term between variables 1 and 2.

The model (4.12) is known as a saturated model as it includes all possible u terms to explain any expected values, $\hat{m}_{ij}$. It is also the general log-linear model for a two way contingency table (Agresti, 1990).

For a three dimensional table with I rows, J columns and K layers, the observed counts will be denoted as n_{ijk} and the expected counts, $\hat{m}_{ijk}$. One log-linear model for a three way table can be written as

$$\log\left(\hat{m}_{ijk}\right) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} \tag{4.13}$$

and including the interaction terms, the saturated model is

$$\log\left(\hat{m}_{ijk}\right) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}. \tag{4.14}$$

Fienberg (1977) points out that for higher dimensional tables, it is quite straightforward, i.e., for a four way contingency table, the model is the same as (4.14) but also including the main effect term for the fourth variable, $u_{4(l)}$, and all of the associated interaction terms, $u_{124(ijl)}, u_{134(ikl)}, u_{234(jkl)}, u_{1234(ijkl)}$.

4.4.2 LOGISTIC REGRESSION

With log-linear models the cells contained Poisson counts. Logit models involve data that have a binomial distribution (also belonging to the exponential family) and that the outcomes are dichotomous, i.e., either 0 or 1. For example, in medical situations where testing a particular drug to aid recovery of patients, the two outcomes to the drug would be 'cured' or 'not cured'. Usually one variable is a response or dependent variable and the others are the explanatory or independent variables. Here the log of the odds of the response variable is modelled. Although many logit models turn out to be similar to log-linear models, the main difference between the two types of models is that the primary interest of a logit model is the response variable. Relationships between the explanatory variables ($\mathbf{x}_1,\ldots,\mathbf{x}_p$) are of no direct interest in the logit models (Christensen, 1990).

Let P(Y = 1) = π_i be the probability of getting an outcome of one and P(Y = 0) = 1- π_i be the probability of getting zero. The odds of getting an outcome of one is given by the odds ratio, $\frac{\pi_i}{1-\pi_i}$.

Suppose that the dependence of π on $\mathbf{x}$ is linear,

$$\pi(\mathbf{x}_i) = \mathbf{XB} = \sum_{j=1}^{p} \mathbf{x}_j \mathrm{B}_j \tag{4.15}$$

for some unknown coefficients B_j. Note that B_1 is the intercept term and $\mathbf{x}_1 = 1$. However, $\pi(\mathbf{x}_i)$ exists only between the interval (0,1) and $\mathbf{x}$ can exist between the interval $(-\infty,\infty)$. Some link function is required to map the interval (0,1) to $(-\infty,\infty)$. There are three such link functions associated with logit models. The first model is the most common, known as the logistic regression model and its link function is given by the log of the odds ratio.

A logistic regression function is given by,

$$\pi\left(\mathbf{x}_i\right) = \frac{\exp(\mathbf{XB})}{1+\exp(\mathbf{XB})}. \tag{4.16}$$

A function of the equation (4.16) implies a curvilinear relationship (Agresti, 1990).

The link function is derived from

$$\frac{\pi\left(\mathbf{x}_i\right)}{1-\pi\left(\mathbf{x}_i\right)} = \exp(\mathbf{XB}), \tag{4.17a}$$

$$\log\left(\frac{\pi\left(\mathbf{x}_i\right)}{1-\pi\left(\mathbf{x}_i\right)}\right) = \mathbf{XB} \tag{4.17b}$$

which gives the logit link,

$$\eta = \log\left(\frac{\pi_i}{1-\pi_i}\right). \tag{4.18}$$

The second model is the probit model with link function

$$\eta = \Phi^{-1}(\pi_i), \tag{4.19}$$

where Φ^{-1} is the inverse of the cumulative normal distribution function.

Thirdly, a complementary log-log link,

$$\eta = \ln(-(\ln(1-\pi_i))). \tag{4.20}$$

The last two models briefly mentioned are alternative models to the logistic regression. Further details about them can be found in Agresti (1990).

4.5 CONTINGENCY TABLES ON COMPLEX SURVEYS

So far, all of the tests for contingency tables mentioned are based on the assumption of binomial, multinomial, product-multinomial or Poisson sampling. However, most of the survey designs involve some stratification or clustering (or both) which violates the assumptions associated with each distribution (independence of observations). This can cause misleading results if no adjustments are made to account for the survey design (Scott and Rao, 1981).

They (Scott and Rao, 1981) have found that clustering and stratification affects independence chi-square tests and more seriously, homogeneity chi-square tests. These results were also found in an earlier paper by Holt, Scott and Ewings (1980).

Thomas and Rao (1987) found, in their Monte Carlo study, that cluster sampling and stratification can lead to greatly inflated type I error rates (a type I error is rejecting the null hypothesis when it is true) on standard test statistics such as the Pearson chi-square and the likelihood ratio.

Rao and Scott (1981) have shown that in two way tables, the standard Pearson chi-square, X_p^2, and the likelihood ratio test statistic, G^2, are distributed asymptotically as χ_{I-1}^2. For complex designs, X_p^2 and G^2 are distributed as a weighted sum $\delta_1 W_1 + \delta_2 W_2 + ... + \delta_{I-1} W_{I-1}$ of I-1 independent χ_1^2 variables W_i ($i = 1,...,I-1$). The weights, $\delta_1 ... \delta_{I-1}$, are the eigenvalues of the misspecification effects (hereafter called the design effects) matrix, $\mathbf{D} = \hat{\mathbf{P}}^{-1}\hat{\mathbf{V}}$, where $\hat{\mathbf{P}}$ is the estimated multinomial covariance matrix under H_0,

$\hat{\mathbf{P}} = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^{\mathrm{T}}$ and $\hat{\mathbf{V}}/\mathrm{n}$ is the estimated covariance matrix under the actual design. When I = 2 categories, the weights simply reduce down to the design effect, d, of the proportion estimate, $p = p_1$. And in the special case where $\delta_i = 1$, then the weighted sum reduces to $\chi_{I-1}{}^2$.

It has been discussed by many authors (e.g., Holt et al., 1980; Rao and Scott, 1981; Bedrick, 1983) that adjustments are necessary to the standard chi-square test statistics when there has been stratification or clustering in the sample design. A simple adjustment to the test statistics is dividing by the design effect, d, in the case of I = 2, $X_p{}^2/d$ or G^2/d.

When the cells in the table exceed two, I > 2, Fellegi (1980) proposed an adjustment by dividing the standard chi-square by the estimated average cell deff,

$$X^2 \big/ \hat{\bar{\delta}} \tag{4.21}$$

where $\hat{\bar{\delta}} = \sum_{i=1}^{I} \frac{\delta_i}{I}$. This is known as a first order correction. Equation (4.21) is now approximately χ^2 with I-1 degrees of freedom.

Rao and Scott (1981) gives a first order correction using an estimate of the mean generalized deff. The adjusted Pearson chi-square and likelihood ratio is given by

$$X_p^2(\hat{\delta}.) = X^2 \big/ \hat{\delta}. \tag{4.22}$$

and

$$G^2(\hat{\delta}.) = G^2 \big/ \hat{\delta}. \tag{4.23}$$

respectively, where $\hat{\delta}. = \mathrm{tr}(\mathbf{D})/(I-1) = \sum_{i=1}^{I} \hat{v}_{ii}/p_i(I-1)$. Note that $\hat{\delta}.$ is not the same as $\hat{\hat{\delta}}$.

We know that generally cluster sampling yields a design effect, d, of greater than one and that stratified samples have a design effect less than one. For a stratified design, the standard chi-square tests will give conservative tests. However this is not the case for cluster designs. First order corrections must be made to the test statistics or the test statistics will become inflated and will no longer be asymptotically χ^2. Suppose we take δ_1 to be the maximal deff, $\delta_1 = \sup_{\mathbf{c}} \frac{\mathbf{c}'\hat{\mathbf{V}}\mathbf{c}}{\mathbf{c}'\hat{\mathbf{P}}\mathbf{c}}$, for any arbitrary (I-1) vector, **c**, of real co-ordinates where sup denotes supremum. If δ_1 can be estimated or found, then a first order correction, X^2/δ_1 will give a conservative test to test statistics, (4.21), (4.22) and (4.23) (Chaudhuri and Stenger, 1992).

A first order correction is satisfactory (Binder et al., 1987) when the full covariance matrix, **V**, is unavailable or cannot be estimated. If the full covariance matrix of the sample design was available, then a better and more powerful adjustment is obtainable. This adjustment is based on the Satterthwaite (1946) method, known as the second order correction and given by

$$X_p^2\left(\hat{\delta}.,\hat{a}\right) = X_p^2\left(\hat{\delta}.\right)/\left(1+\hat{a}^2\right) \qquad (4.24)$$

$$\text{and } G^2\left(\hat{\delta}.,\hat{a}\right) = G^2\left(\hat{\delta}.\right)/\left(1+\hat{a}^2\right) \qquad (4.25)$$

where $\hat{a}^2 = \left\{ \sum_{i=1}^{I-1} \frac{\hat{\delta}_i^2}{\left[(I-1)\hat{\delta}.^2 \right]} \right\} - 1$ which is the coefficient of variation of the eigenvalues, δ_i : $i = 1,\ldots,I$ of the estimated design effects matrix and $\sum_{i=1}^{I-1} \hat{\delta}_i^2 = \mathrm{tr}\left(\mathbf{D}^2\right)$. X_p^2 and G^2 are both asymptotically chi-square with $\{(I-1)/(1+a^2)\}$ degrees of freedom.

Rao and Thomas (1989) point out that when the $\hat{\delta}_i$'s have large variation, first order corrections tend to inflate the type I error rates. Hence second order corrections control the type I error better and provides a much more useful adjustment when $\hat{a}^2$ is significantly different to zero (Binder et al., 1987).

4.5.1 WALD TESTS

An alternative test which adjusts for intracluster correlation among the clusters of a complex design is the design based Wald test. This test differs from the Pearson, Neyman and the likelihood ratio test as it requires the estimate of the covariance matrix, $\hat{\mathbf{V}}$. However the Wald test does not require any first order corrections nor second order corrections as it automatically adjusts for a complex design. In the simple Goodness of Fit test, the Wald test statistic for $I = 2$ cells is given by

$$X_w^2 = \frac{\left(p - p_o\right)^2}{\hat{v}} \tag{4.26}$$

where $\hat{v}$ is the estimate of the variance of p. This is asymptotically χ_1^2 for a complex design.

For I > 2 cells the Wald test statistic (written in matrix form),

$$X_w{}^2 = (\mathbf{p} - \mathbf{p}_o)^T \hat{\mathbf{V}}^{-1} (\mathbf{p} - \mathbf{p}_o) \tag{4.27}$$

where $\hat{\mathbf{V}}$ is the covariance estimator of the true covariance estimator, **V**/n of the proportion vector **p**. $X_W{}^2$ follows a chi-square with I-1 df. Paraphrasing Lehtonen and Pahkinen (1995), with large enough sample clusters and a small number of cells, the covariance estimator will be consistent and (4.27) will give a valid test statistic. The Wald test is not ideal in all situations however. Rao and Thomas (1989) suggests that 'it can lead to inflated type I error rates in finite samples, as the degrees of freedom for estimating the covariance matrix, **V**, decreases and the number of cells in the table, I, increases' (pg 102).

It has been shown that an another alternative test, an F-corrected Wald test statistic, is more stable at controlling the type I error rate (Thomas and Rao, 1987). The F-corrected Wald statistic is given by

$$FX_W{}^2 = \frac{f - (I-1) + 1}{f(I-1)} X_W^2 \tag{4.28}$$

where f = number of clusters - number of strata. This F-corrected Wald test statistic is assumed to follow an F-distribution with (I-1) and (f-(I-1)+1) df.

Lehtonen and Pahkinen (1995) suggest that an F-correction can also be applied to the first order corrections proposed by Rao and Scott (1981), and can sometimes give better results. The F-corrected tests are more robust in situations where there is a small number of sample clusters which is supported by Thomas and Rao (1987)'s Monte Carlo study. The F-corrected first order statistics are given by

$$FX^2\left(\hat{\delta}.\right) = \frac{X^2\left(\hat{\delta}.\right)}{(I-1)} \tag{4.29}$$

$$\text{and } FG^2(\hat{\delta}.) = \frac{G^2(\hat{\delta}.)}{I-1} \tag{4.30}$$

which are treated as F-variables with I-1 and f df.

4.5.2 LOG-LINEAR TEST STATISTICS

The log-linear model of the form (4.10) can also be written as

$$\log(\hat{p}_{ij}) = u(\beta)1 + \mathbf{X}\beta \tag{4.31}$$

where **X** is the I×p model matrix of full rank and β is a p-vector of parameters (to be estimated). **1** is the I-vector of 1's and Rao and Thomas (1989) call u(β) the normalising factor, i.e., ensures that $\sum_i \sum_j \hat{p}_{ij} = 1$. Note that equation (4.31) is now modelling the expected proportion of counts rather than the expected counts.

Parameter estimation by maximum likelihood estimation is possibly the most common method. Difficulties arise in finding the likelihood functions for general log-linear (and logit) models, and so an iterative procedure is used.

Suppose we let all the cells in a contingency table be re-numbered as i = 1,...,I. For example, a 2×2 table where I = 2 and J = 2 have 4 cells in total. Re-numbering these 4 cells will give i = 1,...,4. Then for a log-linear model, the Pearson Goodness of Fit test statistic is

$$X^2_{LL} = n\sum_{i=1}^{I} \frac{\left(p_i - \hat{p}_i(\hat{\beta})\right)^2}{\hat{p}_i(\hat{\beta})} \tag{4.32}$$

and the likelihood ratio test is

$$G^2_{LL} = 2n\sum_{i=1}^{I} p_i \log\left(\frac{p_i}{\hat{p}_i\left(\hat{\beta}\right)}\right) \tag{4.33}$$

where p_i are the cell probabilities (or proportions) and $\hat{p}_i\left(\hat{\beta}\right)$ are the estimated cell probabilities found by the iterative proportional fitting (IPF) method from the likelihood equations (Rao and Thomas, 1989),

$$\mathbf{X}'\hat{\mathbf{p}}\left(\hat{\beta}\right) = \mathbf{X}'\mathbf{p}\,. \tag{4.34}$$

Parameter estimation using the IPF method can be found in Bishop et al. (1975) and also in Read and Cressie (1988).

The equations (4.32) and (4.33) are both shown to be distributed as a weighted sum of $\chi_1{}^2$ variables by Rao and Scott (1984; cited in Rao and Thomas, 1989). The weights are eigenvalues of a more complex design effects matrix than **D** (Rao and Thomas, 1989). This design effects matrix is given by

$$\hat{\Delta} = \left(\mathbf{C}'\mathbf{D}(\mathbf{p})^{-1}\mathbf{C}\right)^{-1}\left(\mathbf{C}'\mathbf{D}(\mathbf{p})^{-1}\mathbf{V}\mathbf{D}(\mathbf{p})^{-1}\mathbf{C}\right) \tag{4.35}$$

where $\mathbf{D}(\mathbf{p})$ = diag($\mathbf{p}$), $\mathbf{V}$ is the covariance matrix under the complex design and $\mathbf{C}$ is any I×(I-1) full rank matrix such that $\mathbf{C}^T\mathbf{1} = \mathbf{0}$ and $\mathbf{C}^T\mathbf{X} = \mathbf{0}$. By choosing an appropriate **C**, the log-linear model of the appropriate terms may be formed. The estimated mean generalized deff for the log-linear model is given by

$$\hat{\delta}_{\cdot LL} = \mathrm{tr}\left(\hat{\mathbf{D}}\right)/(I-1)\,. \tag{4.36}$$

Equations (4.32) and (4.33) can be adjusted with a first order correction using the mean generalized deff, (4.36), as before.

For the independence test in a log-linear model, there are several tests of independence if there are more than two variables, e.g. rows are independent of columns and layers (but the columns and layers are not necessary independent). Here, only the complete independence test will be presented. That is each level is independent of another. The test hypothesis for a three way table is, H_o: $p_{ijk} = p_{i..}\ p_{.j.}\ p_{..k}$, and the Pearson and likelihood test statistics respectively are

$$X^2_{LL,Ind} = \sum_i \sum_j \sum_k \frac{\left(p_{ijk} - \hat{p}_{ijk}\right)^2}{\hat{p}_{ijk}} \tag{4.37}$$

$$G^2_{LL,Ind} = 2\sum_i \sum_j \sum_k p_{ijk} \log\left(\frac{p_{ijk}}{\hat{p}_{ijk}}\right) \tag{4.38}$$

The formula for the mean generalized deff is

$$\hat{\delta}._{LL,Ind} = \frac{\sum_i \sum_j \sum_k \frac{\hat{p}_{ijk}\left(1-\hat{p}_{ijk}\right)}{\hat{p}_{i..}\hat{p}_{.j.}\hat{p}_{..k}} \hat{D}_{ijk} - \sum_i \left(1-\hat{p}_{i..}\right)\hat{D}_i - \sum_j \left(1-\hat{p}_{.j.}\right)\hat{D}_j - \sum_k \left(1-\hat{p}_{..k}\right)\hat{D}_k}{\left(IJK - I - J - K + 2\right)}$$

(4.39)

where $\hat{D}_{ijk} = \dfrac{n\hat{v}\left(\hat{p}_{ijk}\right)}{\hat{p}_{ijk}\left(1-\hat{p}_{ijk}\right)}$ is the estimated deff of $\hat{p}_{ijk}$ and $\hat{D}_t$ is the estimated deff of the row marginal $\hat{p}_t$, for t = i,j,k. Thus with a first order correction, $X^2_{LL,Ind}\big/\hat{\delta}._{LL,Ind}$ and $G^2_{LL,Ind}\big/\hat{\delta}._{LL,Ind}$ are treated as a chi-square variable with (IJK-I-J-K+2) df.

If the full covariance matrix is available or can be estimated, it is also possible to construct a Wald test statistic for the log-linear model providing the estimated proportions are non zero. The Wald test statistic given in Rao and Thomas (1989) is approximately $\chi_{I-1}{}^2$:

$$X^2_{w.LL} = (\mathbf{C}'\log(\hat{\mathbf{p}}))'\left[\mathbf{C}'\mathbf{D}(\hat{\mathbf{p}})^{-1}\hat{\mathbf{V}}\mathbf{D}(\hat{\mathbf{p}})^{-1}\mathbf{C}\right](\mathbf{C}'\log(\hat{\mathbf{p}})). \qquad (4.40)$$

4.5.3 LOGISTIC REGRESSION TEST STATISTICS

For the logistic regression model, likelihood functions are also difficult to obtain for complex designs. Lehtonen and Pahkinen (1995) states that there is '...no convenient likelihood functions available...' (pg 242). A modification of the traditional maximum likelihood estimation method, the 'pseudo' maximum likelihood is used instead. This procedure accounts for the intracluster correlation in the case of complex designs. It is similar to a WLS procedure as it includes a diagonal matrix of estimated subpopulation proportions, i.e., n_i/n where n is the sample size of the ith subpopulation and n is the total sample size. The logistic regression function can also be written as $\pi = \mathbf{f}(\beta)$ i.e., the proportions of π is some function of β. Estimates of β and $\pi = \mathbf{f}(\beta)$ are found iteratively on the likelihood equations (see also Roberts et al., 1987),

$$\mathbf{X}'\mathbf{D}(\mathbf{n})\hat{\mathbf{f}} = \mathbf{X}'\mathbf{D}(\mathbf{n})\mathbf{p}. \qquad (4.41)$$

Here $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_p)^T$ is an I×p matrix of rank p, $\mathbf{D}(\mathbf{n}) = \text{diag}(\mathbf{n})$ ($\mathbf{n}$ is the I-vector of subpopulation sizes), $\hat{\mathbf{f}} = \mathbf{f}(\hat{\beta})$ (I-vector of parameter estimates) and $\mathbf{p} = n_{i1}/n_i$ is the vector of sample proportions (n_{i1} is the cell total (number of success responses) and n_i is the sample size of each cell). For a more detailed discussion, consult Roberts et al. (1987) or Lehtonen and Pahkinen (1995).

A chi-square test statistic for a goodness of fit test can be constructed for a logistic model. Following Roberts et al. (1987) this gives

$$X_{LR}^{2} = n\sum_{i=1}^{I}\frac{\left(p_i - \hat{f}_i\right)^2 w_i}{\hat{f}_i\left(1-\hat{f}_i\right)}, \tag{4.42}$$

and for the likelihood test statistic,

$$G_{LR}^{2} = 2n\sum_{i=1}^{I} w_i\left\{\tilde{p}_i \log\left(\frac{\tilde{p}_i}{\hat{\pi}_i}\right) + \left(1-\tilde{p}_i\right)\log\left(\frac{1-\tilde{p}_i}{1-\hat{\pi}_i}\right)\right\} \tag{4.43}$$

where $\tilde{p}_i = \frac{\hat{n}_{i,1}}{\hat{n}_i}$ is the ratio estimate of **p** and $w_i = \frac{\hat{n}_i}{\hat{n}}$.

Under multinomial sampling, equations (4.42) and (4.43) are distributed as chi-square variables with I-p df. For any complex design, they are distributed as a weighted sum of χ_1^2 variables. The weights are estimated by the eigenvalues of the matrix,

$$\hat{\mathbf{V}}_{LR} = \left(\mathbf{C}'\hat{\Delta}_{LR}^{-1}\mathbf{C}\right)^{-1}\left(\mathbf{C}'\hat{\Delta}_{LR}^{-1}\mathbf{D}(\mathbf{w})\hat{\mathbf{V}}\mathbf{D}(\mathbf{w})\hat{\Delta}_{LR}^{-1}\mathbf{C}\right) \tag{4.44}$$

where $\hat{\Delta}_{LR}^{-1} = \mathrm{diag}\left(w_1\hat{f}_1\left(1-\hat{f}_1\right),\ldots,w_I\hat{f}_I\left(1-\hat{f}_I\right)\right)$, $\hat{\mathbf{V}}$ is the covariance matrix estimate of **p** and **D**(**w**) = diag(**w**).

To take account of the survey design, the statistics (4.42) and (4.43) are adjusted based on the generalized deffs. The mean generalized deff is

$$\hat{\delta}_{\cdot LR} = \sum_{i=1}^{I}\left\{\frac{\mathrm{diag}\left(\hat{\mathbf{V}}_{res}\right)w_i}{\hat{f}_i\left(1-\hat{f}_i\right)}\right\}\Bigg/\left(I-p\right) \tag{4.45}$$

where $\hat{\mathbf{V}}_{res} = \mathbf{A}\hat{\mathbf{V}}\mathbf{A}'$ is the residual covariance matrix estimator of $p_i - \hat{f}_i$, and $\mathbf{A} = \mathbf{I} - \mathbf{D}(\mathbf{w})^{-1}\hat{\Delta}_{LR}\mathbf{X}\left(\mathbf{X}'\hat{\Delta}_{LR}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}(\mathbf{w})$. (**I** is the I×I identity matrix.) The residual covariance matrix is desirable because it is used to calculate standardised residuals which are then used to detect outlying cell proportions (Roberts et al., 1987; Lehtonen and Pahkinen, 1995). The adjusted statistics are $X^2_{LR}\left(\hat{\delta}.\right) = X^2_{LR} \big/ \hat{\delta}._{LR}$ and $G^2_{LR}\left(\hat{\delta}.\right) = G^2_{LR} \big/ \hat{\delta}._{LR}$ for the chi-square and likelihood tests respectively. These have an approximate $\chi_{I-p}{}^2$ distribution.

A second order adjustment is better if there is a large variation of the eigenvalues. $X^2_{LR}\left(\hat{\delta}., \hat{a}\right) = X^2_{LR}\left(\hat{\delta}.\right) \big/ \left(1+\hat{a}^2\right)$ and $G^2_{LR}\left(\hat{\delta}., \hat{a}\right) = G^2_{LR}\left(\hat{\delta}.\right) \big/ \left(1+\hat{a}^2\right)$ are treated as $\chi_{I-p}{}^2$ variables with $\left(I-p\right)/\left(1+\hat{a}^2\right)$ df. Note that $\hat{a}^2 = \left\{ \sum_{i=1}^{I-p} \frac{\hat{\delta}^2_{i(LR)}}{\left(I-p\right)\hat{\delta}.^2_{LR}} \right\} - 1$ and $\sum \hat{\delta}^2_{i(LR)} = \sum_i^I \sum_j^I \frac{\hat{\mathbf{V}}_{res}\left(nw_i\right)\left(nw_j\right)}{\hat{f}_i \hat{f}_j \left(1-\hat{f}_i\right)\left(1-\hat{f}_j\right)}$.

Roberts et al., (1987) also gives a Wald test statistic for complex survey data,

$$X^2_{W,LR} = \hat{\mathbf{v}}'\mathbf{C}\left(\mathbf{C}'\hat{\Delta}^{-1}_{LR}\mathbf{D}(\mathbf{w})\hat{\mathbf{V}}\mathbf{D}(\mathbf{w})\hat{\Delta}^{-1}_{LR}\mathbf{C}\right)\mathbf{C}'\hat{\mathbf{v}} \tag{4.46}$$

where $\hat{\mathbf{v}}$ is the vector of logits.

4.6 SUMMARY

Chi-squared tests commonly used on IID sample survey data yield results that increase the type I error when the survey is not a simple random sample design. For categorical data involving a complex design, there are adjustments that can be made on the hypotheses tests of goodness of fit, homogeneity and independence. Generally first order adjustments are used if the full covariance matrix is unavailable. This involves dividing the chi-square statistic or the likelihood ratio statistic by the average cell deff or the mean generalized deff.

If the full covariance matrix is known, then a Wald test statistic or a second order correction is better. However, the Wald test is not as stable when there are a small number of clusters. This leads to an F-correction of the Wald statistic. The F-correction can also be applied to first order adjusted test statistics.

The log-linear and logit modelling required ‘pseudo MLE’ procedures to estimated the unknown coefficients. This is due to difficulties in finding maximum likelihood functions of these models. To adjust for intracluster correlations, a weight was used in the pseudo maximum likelihood estimation for the logistic model. First and second order adjustments as well as Wald test statistics are available for these generalized linear models.

Although the test statistics discussed are design based, Lehtonen and Pahkinen (1995) point out that if the finite population is large, there is little difference between the cell proportions in a finite population or an infinite, superpopulation. Thus the results given here can be used for model based inferences also.

For examples, see section 6.2.1.

CHAPTER FIVE

MULTIVARIATE ANALYSIS

Often data exists in a multivariate form. Many observations are taken on a number of variables for each individual which are then compared simultaneously. Simple analyses such as z-tests and ANOVAs only deal with one variable at a time. With multivariate methods, the data sets can become quite large and since the usage of computers has advanced, it is much easier now to analyse many variables at one time.

The data is usually arranged in an n×p matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}. \qquad (5.1)$$

This matrix **X** holds all of the information of the p variables on the n individuals. Descriptive statistics are also calculated for multivariate data (Everitt and Dunn, 1992). The mean is a vector of p means, $\mu = (\mu_1, \mu_2, \ldots, \mu_p)^T$ and the estimate of the means is given by the sample means, $\bar{\mathbf{x}} = \left(x_1, x_2, \ldots, x_p\right)^T$. The variance and covariance between the ith and jth variables (i, j = 1,...,p) is arranged in a covariance matrix,

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \tag{5.2}$$

where σ_{ij} = Cov(x_i,x_j) = E($x_i x_j$) - $\mu_i \mu_j$. Similarly, the covariance matrix is estimated by the sample covariance matrix

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)' . \tag{5.3}$$

Another commonly used descriptive statistic is the correlation matrix,

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}, \tag{5.4}$$

where $\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$. The sample correlation coefficient is calculated by

$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$.

The matrix can get very large and methods have been used to summarise the information in the data matrix to something of a smaller dimension. As mentioned earlier, eigenvalues and eigenvectors serve the purpose of reducing the matrix down to scalars (eigenvalues) and p×1 eigenvectors. To find the eigenvalues and eigenvectors, the following equation must be satisfied

$$\mathbf{Sa} = \lambda\mathbf{a} \tag{5.5}$$

$$\text{i.e., } (\mathbf{S} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0} \tag{5.6}$$

where **S** is the p×p correlation or covariance matrix and **I** is the identity matrix. Note that the λ's are the eigenvalues and the corresponding **a**'s are the eigenvectors. However, equation (5.6) is satisfied when **a** = **0**. For a not so obvious solution, the matrix must be of full rank (complete independence of the rows and columns) and this would satisfy

$$|\mathbf{S} - \lambda\mathbf{I}| = \mathbf{0}. \tag{5.7}$$

Some useful properties of the eigenvalues are

$$\lambda_i \geq 0 \quad \text{for } i = 1,\ldots,p,$$

$$\begin{aligned} &\sum_{i=1}^{p} \lambda_i = \text{trace}(\mathbf{S}) \\ &\text{and} \prod_{i=1}^{p} \lambda_i = |\mathbf{S}|\,. \end{aligned} \tag{5.8}$$

Note that the trace (also denoted by tr) of a matrix is the sum of it's diagonal elements.

5.1 THE MULTIVARIATE NORMAL DISTRIBUTION

It is generally assumed that the variables of a multivariate problem are distributed normally and so follow a multivariate normal distribution. The p variables are also assumed independent and have a joint probability density function (pdf),

$$\text{f}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)'\,\Sigma^{-1}(\mathbf{x}-\mu)\right\}. \tag{5.9}$$

This distribution has mean μ and covariance matrix Σ.

An important property of the multivariate normal distribution is that, if

$$\text{Y} = \mathbf{a}'\mathbf{x}\,, \tag{5.10}$$

a linear combination of a multivariate normal variable, then Y also follows a normal distribution,

$$Y \sim N(\mathbf{a}'\mu, \mathbf{a}'\Sigma\mathbf{a}). \tag{5.11}$$

5.1.1 LIKELIHOOD FUNCTION

The likelihood function for a multivariate normal distribution is given by

$$L(\mathbf{x};\mu\Sigma) = \prod_{i=1}^{n} f(\mathbf{x}_i;\mu\Sigma)$$

$$= \prod_{i=1}^{n} (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu)' \Sigma^{-1}(\mathbf{x}_i - \mu)\right\}$$

$$= (2\pi)^{-np/2}|\Sigma|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)' \Sigma^{-1}(\mathbf{x}_i - \mu)\right\} \tag{5.12}$$

and taking logs

$$\ln L(\mathbf{x};\mu\Sigma) = -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)' \Sigma^{-1}(\mathbf{x}_i - \mu),$$

$$\text{but } (\mathbf{x}_i - \mu)' \Sigma^{-1}(\mathbf{x}_i - \mu) = (\mathbf{x}_i - \overline{\mathbf{x}} + \overline{\mathbf{x}} - \mu)' \Sigma^{-1}(\mathbf{x}_i - \overline{\mathbf{x}} + \overline{\mathbf{x}} - \mu)$$

$$= (\mathbf{x}_i - \overline{\mathbf{x}})' \Sigma^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) + (\overline{\mathbf{x}} - \mu)' \Sigma^{-1}(\overline{\mathbf{x}} - \mu) + 2(\overline{\mathbf{x}} - \mu)' \Sigma^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}).$$

$$\text{Note that } \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}}) = 0$$

$$\text{and } \sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})' \Sigma^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) = \sum_{i=1}^{n}\text{trace}(\mathbf{x}_i - \overline{\mathbf{x}})' \Sigma^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})$$

$$= \sum_{i=1}^{n}\text{tr}\Sigma^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$$

$$= \text{tr}\Sigma^{-1}\left(\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'\right).$$

So the log likelihood function is

$$\ln L(\mathbf{x};\mu\Sigma) = -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}\left\{ tr\Sigma^{-1}\left[\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' + n(\overline{\mathbf{x}} - \mu)(\overline{\mathbf{x}} - \mu)'\right]\right\}.$$

(5.13)

5.1.2 MAXIMUM LIKELIHOOD ESTIMATORS OF μ AND Σ

The maximum likelihood estimator for μ is found by differentiating the log likelihood function (5.12) with respect to μ and equating it to zero. The only term that involves μ is $\frac{n}{2} tr\Sigma^{-1}(\overline{\mathbf{x}} - \mu)(\overline{\mathbf{x}} - \mu)'$ and solving $\frac{\partial \ln L}{\partial \mu} = 0$ gives $n\Sigma^{-1}(\overline{\mathbf{x}} - \mu) = 0$ which is at a maximum when $\mu = \overline{\mathbf{x}}$. The maximum likelihood estimator of μ is therefore the sample mean, $\overline{\mathbf{x}}$.

Similarly for the covariance matrix, the maximum likelihood estimator is found by differentiating the log likelihood with respect to Σ. This is found to be $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$.

5.2 METHODS OF MULTIVARIATE ANALYSIS

5.2.1 PRINCIPAL COMPONENTS

This technique is one of the most common of multivariate data analysis. The idea behind principal components is to reduce the number of variables so that the data would be more manageable. To illustrate, a multivariate problem may involve, 50 variables that are measured on some 1000 individuals. The data matrix will appear as a 1000×50 matrix and trying to get any useful information out of such a large matrix will prove quite a difficult task.

Principal component analysis attempts to create linear combinations of the variables such that the original data set can be described in just a few new variables. New variables would be of the form

$$\mathrm{Y} = \mathbf{a}'\mathbf{x} \tag{5.14}$$

where Y is the new variable,

a is the (p×1) vector of weightings transposed

and **x** is the (p×1) vector of original p variables.

The weightings, **a**, are chosen so that the new variable, has maximum variance of the original variables. This Y is the first principal component, and denoted by Y_1. The following principal components (Y_2, Y_3, etc.) are also linear combinations, each accounting for the next greatest amount of variation after the previous principal component.

Derivation of Principal Components

The first principal component, $Y_1 = \mathbf{a}^T\mathbf{x}$, is formed when the variance is at a maximum, that is when $\mathrm{Var}(Y_1) = \mathbf{a}^T\mathbf{S}\mathbf{a}$ is at a maximum. Here **S** is the p×p sample covariance matrix. Obviously without restricting **a**, the variance could keep increasing. Therefore a restriction is placed on the weightings, such that $\mathbf{a}^T\mathbf{a} = 1$. Due to the restriction on the weightings, it can then be shown that the maximum variance is equivalent to the first eigenvalue and the weightings are the eigenvectors associated with the first eigenvalue. Ganesalingam (1993) has outlined this proof.

Define a new function,

$$f = \mathbf{a}'\mathbf{S}\mathbf{a} - \lambda(\mathbf{a}'\mathbf{a} - 1) \qquad (5.15)$$

where λ is Lagrangian Multiplier

then

$$\frac{\partial f}{\partial \mathbf{a}} = 2\mathbf{S}\mathbf{a} - 2\lambda\mathbf{a}\,, \qquad (5.16)$$

to maximise, set differential to zero and solve.

$$\begin{aligned} 2\mathbf{S}\mathbf{a} - 2\lambda\mathbf{a} &= 0 \\ (\mathbf{S} - \lambda\mathbf{I})\mathbf{a} &= 0 \end{aligned} \qquad (5.17)$$

which is exactly the form for solving eigenvalues and eigenvectors. Solving equation (5.17) will give eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$. Putting each eigenvalue back into the equation $(\mathbf{S} - \lambda_i\mathbf{I})\mathbf{a}_i = \mathbf{0}$ and solving for $\mathbf{a}_i$, this will produce corresponding p eigenvectors. Note that

$$(\mathbf{S} - \lambda_i\mathbf{I})\mathbf{a}_i = \mathbf{0}$$

$$\Leftrightarrow \mathbf{S}\mathbf{a}_i = \lambda_i\mathbf{a}_i$$

$$\Leftrightarrow \mathbf{a}_i{}'\mathbf{S}\mathbf{a}_i = \lambda_i \mathbf{a}_i \mathbf{a}_i{}'$$

$$\Leftrightarrow \mathbf{a}_i{}'\mathbf{S}\mathbf{a}_i = \lambda_i$$

so λ_i are the variances and the $\mathbf{a}_i$'s are the weightings.

The total variance of the p principal components will then be equal to the trace of **S**, since we know that $\sum_{i=1}^{p}\lambda_i = \mathrm{tr}(\mathbf{S})$ and the first j principal components will have a variance of $\sum_{i=1}^{j}\lambda_i$. To find the proportion of the variance of each principal component, $\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i}$.

Graphical Representation of Principal Components

For simplicity, consider the two variables from the 1996 survey, mean individual offences and age. The scatter plot is given in figure 5.1.

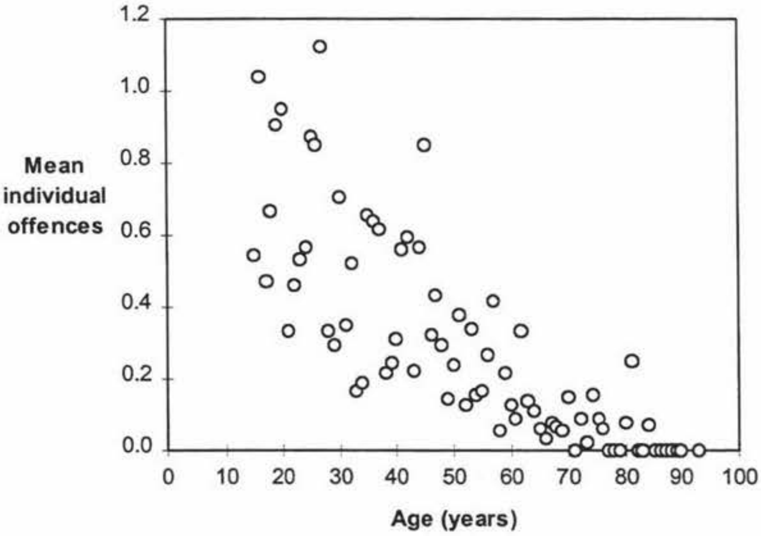

Figure 5.1. Plot of mean incidence of individual offences and age.

The first principal component (which in this case has been adjusted for the survey design), Y_1 is a linear combination that has maximum variance and is represented as in figure 5.2. Note that the line in the plot is not equivalent to the regression line of figure 3.2.

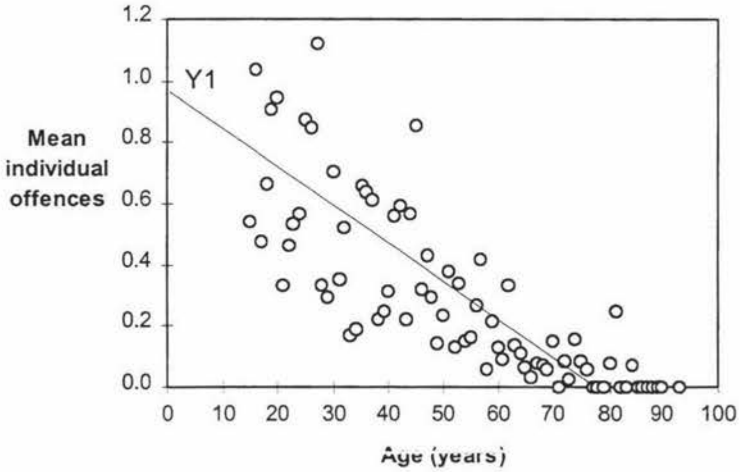

Figure 5.2. Plot showing the first Principal Component, Y_1.

The second principal component, Y_2, is in the direction of maximum variance after Y_1. See figure 5.3. Note that Y_1 and Y_2 are also orthogonal.

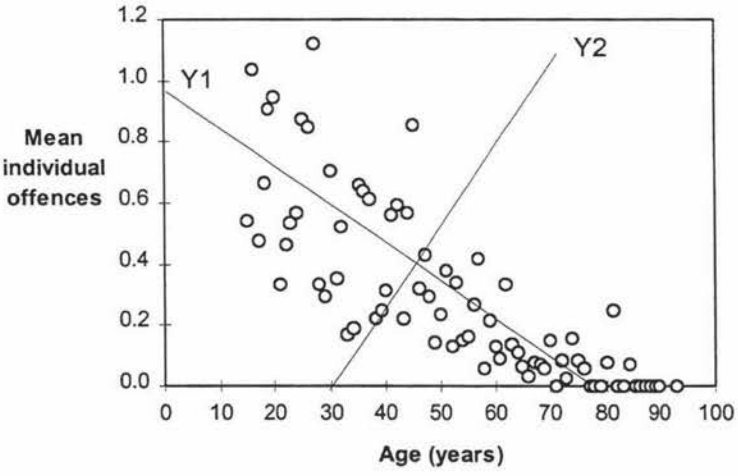

Figure 5.3. Plot showing both Principal Components, Y_1 and Y_2.

5.2.2 FACTOR ANALYSIS

Another common multivariate technique that makes use of the covariance matrix is factor analysis. The main aim of factor analysis is to describe the covariance matrix in some linear relationship of a few unobservable factors. This method is very similar to that of principal components except it uses observable and some underlying 'unobservable' factors (Johnson and Wichern, 1992). An example of an unobservable factor could be intelligence and other variables such as test scores on Mathematics, English, Music etc. could suggest some sort of an intelligence factor.

Relationships between the observed variables can be 'explained' by the unobserved variables. Usually the number of unobserved variables are less than the observed variables, i.e., if m is the number of unobserved variables and p is the total observed variables, then usually m<p.

The model for the orthogonal factor model is given as

$$\begin{aligned} x_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1 \\ x_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2 \\ &\vdots \\ x_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p \end{aligned}$$

and in matrix notation

$$\mathbf{x} - \mu = \mathbf{LF} + \varepsilon \tag{5.18}$$

where $\mathbf{x}$ has mean μ and covariance matrix Σ. The F_i are the unobservable factors, l_{ij} are the factor loadings (the square of the factor loading is the proportion of variance explained by that factor) and ε_i are the random errors associated with the model.

The assumptions of orthogonal factor analysis are

$$\begin{aligned} &\mathrm{E}(\mathbf{F}) = 0, \\ &\mathrm{cov}(\mathbf{F}) = \mathrm{I}, \\ &\mathrm{E}(\varepsilon) = 0, \end{aligned}$$

$$\mathrm{cov}(\varepsilon) = \mathrm{E}(\varepsilon\varepsilon') = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \psi_p \end{bmatrix}$$

$$\text{and } \mathrm{cov}(\varepsilon, \mathbf{F}) = \mathrm{E}(\varepsilon\mathbf{F}') = 0$$

(assuming independence between the unobservable factors and the errors). The covariance matrix, Σ, of the observable variables is calculated by

$$\begin{aligned} \mathrm{cov}(\mathbf{x}) &= \mathrm{E}(\mathbf{x} - \mu)(\mathbf{x} - \mu)' \\ &= \mathrm{E}\left[(\mathbf{LF} + \varepsilon)(\mathbf{LF} + \varepsilon)'\right] \\ &= \mathrm{E}\left(\mathbf{LF}(\mathbf{LF})' + \varepsilon(\mathbf{LF})' + \mathbf{LF}\varepsilon + \varepsilon\varepsilon'\right) \\ &= \mathbf{L}\mathrm{E}(\mathbf{FF}')\mathbf{L}' + \mathrm{E}(\varepsilon\mathbf{F}')\mathbf{L}' + \mathbf{L}\mathrm{E}(\mathbf{F}\varepsilon') + \mathrm{E}(\varepsilon\varepsilon') \\ &= \mathbf{LL}' + \mathrm{diag}\left(\psi_1, \psi_2, \ldots, \psi_p\right). \end{aligned}$$

Hence we get

$$\begin{aligned} &\mathrm{var}(\mathbf{x}_i) = l_{i1}^2 + \ldots + l_{im}^2 + \psi_i \\ &\text{and } \mathrm{cov}(\mathbf{x}_i, \mathbf{x}_k) = l_{i1}l_{k1} + \ldots + l_{im}l_{km}. \end{aligned} \tag{5.19}$$

The squared factor loadings are also known as the communality of x_i, i.e., the variance that is related to the common factors. The variance of the errors (variance not related to the common factors) is also called the specificity of x_i. Factor analysis is most useful when the number of unobservable variables are small compared to the number of observable variables. The data can then be summarised as m factor loadings rather than the original p (p>m) variables.

5.3 MULTIVARIATE ANALYSIS ON COMPLEX SURVEYS

As with other statistical methods, application of multivariate analysis assumes IID samples which are derived from simple random sampling. Estimation procedures based on simple random sample data are not appropriate if the design is a complex one. It has been found by many that surveys involving a complex sampling design affects the variance estimation (e.g., Kish and Frankel, 1974; Nathan and Holt, 1980). In multivariate analysis, procedures such as principal components and factor analysis are primarily concerned with the estimation of the covariance and/or the correlation matrix. Not surprising then, in the sample design such as stratification and clustering, the estimation of the covariance or correlation matrix is no longer unbiased.

Bebbington and Smith (1977) considered the problem of estimating the population correlation coefficient of a finite population with some grouping. It was found that in general, a bias existed for correlations. A further investigation was carried out and estimates of the correlation matrix for the finite population was calculated. Four sampling designs were chosen and samples selected. The clustered sampling designs were found to produce biased correlations, especially if the clustered design had unequal probability of selection (for details of the study, see Bebbington and Smith, 1977).

Consider a finite population, U, which contains N units. With each ith unit ($i = 1,...,N$) is a survey variable y_i (variable of interest) and a design variable z_i (assumed known prior to selection). Assume that the (y_i, z_i) are now IID realisations of a superpopulation, **(Y, Z)** and that **(Y, Z)** are a random sample from a multivariate normal distribution. Assuming a joint distribution, **Y** and **Z** are multivariate normal with parameters

$$\mu = \begin{bmatrix} \mu_y \\ \mu_z \end{bmatrix},\ \Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yz} \\ & \Sigma_{zz} \end{bmatrix}. \tag{5.20}$$

A sample of n units is drawn from U and will have a probability of selection p(s|**z**). The sample inclusion probability (probability that the ith unit is selected into the sample) for the ith unit is π_i and the weight of the ith unit is $w_i = (N\pi_i)^{-1}$.

As early as 1903, Pearson (cited in Smith, 1989) considered effects of selection on populations. Pearson showed that the (before and after) parameters are related by

$$\mu = \begin{bmatrix} \mu_y \\ \mu_z \end{bmatrix} = \begin{bmatrix} \mu_y * + \Sigma_{yz} * \Sigma_{zz} *^{-1} (\mu_z - \mu_z *) \\ \mu_z \end{bmatrix},$$

$$\Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yz} \\ & \Sigma_{zz} \end{bmatrix}$$

$$= \begin{bmatrix} \Sigma_{yy} * + \Sigma_{yz} * \Sigma_{zz} *^{-1} (\Sigma_{zz} - \Sigma_{zz} *) \Sigma_{zz} *^{-1} \Sigma_{zy} * & \Sigma_{yz} * \Sigma_{zz} *^{-1} \Sigma_{zz} \\ & \Sigma_{zz} \end{bmatrix} \tag{5.21}$$

where * estimates are the 'after selection' parameters i.e., conditional on z. Note also that $\Sigma_{yz} * \Sigma_{zz} *^{-1} = \beta_{yz}$. To adjust for selection bias,

$$\mu_y = \mu_y * + \beta_{yz}(\mu_z - \mu_z *),\ \Sigma_{yy} = \Sigma_{yy} * + \beta_{yz}(\Sigma_{zz} - \Sigma_{zz} *)\beta_{yz}'. \tag{5.22}$$

This is also known as the Pearson adjustment (Smith, 1989). The estimators of (5.22) are given respectively by

$$\hat{\mu}_y = \overline{\mathbf{y}}_s + \mathbf{b}_{yz}(\overline{\mathbf{z}} - \overline{\mathbf{z}}_s) \text{ and } \hat{\Sigma}_{yy} = \mathbf{V}_{yy} + \mathbf{b}_{yz}(\hat{\Sigma}_{zz} - \mathbf{V}_{zz})\mathbf{b}_{yz}'. \tag{5.23}$$

However, the Pearson adjusted estimators are assuming linearity of the regression of **y** on **z** and homoscedasticity of the residuals. Smith and Holmes (1989) have found that these estimators are not robust when the assumptions of linearity and homoscedasticity are not met.

Skinner, Holmes and Smith (1986) discusses the effect of selection on several covariance matrix estimators and then extends to discuss the impact of complex designs on Principal Components. They (Skinner et al., 1986) considered two design based covariance estimators and one model based estimator. These are

1. The simple random sample estimator (design based)

$$\hat{\Sigma}_{yy,srs} = \mathbf{V}_{yy} \tag{5.24}$$

2. The probability weighted estimator (design based)

$$\hat{\Sigma}_{yy,\pi} = \sum_{s} w_i \mathbf{y}_i \mathbf{y}_i{}' - w(s)^{-1} \hat{\mu}_{y,\pi} \hat{\mu}_{y,\pi}{}' \tag{5.25}$$

3. The maximum likelihood estimator (model based)

$$\hat{\Sigma}_{yy,ML} = \mathbf{V}_{yy} + \mathbf{b}_{yz}\left(\hat{\Sigma}_{zz} - \mathbf{V}_{zz}\right)\mathbf{b}_{yz}{}' \tag{5.26}$$

where

$$\beta_{yz} = \Sigma_{yz}\Sigma_{zz}^{-1},$$

$$w(s) = \sum_{s} w_i ,$$

$$\hat{\mu}_{y,\pi} = \sum_{s} w_i \mathbf{y}_i ,$$

$$\mathbf{x}_i = \left(\mathbf{y}_i{}^{T}, \mathbf{z}_i{}^{T}\right)^{T},$$

$$\bar{\mathbf{x}}_s = \frac{\sum_s \mathbf{x}_i}{n} = \left(\bar{\mathbf{y}}_s{}^T,\ \bar{\mathbf{z}}_s{}^T\right)^T,$$

$$\bar{\mathbf{x}} = \frac{\sum_N \mathbf{x}_i}{N} = \left(\bar{\mathbf{y}}^T,\ \bar{\mathbf{z}}^T\right)^T,$$

$$\begin{bmatrix} \mathbf{V}_{yy} & \mathbf{V}_{yz} \\ \mathbf{V}_{zy} & \mathbf{V}_{zz} \end{bmatrix} = \mathbf{V}_{xx} = \frac{\sum_s (\mathbf{x}_i - \bar{\mathbf{x}}_s)(\mathbf{x}_i - \bar{\mathbf{x}}_s)^T}{(n-1)},$$

$$\mathbf{b}_{yz} = \mathbf{V}_{yz}\mathbf{V}_{zz}{}^{-1}$$

$$\text{and } \Sigma_{zz} = \frac{\sum_N (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T}{N}.$$

The first estimator (5.24) does not take into account the sample design. It merely estimates the covariance matrix as if the sample s was selected by simple random sampling. Estimator (5.25) does include the use of prior information, i.e., the design variance is not used in the estimation. The model based estimator is found by the maximum likelihood method, based on the superpopulation model under the assumption of a multivariate normal distribution. Note that this is equivalent to the covariance matrix estimator based on Pearson adjustments of equation (5.23).

A fourth estimator proposed initially by Nathan and Holt (1980) is given in Smith and Holmes (1989) which is a combination of (5.25) and (5.26). This estimator is both design and model based. It is given by

4. Probability weighted maximum likelihood estimator

$$\hat{\Sigma}_{yy,\pi ML} = \mathbf{V}_{yy}* + \mathbf{b}_{yz}*\left(\hat{\Sigma}_{zz} - \mathbf{V}_{zz}*\right)\mathbf{b}_{yz}*^T \qquad (5.27)$$

where

$$\begin{bmatrix} \mathbf{V}_{yy} * & \mathbf{V}_{yz} * \\ \mathbf{V}_{zy} * & \mathbf{V}_{zz} * \end{bmatrix} = \mathbf{V}_{xx} * = \sum_{s} w_i \mathbf{x}_i \mathbf{x}_i^{T} - \frac{\bar{\mathbf{x}}_s * \bar{\mathbf{x}}_s *^{T}}{w(s)},$$

$$\bar{\mathbf{x}}_s * = \sum_{s} w_i \mathbf{x}_i$$

$$\text{and } \mathbf{b}_{yz} * = \mathbf{V}_{yz} * \mathbf{V}_{zz} *^{-1}.$$

Since two of these estimators are design based, one model based and one design and model based, to be able to compare and assess these estimators, the superpopulation model is adopted. Their conditional expected values, under a superpopulation model is given in table 5.1.

Table 5.1. Conditional expectations of the covariance estimators with respect to the superpopulation model.

$$E_{\xi}\left(\hat{\Sigma}_{yy,srs} | s, \mathbf{z}\right) = \Sigma_{yy} + \beta_{yz}\left(\mathbf{V}_{zz} - \Sigma_{zz}\right)\beta_{yz}^{T}$$

$$E_{\xi}\left(\hat{\Sigma}_{yy,\pi} | s, \mathbf{z}\right) = \alpha_{w}\Sigma_{yy} + \beta_{yz}\left(\mathbf{V}_{zz} * - \alpha_{w}\Sigma_{zz}\right)\beta_{yz}^{T}$$

where $\alpha_{w} = w(s) - \sum_{s} w_i^2 / w(s)$

$$E_{\xi}\left(\hat{\Sigma}_{yy,ML} | s, \mathbf{z}\right) = \alpha\Sigma_{yy} + \beta_{yz}\left(\hat{\Sigma}_{zz} - \alpha\Sigma_{zz}\right)\beta_{yz}^{T}$$

where $\alpha = n^{-1}\left[n - q - 1 + tr\left(\hat{\Sigma}_{zz}\mathbf{V}_{zz}^{-1}\right)\right]$ and q = number of design variables

$$E_{\xi}\left(\hat{\Sigma}_{yy,\pi ML} | s, \mathbf{z}\right) = \alpha * \Sigma_{yy} + \beta_{yz}\left(\hat{\Sigma}_{zz} - \alpha * \Sigma_{zz}\right)\beta_{yz}^{T}$$

where $\alpha *$ is a weighted α.

Source: Smith and Holmes (1989)

A simulation study in Skinner et al. (1986)'s paper, based on the point estimators found that the simple random sample estimator is generally biased but more so if the sample design is not self weighting. The probability weighted estimator, although a better estimator than the simple random sample estimator given a complex design, was found to be model biased. That is, under the assumption of a superpopulation, the probability weighted estimator was biased. The better estimator of the three which Skinner et al. (1986) compared was the model based estimator. This comes as no surprise since the model based estimator assumes a multivariate normal distribution of the superpopulation. Smith and Holmes (1989) points out that this estimator is not necessarily the best. The maximum likelihood estimator relies heavily on the assumptions of linearity and homoscedasticity of the regression of **Y** on **Z** and when these assumptions are not met, this estimator is also found to be biased. This is demonstrated in a 'real data' study by Smith and Holmes (1989). The fourth estimator (5.27), the probability weighted model based estimator, appears to produce unbiased results for most survey designs (Smith and Holmes, 1989).

5.3.1 PRINCIPAL COMPONENTS

The most common multivariate technique is principal components. The eigenvalues and eigenvectors will be affected if the covariance matrix is biased. Skinner et al. (1986) uses the Taylor series expansion to illustrate the bias of the design based estimators.

Define

$$\rho_j = \mathrm{corr}\left(X_j, \mathbf{Z}\right),$$

$$\Delta\Sigma = \left(\mathbf{V}_{zz} - \Sigma_{zz}\right)/\Sigma_{zz},$$

$$\Delta\Sigma_{\pi} = \left(\hat{\Sigma}_{zz,\pi} - \Sigma_{zz}\right)/\Sigma_{zz},$$

then the expected values, conditional on s and **z** are given by

$$E_{\xi}\left(\hat{\lambda}_{j,srs} | s, \mathbf{z}\right) \cong \lambda_j + \mathbf{a}_j{}^{T} \beta_{yz} \left(\mathbf{V}_{zz} - \Sigma_{zz}\right) \beta_{yz}{}^{T} \mathbf{a}_j$$

$$\cong \lambda_j \left(1 + \rho_j^2 \Delta\Sigma\right) \tag{5.28}$$

and

$$E_{\xi}\left(\hat{\lambda}_{j,\pi} | s, \mathbf{z}\right) \cong \lambda_j + \mathbf{a}_j{}^{T} \beta_{yz} \left(\hat{\Sigma}_{zz,\pi} - \Sigma_{zz}\right) \beta_{yz}{}^{T} \mathbf{a}_j$$

$$\cong \lambda_j \left(1 + \rho_j^2 \Delta\Sigma_{\pi}\right) \tag{5.29}$$

Note that $\mathbf{a}_j{}^{T} \beta_{yz} = \mathrm{cov}\left(X_j, Z\right)/\Sigma_{zz}$, $\mathrm{var}\left(X_j\right) = \lambda_j$ and $\alpha_w = 1$.

The eigenvectors are also estimated by the Taylor series and their conditional expected values are

$$E_{\xi}\left(\hat{\mathbf{a}}_{j,srs} | s, \mathbf{z}\right) = \mathbf{a}_j + \rho_j \Delta\Sigma \sum_{k \neq j} \phi_{jk} \rho_k \mathbf{a}_k \tag{5.30}$$

$$E_{\xi}\left(\hat{\mathbf{a}}_{j,\pi} | s, \mathbf{z}\right) = \mathbf{a}_j + \rho_j \Delta\Sigma_{\pi} \sum_{k \neq j} \phi_{jk} \rho_k \mathbf{a}_k \tag{5.31}$$

where $\phi_{jk} = \left(\lambda_j \lambda_k\right)^{1/2} / \left(\lambda_j - \lambda_k\right)$ and $\alpha_w = 1$.

The model based estimator has a smaller conditional bias and the estimated eigenvalues and eigenvectors are approximately model unbiased (Skinner et al., 1986). I.e., $E_{\xi}\left(\hat{\lambda}_{j,ML} | s, \mathbf{z}\right) \cong \lambda_j$ and $E_{\xi}\left(\hat{\mathbf{a}}_{j,ML} | s, \mathbf{z}\right) \cong \hat{\mathbf{a}}_j$.

The simulation study by Skinner et al. (1986) demonstrates the bias of the eigenvalues and eigenvectors. In their paper, they found that unconditional on s and **z**, the simple random sample estimator of the first eigenvalue, $\lambda_{1,srs}$, was biased when the design is not self weighting. Both the model based and probability weighted estimators were approximately unbiased for all of the designs. These results are replicated in Smith and Holmes (1989). In addition the probability weighted, model based estimator also gave approximately unbiased results. They (Smith and Holmes, 1989) also found that for the second eigenvalue, λ_2, all estimators showed little bias due to the low correlation between the second principal component and the design variable. This would make sense since the design would have almost no effect on the principal component, hence no effect on the eigenvalue and eigenvector.

The conditional results of Skinner et al. (1986) and of Smith and Holmes (1989) show that both the eigenvalues of the simple random sample and probability weighted estimators are conditionally biased when $\Delta\Sigma_{zz} \neq 0$. The two model based estimators have only a small bias, $O_p\left(n^{-1}\right)$.

The eigenvectors associated with each eigenvalue are more complicated to determine their bias. Skinner et al. (1986) defines a normalised mean eigenvector, $\bar{\mathbf{a}}^N = \bar{\mathbf{a}}\left(\bar{\mathbf{a}}^T\bar{\mathbf{a}}\right)^{-1/2}$ where $\bar{\mathbf{a}} = \frac{1}{r}\sum_{j=1}^{r}\hat{\mathbf{a}}_j$ and $\hat{\mathbf{a}}_j$ is the estimate of the first eigenvector associated with the first eigenvalue for the jth sample (j = 1,...,r). The bias of $\hat{\mathbf{a}}_1$ is then defined by the Euclidean distance

$$d\left(\bar{\mathbf{a}}^N, \hat{\mathbf{a}}_1\right) = \left[\left(\bar{\mathbf{a}}^N - \hat{\mathbf{a}}_1\right)^T\left(\bar{\mathbf{a}}^N - \hat{\mathbf{a}}_1\right)\right]^{-\frac{1}{2}}. \tag{5.32}$$

It was found by Skinner et al. (1986) that the bias to the eigenvectors were much less of a problem than for the eigenvalues. Both unconditional and conditional results were similar. Again the simple random sample estimator resulted in the most severe bias for the non-self weighting, more extreme designs. The model based and probability weighted estimator discussed had a fairly small bias.

The fourth estimator, in Smith and Holmes (1989), also showed small biases. The second eigenvector which is associated with the second eigenvalue was found to be approximately unbiased for all four estimators. This follows from the small biases of the second eigenvalue.

The simulations (Skinner et al., 1986 and replicated in Smith and Holmes, 1989) showed that the simple random sample estimator will give biased estimates especially if the sample design is not self weighting. The probability weighted estimator is generally unbiased but can be when conditioned on s and **z**. Both the model based estimators showed small biases in the simulations. This is only to be expected as Smith and Holmes (1989) points out that the simulations are based on a multivariate normal population.

The question then arises, how useful are these model based estimators to 'real' data? Smith and Holmes (1989) drew samples from a 'real' population, namely the 1975 UK Family Expenditure Survey. The design variable in this population was total expenditure and the other (**Y**) variables included housing, food, clothing, transport, lighting/fuel/power and services. However, to meet the assumption of normality, logs were taken of all of the **Y** variables.

The results for the simple random sample estimator was, as before, biased and it is suggested that this estimator should not be used when dealing with complex data. Both of the probability weighted estimators resulted in smaller biases than for the model based estimator. The weighted model based estimator gave the smallest bias. This confirms that the probability weighted estimators are overall better than the model based estimator, especially when the assumptions of linearity and homoscedascity is not met. When dealing with real data, it is not often that the data is multivariate normally distributed. Even when normality is approximated (using transformations), the model based estimator still gave biased estimates. This shows that the model based estimators relies heavily on their assumptions and it is better to use the probability weighted estimators, or model/design estimators. Note that, in essence, the problem of principal components for complex data, reduces to the problem of estimating the covariance matrix for the variables of interest.

5.3.2 FACTOR ANALYSIS

Factor analysis is very similar to principal components. Both multivariate techniques aim to reduce the covariance matrix into a more manageable, comprehensible form. Thurstone (1945) considered how selection can affect factor analysis but Skinner (1986) and Fuller (1987) have considered estimating covariance matrices given survey data. Finding factors for complex survey data, as for principal components, reduces to the problem of estimating the covariance matrix.

Skinner (1986) used regression estimation to adjust for selection and to estimate covariance matrices for the factor analysis model. He points out that regression estimation 'may give consistent estimators of parameters...' (pg347) if the sample is not a simple random sample selected from the finite population. In his paper, a two stage estimation procedure of the covariance matrix was proposed. The first stage estimates the sample covariance matrix, for example using regression estimation. The second stage modified the estimated covariance matrix and this modified version is put into the factor analysis model. Skinner (1986) also suggests an alternative one stage maximum likelihood procedure which is more efficient over the two stage estimation but relies on more normal distributional assumptions. This method can be implemented in the computer package LISREL.

Fuller (1987) also suggests two approaches to the estimation of the covariance matrix. One approach is by generalized least squares using the sample covariance matrix to form an unbiased estimate of the covariance matrix. This method is good for large sample sizes but not so good as the number of variables p increases, since the dimension of the problem is p(p-1) and hence requires a lot of mathematical computation. An alternative approach which requires less computations is by maximising the Wishart likelihood function. This approach is based heavily on assumptions of normality. Fuller illustrates the need for alternative covariance estimators other than the IID estimator using a language study example with self weighting data.

5.4 SUMMARY

It is generally agreed that the covariance matrix for IID data (generally the option in statistical packages) will be biased when used on complex survey data. The bias is usually larger if the design is not self-weighting. Alternative estimators include a probability weighted estimator, a maximum likelihood estimator and a weighted maximum likelihood estimator. Under a multivariate normal distribution, the maximum likelihood estimators perform quite well as long as the model holds. The weighted estimators were found to be more robust in simulation studies than the unweighted estimators (Smith and Holmes, 1989). Smith (1984) suggests that methods not adjusted for the design should not be used with complex data.

CHAPTER SIX

6.1 COMPUTING

With the aid of computing packages, statisticians have had their workload cut down a lot. Now analysing hundreds of data points can take a fraction of the time. No longer are statisticians required to calculate the covariance matrices by hand. This could require many hours if the covariance matrix had a lot of data points. Packages such as SPSS, MINITAB, and SAS are but a few of the available statistical packages. These packages allow even the most inexperienced to produce output. However, the output produced may not necessarily be correct since if data is entered into a computer and some right keys punched, the computer will produce output. Whether or not this output is the right output is left up to the analyst. Here, people with little or no training in statistics may fail to pick up on the 'wrong' output and come to a wrong conclusion.

The data itself can produce inaccurate results. Standard statistical packages such as those mentioned above, generally assumes the data is from an IID distribution and that it comes from a simple random sample design. Given a population that has information regarding the structure, the estimates from a sample, which makes use of that population structure in its design is more precise than one that does not take account of the population structure (e.g., a stratified sample is generally more precise than a simple random sample). Also dependence in the sample among individuals (e.g., cluster samples) will cause the estimate to be less precise than a simple random sample. Statistical packages that allow for complex sampling procedures (e.g., stratification and clustering within a sample) have been developed to deal with analysis of complex survey data. PC CARP, SUPERCARP and SUDAAN are all computer packages that allow estimation of population parameters given a complex design.

In this study, the computing packages used to deal with complex survey data will be PC CARP and SUDAAN. The aim is to compare the standard errors and the conclusions of these computing packages with a standard statistical package, SAS.

6.1.1 SAS

SAS is a powerful statistical package with great data handling capabilities. It comprises of many software products but the core component is 'base SAS'. To become familiar with SAS, a working knowledge of base SAS is required. This knowledge can then be extended to the other SAS software products.

In this study, SAS is used in a windows environment. The basic structure of SAS contains a program editor window, a log window and an output window. Programs are written in the program editor window, submitted and the output of the analyses (if any) are produced in the output window. The log window displays the program which is submitted and messages from the SAS system about the program.

The programming language in SAS is C+. A SAS program is a set of statements of which there are two parts, a DATA and a PROC part. DATA statements are where the data is read into SAS (either directly or from a file) and any data manipulation procedures occur in this step. PROC statements are for analysing the data. The range of analyses in SAS are vast. From descriptive means, totals, standard errors to two way tables, regression, logistic and log-linear models and multivariate data.

6.1.2 PC CARP

As the name suggests, PC CARP is designed to be used on a PC. It does not require much memory (in today's standards) to run the program (450K of memory).

It was developed at Iowa State University as an extension of the SUPERCARP to include analyses for subpopulations and two way tables. PC CARP is a program specifically designed to analyse survey data. Within this program, options of analyses include population estimates, stratum estimates, subpopulation analyses, two way tables, regression and univariate analyses. Later an addition analysis was written for PC CARP to perform logistic regression.

The data must be entered in a unique format so that it can be 'read' by the program to be analysed. Since it is a package designed to deal with structure within a population, the data must be read starting with a stratum identification first, then a cluster identification and a weight. The data for the variables can then be entered. The stratum must be ordered and also the clusters within each stratum, these too must be ordered. The weights are usually the reciprocal of the selection probability. These weights can be equal and in such a case, they do not have to be entered. Data can either be read from the hard drive or disk or entered straight from the keyboard as the program is running.

PC CARP is, on the whole, relatively user friendly. It is designed such that there are two parts to the program. The first is called 'Problem Specification'. This part prompts the user to enter in the data set and variable names to the point of ready for analysis. Note that PC CARP will only run if there are no missing values in the data set. The second part of the program is called 'Analysis Specification' which prompts the user for a specific analysis that is to be performed on the data set. Within each of the two parts, there are several stages. At each stage, the screen displays a menu. This menu is self explanatory and requires the user to enter some information before moving to the next screen. If a mistake is entered, in some cases, it is possible to go back to the previous screen. Online help is available in each screen if the user is unsure of the information required. Output can be requested to be printed, displayed on screen and/or saved into a file. This output contains a summary of the information of the data entered and the output from the analyses run.

6.1.3 SUDAAN

This statistical package was developed from a need for software for analysing complex sample surveys. SUDAAN was developed at the Research Triangle Institute (RTI) in the 1970s and it has been updated since. Originally it evolved from STDERR, a package designed specifically to compute standard errors of complex surveys as none of the standard packages did this. STDERR was similar to the way in which the SAS program operates, writing programs and submitting them. Since the need for a multi procedure package increased, eventually the statistical package, SUDAAN came about.

SUDAAN is designed for complex survey data. Some of the analyses include descriptive statistics such as calculating means, totals, proportions, percentages, ratios, tabulating contingency tables etc. Analytic statistics include regression, logistic regression, log-linear models for contingency tables, fitting models to failure time data and testing hypotheses. The data must be numeric to be read into SUDAAN. It also has to be of a certain format. For each survey observation there is to be a record and each variable appearing on each record must be in the same position. SAS data files are also compatible in some versions of SUDAAN.

To an unfamiliar programmer, this package is not as user friendly as PC CARP. Like SAS, SUDAAN has PROC (procedure) steps, and some programming is involved to carry out the analyses. There is however, online help. 'Help' is typed at the prompt and it lists subjects for which help is required. It is a more powerful package than PC CARP as it can deal with multiple analyses by including a few PROC statements. In PC CARP, each analysis requires the user to go through the 'Analysis Specification' section. For different data sets, PC CARP requires specifying the problem through the numerous screens in the 'Problem Specification' part, but in SUDAAN, a different data set can be easily identified by a design statement.

6.1.4 OTHER PACKAGES

It has been shown in the previous chapters that standard testing procedures for complex survey data can lead to misleading conclusions. Since almost all sample surveys involve a large number of data points, statistical analyses are generally done on computers. The output from standard computing packages may then give inaccurate results but be presented as the correct ones by researchers not aware of the effect of complex survey data. Although the standard statistical packages that assume simple random sampling are still more popular (SAS and SPSS are much more capable at handling and manipulating large data sets this may be why researchers choose to use these statistical packages), there are an increasing number of packages which allow for a complex design. On top of SUPERCARP, PC CARP and SUDAAN, there is also a package by Westat called Blaise. This piece of software was designed specifically for use on survey data. For multi level modelling, MLn for WINDOWS includes analyses for repeated measures, logistic and log-linear models and multivariate data. Recently there have also been macros written in SAS for handling multiple response data.

6.2 APPLICATION TO 'REAL' DATA

Following are three examples of analyses on survey data from the 1986 Community Questionnaire and the 1996 New Zealand National Survey of Crime Victims to illustrate the statistical methods from the previous chapters and some possible consequences of using a standard statistical package on complex data. The first example uses PC CARP as the statistical package for dealing with complex surveys. The output from PC CARP is compared to the output from SAS. The last two examples compares the package SUDAAN with SAS. Note that in all three examples, the weighting option is used in SAS. This gives exactly the same parameter estimates as in PC CARP and SUDAAN, i.e., the fitted model is exactly the same. However, the conclusions are not always the same. All SAS and SUDAAN statistical programs can be found in the appendix.

6.2.1 EXAMPLE ONE: CONTINGENCY TABLES

A two way table was produced for gender and 'whether an individual reported a crime to the police in the last 12 months'. The survey used in this analysis is the 1986 Community Questionnaire, and the statistical packages used in this analysis are SAS and PC CARP. Table 6.1 show the weighted counts from both SAS and PC CARP.

Table 6.1. Two way table of Gender by 'Reported a crime to police'

	Reported a Crime		
Gender	No	Yes	Total
Male	321.89	112.19	434.07
Female	482.42	144	626.42
Total	804.305	256.187	1060.49

Both packages tested for independence of the rows (gender) and columns (reported a crime). The chi-square statistic by SAS is, $\chi_1^2 = 1.142$ with a p-value of 0.285. For PC CARP, $\chi_1^2 = 1.024$ with a p-value of 0.3116. Both tests indicate non significance which would lead to the same conclusion, that gender is independent of reporting a crime. However, the chi-square test in SAS is slightly inflated. The design effect (given in PC CARP), deff = 0.9123, indicates that the sampling design will have little effect on the analysis. This is evident from the analysis as the SAS and PC CARP output are similar.

A second, two way contingency table illustrates the effect of a complex design when using the wrong package. The design effect in this case is 0.4767 which indicates that the standard package, SAS, is more likely to reject the null hypothesis when the null hypothesis is true (i.e., increase type I error). Gender was cross classified with 'whether an individual had asked a police officer for directions in the last 12 months'. Table 6.2 show the weighted estimates of both statistical packages.

Table 6.2. Two way table of Gender by 'Asked police for directions'.

	Asked for directions		
Gender	No	Yes	Total
Male	398.95	35.122	434.07
Female	599.56	26.858	626.42
Total	998.511	61.9808	1060.49

Again the chi-square is testing for independence between the rows and columns. SAS resulted in a chi-square value of 6.742 (df = 1, p = 0.009), and PC CARP resulted in a chi-square value of 6.0455 (df = 1, p = 0.0139). Comparing the two chi-square values, they are similar and both significant at the 0.05 level. However, at a fixed 0.01 significance level SAS shows that the chi-square test is still significant but the output from PC CARP is not. Due to the inflated χ_1^2 in SAS, this will lead to a wrong conclusion at the p = 0.01 significance level. We would conclude that gender is not independent of asking for directions if the analysis is run in SAS and that gender is independent of asking for directions if the analysis is run in PC CARP.

6.2.2 EXAMPLE TWO: REGRESSION

The data used in this linear regression example is from the 1996 Crime Victims questionnaire. Although there weren't very many variables suited for a linear regression, this example is mainly for illustrative purposes. The dependent variable is the total number of household offences (hhoffi). The independent variables are household size (hhsize), type of household (d4), rent/own the home and if rented, who from (d8d9a-d8d9e), occupation (d11), what kind of neighbourhood (w1), degree of different problems in the neighbourhood (w6a1 to w6a7) and how often an individual goes out at night, after dark (w7). The dependent variable was regressed on the independent variables through both SAS and SUDAAN. The household weighting variable, (hhwgtpos) is also used in the analyses of both packages.

Linear regression is not very appropriate for predicting the total number of household offences as the linear model fitted to the data may give negative fitted values of the dependent variable. This is logically incorrect as it is impossible to have a negative number of household offences. To overcome this 'problem', a log transformation was performed on the dependent variable. Also 0.1 was added to all of these values so there would not be any log(0) calculations. Nevertheless, both hhoff and log(hhoff + 0.1) was modelled by a linear regression. The results were compared and the same variables were significant in both analyses, which was expected. Hence for ease of interpretation, the results of untransformed variable is presented in table 6.3. Note that only the significant variables (either in SAS, SUDAAN or both) are shown.

Table 6.3. Linear regression estimates from SAS and SUDAAN.

Independent Variable	Beta Estimate	Standard Error SAS	Standard Error SUDAAN	P-Value SAS	P-Value SUDAAN
HHSIZE	0.0532	0.0211	0.0211	0.0119*	0.0117*
d4					
Couple with children	-0.2344	0.0892	0.1326	0.0087**	0.0773
w6a5					
Very big Problem	0.55061	0.0980	0.1616	0.0001**	0.0007**
w6a6					
Very big Problem	0.3829	0.1545	0.3231	0.0133*	0.2362
w6a7					
Very big Problem	0.5287	0.1456	0.3300	0.0003**	0.1094
w7					
At least once a week	0.1761	0.0707	0.0428	0.0127*	0.0000**

* Significant at 0.05 level
** Significant at 0.01 level
Note: Significance levels for SAS are nominal only.

The results shown in table 6.3 indicate that SAS produced twice as many significant variables than SUDAAN. The standard errors from SAS are generally much smaller than those from SUDAAN. This leads to the problem of smaller p-values which causes variables to become significant when they really are not (e.g., d4 is significant in SAS but not in SUDAAN).

6.2.3 EXAMPLE THREE: LOGISTIC REGRESSION

For the logistic regression, the binary response variable (burgp) indicates whether the household had been burgled (coded 1) or not (coded 0). The independent variables are the same independent variables used in the regression example. All of the variables are taken to be binary except for household size which is taken to be continuous.

The results from the SAS and SUDAAN programs are in tables 6.4 and 6.5 respectively. Table 6.4 show that SAS have found almost all of the variables to be significant when modelling the incidence of burglary. Compare this with the 'correct' output (table 6.5), SUDAAN has found only the intercept, hhsize and w6a1 and w7 to be significant in predicting burglary. A positive coefficient means that each increase in the variable, the probability of being burgled increases, e.g., as household size increases, the chances of being burgled increases also.

SAS gave smaller standard errors than SUDAAN, causing more variables in SAS to appear significant when they are not. For example, hhsize with a coefficient of 0.1261 has a standard error of 0.00404 in SAS and a p-value of 0.0001. SUDAAN gives a much larger standard error of 0.06157, p-value of 0.041, just significant at the 5 percent level.

Table 6.4. Logistic regression estimates from SAS, modelling burglary.

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCPT	1	-3.9127	0.0729	2880.3921	0.0001**
HHSIZE	1	0.1261	0.00404	977.0219	0.0001**
D4_1	1	0.3747	0.0233	258.4044	0.0001**
D4_2	1	0.5674	0.0223	647.1819	0.0001**
D4_3	1	0.3408	0.0203	281.0279	0.0001**
D4_4	1	0.0222	0.0193	1.3317	0.2485
D4_5	1	-0.4516	0.0288	245.4556	0.0001**
D4_6	1	0.1654	0.0252	43.1329	0.0001**
D8D9A	1	0.5772	0.0637	82.2226	0.0001**
D8D9B	1	-0.0909	0.0852	1.1401	0.2856
D8D9C	1	0.4312	0.0652	43.7239	0.0001**
D8D9D	1	0.5004	0.0632	62.6730	0.0001**
D11_1	1	-0.4672	0.0272	295.2006	0.0001**
D11_2	1	-0.2288	0.0246	86.2808	0.0001**
D11_3	1	-0.7734	0.0257	904.7901	0.0001**
D11_4	1	-0.4392	0.0242	330.1297	0.0001**
D11_5	1	-0.6003	0.0263	520.1608	0.0001**
D11_6	1	-0.7937	0.0307	669.9391	0.0001**
D11_7	1	-0.3149	0.0258	148.7906	0.0001**
W1_1	1	-0.0700	0.0114	37.6633	0.0001**
W1_2	1	0.1132	0.0105	116.7343	0.0001**
W6A1_1	1	-0.2642	0.0191	191.1731	0.0001**
W6A1_2	1	-0.0708	0.0139	26.0077	0.0001**
W6A1_3	1	0.0360	0.0101	12.5971	0.0004**
W6A2_1	1	0.6010	0.0234	659.4512	0.0001**
W6A2_2	1	0.5387	0.0152	1256.6604	0.0001**
W6A2_3	1	0.3662	0.0105	1211.1597	0.0001**
W6A3_1	1	0.0410	0.0164	6.2110	0.0127*
W6A3_2	1	-0.0135	0.0132	1.0387	0.3081
W6A3_3	1	-0.1933	0.0109	316.3858	0.0001**
W6A4_1	1	0.2729	0.0144	359.8380	0.0001**
W6A4_2	1	-0.3441	0.0141	593.1860	0.0001**
W6A4_3	1	0.0839	0.0105	64.3327	0.0001**
W6A5_1	1	0.2595	0.0174	223.4313	0.0001**
W6A5_2	1	-0.0358	0.0146	6.0404	0.0140*
W6A5_3	1	-0.0425	0.0106	16.0567	0.0001**
W6A6_1	1	-0.2151	0.0257	70.2003	0.0001**
W6A6_2	1	0.2325	0.0181	165.3868	0.0001**
W6A6_3	1	-0.0294	0.0113	6.7379	0.0094**
W6A7_1	1	0.9033	0.0219	1694.1742	0.0001**
W6A7_2	1	0.6592	0.0162	1655.7383	0.0001**
W6A7_3	1	0.2838	0.0111	650.5322	0.0001**
W7_1	1	-0.6589	0.0258	651.8501	0.0001**
W7_2	1	0.2975	0.0163	334.0527	0.0001**
W7_3	1	0.1489	0.0208	51.1607	0.0001**
W7_4	1	-0.3791	0.0237	256.3066	0.0001**

* Significant at 0.05 level

** Significant at 0.01 level

Note: Significance levels for SAS are nominal only.

Table 6.5. Logistic regression estimates from SUDAAN, modelling burglary.

Independent Variables and Effects	Beta Coeff.	SE Beta	T-Test B=0	P-Value for test B=0
Intercept	-3.91272	1.24937	-3.13176	0.00178**
HHSIZE	0.12614	0.06157	2.04888	0.04069*
D4				
One person living alone	0.37473	0.33598	1.11534	0.26492
Solo parent with child/children	0.56735	0.35026	1.61982	0.10553
Couple without children	0.34080	0.29697	1.14757	0.25137
Couple with children	0.02223	0.30795	0.07219	0.94247
Extended family/Whanau	-0.45161	0.44627	-1.01196	0.31176
Flatmates	0.16543	0.40668	0.40677	0.68425
Family other combination	0.00000	0.00000	.	.
D8D9A				
1	0.57724	1.10206	0.52378	0.60052
2	0.00000	0.00000	.	.
D8D9B				
1	-0.09093	1.55646	-0.0584	0.95342
2	0.00000	0.00000	.	.
D8D9C				
1	0.43119	1.13518	0.37984	0.70413
2	0.00000	0.00000	.	.
D8D9D				
1	0.50044	1.09712	0.45613	0.64838
2	0.00000	0.00000	.	.
D11				
Level1	-0.46724	0.46050	-1.01464	0.31048
Level2	-0.22882	0.42275	-0.54126	0.58843
Level3	-0.77341	0.43477	-1.77892	0.07550
Level4	-0.43917	0.40779	-1.07697	0.28171
Level5	-0.60034	0.43687	-1.37418	0.16964
Level6	-0.79375	0.50670	-1.56649	0.11749
Level7	-0.31491	0.43528	-0.72346	0.46953
No main income earner	0.00000	0.00000	.	.
W1				
help each other	-0.07004	0.18452	-0.37958	0.70432
Go own way	0.11318	0.17033	0.66445	0.50653
Mixture	0.00000	0.00000	.	.
W6A1				
Very big problem	-0.26415	0.30603	-0.86317	0.38821
Fairly big problem	-0.07081	0.22206	-0.31886	0.74989
Not a very big problem	0.03597	0.16973	0.21192	0.83220
Not a problem at all	0.00000	0.00000	.	.
W6A2				
Very big problem	0.60097	0.40951	1.46752	0.14249
Fairly big problem	0.53866	0.24421	2.20573	0.02759*
Not a very big problem	0.36617	0.18155	2.01691	0.04392*
Not a problem at all	0.00000	0.00000	.	.

* Significant at 0.05 level
** Significant at 0.01 level

Table 6.5. Logistic regression estimates from SUDAAN, modelling burglary (continued).

Independent Variables and Effects	Beta Coeff.	SE Beta	T-Test B=0	P-Value for test B=0
W6A3				
Very big problem	0.04095	0.28143	0.14552	0.88432
Fairly big problem	-0.01348	0.22506	-0.05988	0.95226
Not a very big problem	-0.19327	0.19666	-0.98275	0.32593
Not a problem at all	0.00000	0.00000	.	.
W6A4				
Very big problem	0.27286	0.23077	1.18239	0.23728
Fairly big problem	-0.34414	0.23294	-1.47739	0.13983
Not a very big problem	0.08390	0.17378	0.48278	0.62934
Not a problem at all	0.00000	0.00000	.	.
W6A5				
Very big problem	0.25951	0.29145	0.89041	0.37342
Fairly big problem	-0.03578	0.22683	-0.15774	0.87469
Not a very big problem	-0.04253	0.17425	-0.24409	0.80720
Not a problem at all	0.00000	0.00000	.	.
W6A6				
Very big problem	-0.21511	0.39192	-0.54886	0.58320
Fairly big problem	0.23254	0.30085	0.77295	0.87469
Not a very big problem	-0.02945	0.18827	-0.15641	0.87573
Not a problem at all	0.00000	0.00000	.	.
W6A7				
Very big problem	0.90334	0.37035	2.43915	0.01486*
Fairly big problem	0.65918	0.29112	2.26429	0.02373*
Not a very big problem	0.28377	0.18939	1.49836	0.13430
Not a problem at all	0.00000	0.00000	.	.
W7				
Never	-0.65893	0.40790	-1.61542	0.10648
At least once a week	0.29749	0.26195	1.13568	0.25631
At least once a fortnight	0.14888	0.33804	0.44042	0.65971
At least once a month	-0.37914	0.38934	-0.97379	0.33035
Less often than once a month	0.00000	0.00000	.	.

* Significant at 0.05 level

** Significant at 0.01 level

A second logistic regression was run for modelling violence (totalp) on variables, gender (d5), age (age), ethnicity (ethnic), work situation (d10), occupation (d11), marital status (d13) and how often an individual goes out at night after dark (w7). The estimated parameters, standard errors and p-values are presented in tables 6.6 and 6.7.

Table 6.6. Logistic regression estimates from SAS, modelling violence.

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCPT	1	-1.7582	0.0587	895.8769	0.0001**
D5	1	0.0537	0.00443	146.7981	0.0001**
Age_1	1	0.4200	0.0339	153.8521	0.0001**
Age_2	1	0.00615	0.0336	0.0336	0.8546
Age_3	1	-0.5007	0.0337	221.1497	0.0001**
Age_4	1	-0.9596	0.0373	663.3537	0.0001**
Age_5	1	1.3345	0.0637	438.2501	0.0001**
Ethnic_1	1	-0.1127	0.0101	123.8911	0.0001**
Ethnic_2	1	0.1323	0.00886	222.8473	0.0001**
Ethnic_3	1	0.2481	0.0106	551.0432	0.0001**
Ethnic_4	1	0.2687	0.0120	505.0895	0.0001**
D11_1	1	-0.3085	0.0125	604.9007	0.0001**
D11_2	1	-0.2525	0.0111	519.6081	0.0001**
D11_3	1	-0.1438	0.0111	168.1167	0.0001**
D11_4	1	-0.1009	0.0105	92.5909	0.0001**
D11_5	1	-0.1124	0.0116	94.0686	0.0001**
D11_6	1	0.1661	0.0126	174.5194	0.0001**
D11_7	1	0.1084	0.0122	79.2571	0.0001**
D10_1	1	0.1544	0.0456	11.4630	0.0007**
D10_2	1	0.0857	0.0463	3.4240	0.0643
D10_3	1	-0.7532	0.0494	232.3421	0.0001**
D10_4	1	0.5659	0.0461	150.4697	0.0001**
D10_5	1	0.2203	0.0475	21.4871	0.0001**
D10_6	1	0.6570	0.0516	161.8403	0.0001**
D10_7	1	0.1358	0.0460	8.7036	0.0032**
D13_1	1	-0.8574	0.00842	10379.3893	0.0001**
D13_2	1	-0.3471	0.00977	1263.0317	0.0001**
D13_3	1	-0.3540	0.00934	1435.3487	0.0001**
D13_4	1	-1.2230	0.0229	2847.7592	0.0001**
W7_1	1	-0.2302	0.0148	242.5906	0.0001**
W7_2	1	0.1057	0.0101	110.0938	0.0001**
W7_3	1	-0.0790	0.0128	38.0974	0.0001**
W7_4	1	-0.2775	0.0136	418.5061	0.0001**

** Significant at 0.01 level
Note: Significance levels for SAS are nominal only.

Table 6.7. Logistic regression estimates from SUDAAN, modelling violence.

Independent Variables and Effects	Beta Coeff.	SE Beta	T-Test B=0	P-Value for test B=0
Intercept	-1.65077	1.42803	-1.15597	0.24790
D5				
Male	-0.05373	0.12852	-0.41803	0.67599
Female	0.00000	0.00000	.	.
AGE				
1	0.41998	0.93147	0.45088	0.65215
2	0.00615	0.93484	0.00658	0.99475
3	-0.50072	0.93967	-0.53287	0.59421
4	-0.95961	1.07329	-0.89408	0.37144
5	1.33454	1.34383	0.99309	0.32085
6	0.00000	0.00000	.	.
ETHNIC				
European	-0.11267	0.28164	-0.40005	0.68918
NZ European	0.13226	0.23919	0.55293	0.58041
NZ Maori	0.24806	0.25788	0.96190	0.33627
Pacific Islander	0.26869	0.41474	0.64784	0.51720
Other	0.00000	0.00000	.	.
D10				
Working in paid employment	0.15444	1.02633	0.15047	0.88041
Home duties	0.08574	1.03609	0.08276	0.93406
Retired/Supt	-0.75319	1.12358	-0.67035	0.50275
Benefit/Unemployed	0.56590	1.03199	0.54836	0.58354
Sick/Disabled and unable to work	0.22032	1.09267	0.20164	0.84023
Unpaid work outside the home	0.65695	1.20635	0.54458	0.58614
Student	0.13585	1.04818	0.12960	0.89690
Other	0.00000	0.00000	.	.
D11				
Level 1	-0.30845	0.44907	-0.68687	0.49228
Level 2	-0.25249	0.40535	-0.62289	0.53346
Level3	-0.14377	0.41489	-0.34651	0.72901
Level4	-0.10093	0.39763	-0.25382	0.79968
Level5	-0.11235	0.43069	-0.26087	0.79423
Level6	0.16607	0.42271	0.39286	0.69448
Level7	0.10843	0.44079	0.24600	0.80572
No main income earner	0.00000	0.00000	.	.
D13				
Legally married	-0.85743	0.19659	-4.36155	0.00001**
Defacto relationship	-0.34710	0.23580	-1.47197	0.14126
Single/Never married	-0.35403	0.21548	-1.64303	0.10061
Widowed	-1.22301	0.40351	-3.03095	0.00248**
Div/Sept	0.00000	0.00000	.	.
W7				
Never	-0.23020	0.36323	-0.63377	0.52634
At least once a week	0.10566	0.25725	0.41075	0.68132
At least once a fortnight	-0.07895	0.32937	-0.23971	0.81059
At least once a month	-0.27748	0.35522	-0.78113	0.43486
Less often than once a month	0.00000	0.00000	.	.

** Significant at 0.01 level

Looking at tables 6.6 and 6.7, note that in SAS (table 6.6), gender (d5) is coded as 1 for males and 2 for females whereas in SUDAAN (table 6.7), gender is coded 1 for males and 0 for females. This will give different intercepts and coefficients for gender but will lead to the same linear model, i.e., totalp = intercept + gender coefficient*(gender) + (the other variables). So for males, the linear equations are:

$$\begin{aligned} \text{totalp} &= -1.7582 + 0.0537*1 + (\text{the other variables}) \\ &= -1.7045 + (\text{the other variables}) \quad (\text{SAS}) \end{aligned} \tag{6.1}$$

$$\begin{aligned} \text{and}\quad \text{totalp} &= -1.65077 + (-0.05373)*1 + (\text{the other variables}) \\ &= -1.7045 + (\text{the other variables}) \quad (\text{SUDAAN}) \end{aligned} \tag{6.2}$$

which are equivalent. Similarly for females, equations (6.3) and (6.4) are equivalent:

$$\begin{aligned} \text{totalp} &= -1.7582 + (0.0537)*2 + (\text{the other variables}) \\ &= -1.6508 \quad (\text{SAS}) \end{aligned} \tag{6.3}$$

$$\begin{aligned} \text{and}\quad \text{totalp} &= -1.65077 + (-0.05373)*0 + (\text{the other variables}) \\ &= -1.65077 \quad (\text{SUDAAN}) \end{aligned} \tag{6.4}$$

This shows some discrepancies not related to complex surveys when using SAS and SUDAAN.

Apart from the slight differences in coding gender, the beta coefficients of all the other variables are equivalent. The problem with using a standard package such as SAS to analyse complex survey data is that it results in smaller (and incorrect) estimated standard errors than in a statistical package that allows for the complex design. This in turn inflates the chi-square statistic and causes the nominal p-values to be much smaller. If we were using only the SAS package for this analysis, we would conclude that all of the variables are significant except for age (group 2) and d10 (retired/supt) (see table 6.6). However, if we were using a package that adjusts for complex designs, such as we have used SUDAAN, none of the variables are significant except for marital status (legally married and widowed) (see table 6.7). These results from both statistical packages illustrate the consequences that can happen when using statistical packages which do not account for complex survey data and how serious the consequences can be.

6.2.4 GENERAL CONCLUSIONS

The first contingency table analysis from example one (Gender by 'Reported a crime') was found to be not significant at the 0.05 significance level by both of the statistical packages (SAS and PC CARP). That is, gender is independent of whether an individual reports a crime to the police. The second contingency table analysis tested the independence of gender and 'asked police for directions'. This was found to be significant in SAS at the nominal 0.01 significance level but was not significant in PC CARP at the same true significance level. However, both statistical packages showed significance at the 0.05 significance level. Hence at both the true and nominal 0.05 level, gender effects whether people ask the police for directions. At this level for this data and question, the results of either packages would not lead to a wrong conclusion.

The linear regression (example two) from SAS resulted in hhsize, d4 (couple with children), w6a5 (very big problem), w6a6 (very big problem), w6a7 (very big problem) and w7 (at least once a week) all being significant variables in predicting the number of household offences. Compared to the results from SUDAAN, only hhsize, w6a5 (very big problem) and w7 (at least once a week) are significant in predicting the number of household offences.

In example three, two logistic regression analyses were carried out. The first tried to predict the prevalence of burglary. SAS found that all of the variables except for d4_4, d8d9B and w6a3_2 were significant in the model for predicting the prevalence of burglary. This results from the complex design of the 1996 survey which included clustering. The results from SUDAAN shows only the intercept, hhsize, w6a2 and w6a7 to be significant variables for determining the prevalence of burglary. The second logistic regression tried to model the prevalence of violence. The conclusions from the SAS analyses (table 6.6) indicate that all of the variables are significant except for age_2 and d10_2. This is different to the conclusions produced from SUDAAN. The results from SUDAAN (table 6.7) show that d13 is only the important (significant) variable in predicting the prevalence of violence.

Generally, the results from SAS produce standard errors which are too small which lead to significance of variables when they really are not significant. Packages which take into account of the design of the survey are strongly recommended so that the correct conclusions are reached.

6.3 SUMMARY

Although the more popular statistical packages such as SAS, SPSS and MINITAB are taught at universities and other workplaces, they are insufficient for analysing complex data. Almost all sample surveys involve some complex design. Analyses on these surveys should be carried out using statistical techniques developed primarily for dealing with complex designs. This would ensure that the standard errors are correct and that the significance tests do not lead to a wrong conclusion.

Computers are very powerful machines, capable of handling a large amount of data with ease. Many researchers enter the data collected from their surveys into computers and use statistical packages to analyse the data. Most researchers are not aware that the standard statistical packages are designed for use on simple random sample data and not for complex design data. For simple calculations such as totals or means, packages that have a weighting option can be used on the survey data. For more complicated calculations such as estimating covariance matrices or modelling analyses, a specialised package is necessary for the correct estimates.

The examples presented in this chapter have included contingency table analyses, linear and logistic regression. Two specialised packages were used, PC CARP and SUDAAN. The results from these packages were compared to a popular, standard statistical package namely SAS. In the contingency tables example, when the deff was not close to one, it was found that SAS produced inflated chi-squared values. This produced significant results which lead to a rejection of the null hypothesis. The corresponding test in PC CARP was not significant at the same fixed level of significance (p = 0.01). Example two illustrated that the standard errors in SAS were much smaller than in SUDAAN, which adjusted for the design. There were also more variables which were significant in the regression equation when using SAS. The logistic regression examples also showed this trend. The standard errors were generally very small compared to the correct standard errors (calculated in SUDAAN). This led to the variables being significant (in both logistic regressions, almost all of the variables used were significant) when they were not.

It is important to use the correct statistical package, just as it is important to use the correct statistical techniques on complex data. If the complex survey data is not properly accounted for, the estimates may become grossly biased and the conclusions may become misleading.

CHAPTER SEVEN

SUMMARY AND CONCLUSIONS

Sampling is a very common method of gathering information about a certain population. What makes sampling so popular is that it is efficient, economical and takes less time. Of course the precision of the sample depends on the sampling design used and the sample size. There are many sampling designs and some may be more efficient than others depending on the population one wishes to sample from. The precision of a sampling design depends on how the elements are selected and also other possible errors such as non response, coding and data entry errors. This thesis is not concerned about non response errors or coding and data entry errors. The focus is primarily on statistical techniques and their estimates on complex sampling designs.

The simplest of all sampling designs is the simple random sample. Traditionally this formed the basis of all statistical techniques and all statistics-related computer software. Other sampling designs were introduced to increase the efficiency of sampling. That is, making use of auxiliary information to get a better estimated, or reducing the cost of the sampling for a loss of precision (e.g., cluster sampling). Most sample surveys are not simple random designs, mainly due to the availability of resources that can aid in the collection of a sample, for example, if only a list of street names or a list of urban/rural areas is available for designing the sample survey.

The sampling designs discussed in this thesis are not the only possible designs. See for example Cochran (1963). Stratified, cluster and multistage sampling are the most frequently used designs other than simple random sampling. Stratified sampling is very useful and estimates are generally more precise than those from a simple random sample. Cluster sampling, however, is very convenient and sometimes necessary but the estimates are generally less precise than for a simple random sample of the same size. This is due to the intracluster correlation which elements exhibit when cluster sampling. Multistage sampling is a design which is carried out in several stages. Each stage can be any sampling design. In practice, it is usually not ideal to simple random sample but to use a stratified, cluster or multistage design. Any design which is not a simple random sample design is known as a complex design.

Most researchers use the statistical techniques developed for simple random sample data, even if their sample was of a complex design. Since the statistical methods assume the observations are IID, this assumption is generally not met especially when cluster sampling has been used. Historically, concern grew over the usage of statistical techniques when the data was not IID. A design effect was introduced to compare the efficiency of a design against a simple random design.

The estimates obtained are quite often design based. Samples are selected from a finite population and descriptive statistics found which relate only to that particular population at that particular time. To be able to generalize to another population or another time, a theoretical model is required. This is also known as the model based approach, based on an infinite superpopulation. The model based approach offers flexibility in forming estimators depending on the assumptions imposed on that superpopulation.

In a design based approach, the sampling design is considered important as this dictates the selection of the elements. The model based approach, on the other hand, assumes that the sample is already given and the design is no longer considered as important. Under a superpopulation the design or auxiliary variables are incorporated into the proposed model. Hence these variables are not important at the design stage but as independent variables.

In chapter two, a regression model was proposed. In a purely design based approach, estimation of the coefficients is most commonly by the ordinary least squares method. Alternatives to the OLS are weighted least squares, either by the inverses of the respondent variances or by the inverses of the respondent inclusion probabilities. Model based approaches include the design variables into the regression model, a maximum likelihood estimation method and in the case of clustering among individuals, a diagonal block matrix of correlations can be used as weights.

For contingency tables, the results given are applicable to either design or model based approaches. Complex survey designs generally inflate the chi-square statistic. To compensate, first and second order adjustments are used. Alternatively, Wald tests (F-corrected or not) can also be used if the full covariance matrix is available. Log-linear and logistic regression models for complex surveys are more complicated. Generally the cell estimates are not easily derived and an iterative method is required. The first and second order adjustments use a more complex design matrix and tend to give more accurate results.

Since many multivariate techniques require the computation of the covariance or correlation matrix, both design based and model based estimates are suggested as alternatives for complex survey data. Principal components has been discussed widely throughout the literature related to complex survey analyses as a common multivariate method.

Statistical calculations have become easier and easier with the advancement of computer software. There now exist powerful statistical packages capable of handling hundreds of data points and performing statistical functions with ease. However, it is only in the last 20 years that software for complex designs has been developed. Specialised packages such as those used in this thesis (PC CARP and SUDAAN), are sufficient for analysing the data but still lack certain data management abilities of SAS or SPSS. Ideally, developing a standard package such as SAS, which is very popular among social scientists, to include complex analyses would be very useful. The computing examples on two actual sample surveys highlighted the fact that standard statistical procedures are not recommended for analysing complex survey data. This is especially evident in the logistic regression examples where the conclusions from SAS were very different to the results obtained from SUDAAN. Complex survey data requires software designed specifically with complex designs in mind.

People that carry out sample surveys are not always statisticians. Some of the statistical software are quite user friendly and people with no or little training in statistics can still be able to do a statistical analysis on a computer.

However, these people may not be aware of incorrect output and thus will come to wrong conclusions.

This thesis attempted to illustrate the problems under the standard assumptions for simple random data when dealing with complex designs. It also tried to make researchers aware that the assumptions of a simple random sample do not readily apply to complex designs.

APPENDIX

A.1 1986 WELLINGTON COMMUNITY QUESTIONNAIRE

1986 COMMUNITY QUESTIONNAIRE

MAIN QUESTIONNAIRE: PART 1

INTERVIEWER TO COMPLETE AT EACH ELIGIBLE ADDRESS	1 2 0 0 CARD NO.
(a) Questionnaire ID Number	4 5 6 7 8
(b) How many houses/flats are there at this address?	10 11
(c) The house/flat selected is: The only house at this address..... 1 One of a number of houses at this address................... 2 One of a number of flats at this address................... 3	13
(d) The selected dwelling: Occupies only 1 section............ 1 Appears to occupy more than 1 section.................... 2	15
(e) Street number	17 18 19 20 21 22 23 /
(f) Date	25 26 27 28 29 30 DATE
(g) Interviewer number	32 33 INTERVIEWER NO.

THIS QUESTIONNAIRE IS USED FIRST
WITH ALL RESPONDENTS

1 2
0 1
CARD NO.

Q1 First of all, could you tell me how long you have lived in ____________________? (WRITE IN NAME OF AREA e.g. WHITBY)

Less than 1 month	1	ASK Q2
1 but less than 3 months..	2	
3 but less than 6 months..	3	
6 but less than 12 months.	4	
1 but less than 2 years...	5	GO TO Q3
2 but less than 5 years...	6	
5 but less than 10 years..	7	
10 years or more	8	

4
□

Q2 Where was the last place you lived before that?

____

6 7
□ □

INTERVIEWER: CODE WHETHER RESPONDENT LIVED INSIDE OR OUTSIDE THE AREA (EITHER TAWA/PORIRUA BASIN OR PETONE/LOWER HUTT)

Inside 1
Outside.........0

9
□

Q3 How well do you know your neighbours?

READ OUT

Very well 1

Quite well 2

Some very well, others
not very well..............3

Not very well..............4

Not at all 5

No response 8

Don't know 9

11 □

Now I'd like to ask you a few questions about crime in the area.

Q4 Do you think that in the last 12 months there has been more or less crime in ____________ or has it stayed the same?

READ OUT

13 □

More 1

Less 2

Stayed the same 3

No crime around here..... 0

No response 8

Don't know 9

Q5 What do you think are the biggest crime problems in this area?

IF DON'T KNOW ASK: Well, what would you think?

(DO NOT PROMPT) CODE ALL THAT APPLY

Burglary, break-ins	1	
Vandalism	2	
Street attacks...........	3	
Petty thefts	4	
Assault	5	ASK
Domestic Violence	6	Q6
Sexual crimes	7	
Car theft	8	
Theft from cars	10	
Damage to cars	11	
Other (WRITE IN)		
______________________		
______________________		
______________________		
No crime around here	0	GO TO
Don't know	9	Q7

15 16

18 19

21 22

24 25

27 28

30 31

33 34

36 37

38 39

41 42

Q6 FOR EACH THAT APPLIES ASK:

How often do ______________ happen in this area - a lot, quite a lot, or not very much.

N.B. IF MORE THAN 5 ANSWERS GIVEN TO Q5, ASK FOR THE 5 WHICH OCCUR MOST OFTEN.
WRITE IN AND CODE.

	A lot	Quite a lot	Not very much	Don't know		
1. ____________________________	..1..	..2..	..3..	..9..	44 45 □□	47 □
2. ____________________________	..1..	..2..	..3..	..9..	49 50 □□	52 □
3. ____________________________	..1..	..2..	..3..	..9..	54 55 □□	57 □
4. ____________________________	..1..	..2..	..3..	..9..	59 60 □□	62 □
5. ____________________________	..1..	..2..	..3..	..9..	64 65 □□	67 □

ASK ALL

Q7 As I read out the following, can you tell me if they are very common, quite common, or not at all common in your area.

Firstly,... Damaged public telephones.. are they very common, quite common, or not at all common?

READ OUT EACH IN TURN AND CODE BELOW.

	Very common	Quite common	Not common	Not applicable	No response	Don't know	
a) Damaged public telephones	..1..	..2..	..3..	..7..	..8..	..9..	69 □
b) Rubbish and litter lying about on streets or empty sections	..1..	..2..	..3..	..7..	..8..	..9..	71 □
c) Broken windows in shops, public buildings etc	..1..	..2..	..3..	..7..	..8..	..9..	73 □
d) Graffiti on walls, schools, shops, churches etc.	..1..	..2..	..3..	..7..	..8..	..9..	75 □
e) Children hanging around the streets during school hours	..1..	..2..	..3..	..7..	..8..	..9..	77 □

Q8 Is there anywhere in this neighbourhood where you would be afraid to walk alone at night?

Yes..........	1 ASK Q9	
No..,........	2	79 □
No response..	9 GO TO Q11	
Don't know..	8	

1 2
0 2
CARD NO.

Q9 What are you afraid might happen?

(DO NOT PROMPT)

CODE ALL THAT APPLY
(use any pair of boxes)

Being attacked and seriously injured or perhaps killed..............01

Being raped....................02

Being robbed...................03

Being assaulted or threatened....................04

Being hassled by other people..................05

General fear something may happen....................06

Other (WRITE IN)

____

____

No response.................... 88

Don't know..................... 99

4 5

7 8

10 11

13 14

16 17

19 20

22 23

Q10 Why do you think that this/these things might happen to you?

CODE ALL THAT APPLY
(use any pair of boxes)

No-one else around............01

Too dark/no street lighting....02

Read about crime in the papers/see it on TV...........03

Hearsay (what other people say)...................04

People hanging about on the streets.................05

Other (WRITE IN)

____

____

No response................... 88

Don't know.................... 99

25 26

28 29

31 32

34 35

37 38

40 41

ASK ALL

Q11 People often worry about being the victim of a crime, and I would like you to tell me if you are worried about being the victim of different types of crime.

SHOWCARD A Using one of the phrases on this card, could you tell me how worried you are about.. having your home broken into....

READ OUT EACH CRIME IN TURN

	Very Worried	Fairly Worried	Not very Worried	Not at all Worried	Never Thought About it	No Response	Don't Know	
a) Having your home broken into or entered by strangers	..1..	..2..	..3..	..4..	..5..	..8..	..9..	43 □
b) Having some of your belongings stolen.	..1..	..2..	..3..	..4..	..5..	..8..	..9..	45 □
c) Being attacked and robbed	..1..	..2..	..3..	..4..	..5..	..8..	..9..	47 □
d) Having your home or property damaged by vandals	..1..	..2..	..3..	..4..	..5..	..8..	..9..	49 □
e) Having your car stolen	..1..	..2..	..3..	..4..	..5..	..8..	..9..	51 □
f) Having your car damaged or broken into	..1..	..2..	..3..	..4..	..5..	..8..	..9..	53 □
g) Being assaulted by strangers	..1..	..2..	..3..	..4..	..5..	..8..	..9..	55 □
h) Being assaulted by people you know	..1,,	..2..	..3..	..4..	..5..	..8..	..9..	57 □

Q12 Is there any other crime you are worried about being the victim of which we haven't talked about already?

Yes.................... 1

No 2

No response .... 9 59 □

IF YES, SPECIFY TYPE OF CRIME ______________________________ 61 62 ⧅⧅

____ 64 65 ⧅⧅

Q13 People also often worry about members of their families being victims of crime.

SHOWCARD A Using one of the phrases on this card, could you tell me how worried you are about members of your family... being attached and robbed....

READ OUT EACH IN TURN

					Not			
		Very Worried	Fairly Worried	Not very Worried	at all Worried	No Response	Don't Know	
a)	Being attacked and robbed	..1..	..2..	..3..	..4..	..8..	..9..	67 □
b)	Being raped or sexually attacked	..1..	..2..	...3..	..4..	..8..	..9..	69 □
c)	Being assaulted by strangers	..1..	..2..	..3..	..4..	..8..	..9..	71 □
d)	Being assaulted by people they know	..1..	..2..	..3..	..4..	..8..	..9..	73 □
e)	Having drugs pushed on them	..1..	..2..	..3..	..4..	..8..	..9..	75 □

EXPERIENCE AS A VICTIM

The next few questions are about things that might have happened to you over the last 12 months in which you may have been the victim of a crime or offence. I only want to know about things which have happened to you personally, or to other people living in the same house.

I don't just want to know about serious things - I want to know about small things too. It is often difficult to remember exactly when things happen, so I will take the questions slowly and I would like you to think carefully about them.

ASK ALL

1	2
0	3

CARD NO.

Q14 During the last 12 months have any of the following things happened to you or anyone else living in the same house.

WRITE IN NUMBER OF TIMES. IF NONE OR NOT APPLICABLE CODE 00. IF DON'T KNOW CODE 99.

Has anyone had a motor vehicle or motorcycle:

a) stolen or taken away without permission? □□ 25 26

b) tampered with, damaged or vandalised? □□ 28 29

Q15 Has anyone had anything stolen off a vehicle or out of it? □□ 35 36

Q16 Has anyone had a bicycle stolen or taken away without permission? □□ 38 39

Q17 During the last 12 months have any of the following things happened at your house?

WRITE IN NUMBER OF TIMES. IF NONE OR NOT APPLICABLE CODE 00. IF DON'T KNOW CODE 99.

Has anyone ever tried to get or succeeded in getting inside without permission? □□ 41 42

IF YES, ASK Q18
IF NO, GO TO Q20

Q18 Has anyone got in without permission and stolen anything? □□ 44 45

Q19 Has anyone got in without permission and caused damage? □□ 47 48

Q20 Has anything been stolen by someone who was there when they were allowed to be (eg. tradespeople)? □□ 50 51

Q21 Has anyone stolen milk, milk money or tokens from outside your house or flat? □□ 52 53

Q22 Has anyone stolen anything from outside the house/flat, for example, from the doorstep, garden or garage? □□ 55 56

Q23 Has anyone done any damage on purpose to the outside of the house/flat or to anything on the section? □□ 58 59

The next few questions are about things that may have happened to you personally - not the other people in your household - during the last 12 months. This is about anything that happened to you during the time - at home, in the street, at work, in a shop, in a park, on a train or anywhere else.

Q24 Apart from anything you have mentioned already, in the last 12 months have any of the following things happened to you?

WRITE IN NUMBER OF TIMES. IF NONE CODE 00. IF DON'T KNOW CODE 99

Has anyone stolen anything you were carrying - out of your hands, or from your pocket, or from a bag or case? □□ 61 62 □□

Q25 Has anyone tried to steal something you were carrying? □□ 64 65 □□

Q26 Has anyone stolen anything else - from a cloakroom, an office, a car, or anywhere else you left it? □□ 67 68 □□

Q27 Has anyone tampered with or damaged any of your things on purpose? □□ 70 71 □□

Q28 Has anyone (including people you know) hit you, kicked you or hurt you in any other way? □□ 73 74 □□

Q29 Has anyone frightened you with threats of force or violence in any way? □□ 76 77 □□

Q30 Now I am going to write the things you told me on this page.
CHECK BACK TO QUESTIONS LISTED BELOW. FOR EACH, CODE THE NUMBER OF INCIDENTS - IF NONE CODE 0. NOTE THAT Q17 AND Q21 ARE NOT RE-RECORDED HERE.

1 2
0 4
CARD NO.

Q14(a)	Vehicle theft	□□	4 5 □□
Q14(b)	Damage to vehicle	□□	7 8 □□
Q15	Theft from vehicle	□□	10 11 □□
Q16	Bicycle theft	□□	13 14 □□
Q18	Burglary	□□	16 17 □□
Q19	Break-in with damage	□□	19 20 □□
Q20	Theft from dwelling	□□	22 23 □□
Q22	Theft outside dwelling (NOT MILK BOTTLES, MILK MONEY OR TOKENS)	□□	25 26 □□
Q23	Damage to dwelling	□□	28 29 □□
Q24	Theft from person	□□	31 32 □□
Q25	Attempted theft from person	□□	34 35 □□
Q26	Other theft	□□	37 38 □□
Q27	Other damage	□□	40 41 □□
Q28	Assault	□□	43 44 □□
Q29	Threats	□□	46 47 □□

CHECK Q30. IF ALL CRIMES ARE 0 OR NOT APPLICABLE GO TO Q34.

IF NO CRIME HAS MORE THAN ONE INCIDENT, GO TO Q32.

IF ANY CRIME HAS TWO OR MORE INCIDENTS ASK Q31 FOR EACH CRIME WITH TWO OR MORE INCIDENTS.

Q31 You mentioned...(NUMBER) incidents of..(TYPE OF OFFENCE). Were any of these very similar incidents, where the same thing was done under the same circumstances and probably by the same people?

Yes......	1 RECORD DETAILS BELOW
No.......	0 GO TO Q32

49 □

QUESTION NUMBER · NUMBER OF SIMILAR INCIDENTS IN SERIES

______________ ______________________________

______________ ______________________________

______________ ______________________________

______________ ______________________________

QUESTION NO	NO OF INCIDENTS
51 52 □□	54 55 □□
57 58 □□	60 61 □□
63 64 □□	66 67 □□
69 70 □□	72 73 □□

IF ANY INCIDENTS RECORDED AT Q30

Q32 So can I just check ...

1 2
0 5
CARD NO.

(a) RE-RECORD TOTAL NUMBER OF SERIES OF INCIDENTS FROM Q31.	□□	Series	4 5 □□
(b) RE-RECORD TOTAL NUMBER OF OTHER SINGLE INCIDENTS FROM Q30.	□□	Incidents	7 8 □□
(c) CODE OVERALL TOTAL - (a) + (b)	□□	Series or Incidents	10 11 □□

IF TOTAL AT (c) IS 1, 2, 3 OR 4 - COMPLETE VICTIM FORM FOR EACH INCIDENT OR SERIES OF INCIDENTS - AFTER CHECKING Q34.

IF TOTAL AT (c) IS 5 OR MORE - COMPLETE ONLY FOUR VICTIM FORMS. WORK BACK FROM THE END OF THE LIST AT Q30 AND COMPLETE VICTIM FORMS FOR THE FOUR INCIDENTS OR SERIES YOU COME TO FIRST.

- IF THIS MEANS CHOOSING OUT OF INCIDENTS AT THE SAME QUESTION, TAKE THE MOST RECENT.
- RECORD AT Q33 OVERLEAF THE INCIDENTS OR SERIES RECORDED AT Q 30 FOR WHICH A VICTIM FORM WAS NOT COMPLETED.

IF 5 OR MORE INCIDENTS OR SERIES RECORDED AT Q32(c)

Q33 RECORD BELOW ANY INCIDENTS OR SERIES FOR WHICH VICTIM FORMS WERE NOT COMPLETED. WRITE IN NUMBER OF INCIDENTS AT EACH QUESTION FOR WHICH A VICTIM FORM WAS NOT COMPLETED.

Q14(a)	Vehicle Theft	□□	13 14 □□
Q14(b)	Damage to vehicle	□□	16 17 □□
Q15	Theft from vehicle	□□	19 20 □□
Q16	Bicycle theft	□□	22 23 □□
Q18	Burglary	□□	25 26 □□
Q19	Break-in with damage	□□	28 29 □□
Q20	Theft from dwelling	□□	31 32 □□
Q22	Theft outside dwelling (NOT MILK BOTTLES, MILK MONEY OR TOKENS)	□□	34 35 □□
Q23	Damage to dwelling	□□	37 38 □□
Q24	Theft from person	□□	40 41 □□
Q25	Attempted theft from person	□□	43 44 □□
Q26	Other theft	□□	46 47 □□
Q27	Other damage	□□	49 50 □□
Q28	Assault	□□	52 53 □□
Q29	Threats	□□	55 56 □□

Q34 INTERVIEWER RECORD - ARE THERE ANY VICTIM FORMS TO BE COMPLETED FOR THIS RESPONDENT?

YES....	1	GO TO VICTIM FORM
NO	0	GO TO MAIN QUESTIONNAIRE (PART 2)

58 □

1 2
0 6
CARD NO

1986 COMMUNITY SURVEY

VICTIM FORM

4 5 6 7 8
ID NO.

COMPLETE ONE VICTIM FORM FOR EACH INCIDENT OR SERIES OF INCIDENTS INDICATED BY MAIN QUESTIONNAIRE Q32.

10 11
VICTIM FORM NO.

INTERVIEWER

RE-RECORD FROM MAIN QUESTIONNAIRE Q30

Q35(a) SCREENING QUESTION AT WHICH THIS INCIDENT/ SERIES OF INCIDENTS WAS MENTIONED Q □□ 13 14

Q35(b) This form refers to : One incident only...........1 16
A series of incidents.......2

ASK ALL

Q36 In which month did the incident (or the most recent incident in the series) happen?

INTERVIEWER

CODE IN THE APPROPRIATE QUARTER

October - December 1985.....1
January - March 1986........2
April - June 1986...........3
July - September 1986.......4
October - November 1986.....5
Can't say...................9
Before October 1985.........0

18

NB: IF YOU FIND THAT INCIDENT OCCURRED BEFORE OCTOBER 1985, NOTE THIS AND CLOSE VICTIM FORM HERE, THEN GO TO MAIN QUESTIONNAIRE: PART 2

Q37 Can I just check : Did it happen in New Zealand or did it happen somewhere else?:

New Zealand........	1	ASK Q38
Somewhere else.....	2	CLOSE VICTIM FORM HERE

21 □

Q38 Can you tell me, very briefly, what happened? PROBE FOR OUTLINE DETAILS OF NATURE AND CIRCUMSTANCES OF INCIDENT. RECORD KEY DETAILS ONLY.

22 23 24 25 26

INTERVIEWER: THROUGHOUT THE REST OF THE VICTIM FORM CODE WHEREVER THE ANSWER IS OBVIOUS, OTHERWISE ASK THE QUESTION

ASK ALL

Q39 At what time of day did it happen? (PROMPT WITH CATEGORIES IF NECESSARY)

Daytime (6am - 6pm)................1
Evening (6pm - midnight)...........2
Night (Midnight - 6am).............3
Evening/night (can't say which).....4
Don't know.........................9

28 □

Q40 Did it happen during the week or at a weekend?
NOTE: TAKE WEEKEND AS FRIDAY 6PM TO MONDAY 6AM

During the week....................1
At weekend.........................2
Don't know.........................9

30 □

Q41 Do you know or think you know anything at all about the people who did it?

Yes.....	1 ASK Q42	32
No......	0 ASK Q44	

Q42 Was it someone/were any of them people you knew b fore it happened or was it a stranger/were they all strangers?

All known..................	1	ASK Q43	34
Some known, some not known..	2		
None known..................	3	GO TO Q44	

IF ANY KNOWN

Q43 How well did you know them? Just by sight or just to speak to casually, or did you know (any of) them well?

CODE ALL THAT APPLY

(All/some) just by sight..............	1	36
(All/some) just to speak to casually..	2	38
(All/some) known well.................	3	40

Q44 INTERVIEWER TO CODE: Was anything at all stolen or not?

Yes.......	1 ASK Q45	42
No........	0 GO TO Q48	

Q45 (Including cash), what do you think was the total value of what was stolen? PROMPT WITH PRECODED CATEGORIES IF NECESSARY. 46 □

INTERVIEWER: IF ASKED SAY WE MEAN REPLACEMENT VALUE. CHEQUES/CREDIT CARDS COUNT AS NO VALUE.

Nothing...............................1
Under $10.............................2
$10 but under $50.....................3
$50 but under $200....................4
$200 but under $500...................5
$500 but under $1,000.................6
$1,000 but under $2,000...............7
$2,000 but under $5,000...............8
$5,000+ (enter approximate value)....9
____
No idea...............................0

Q46 Was anything stolen which was of sentimental value to you? IF YES: Did it have a lot of sentimental value for you or just a little? 47 □

No, nothing..........0
Yes - a lot..........1
Yes - a little.......2

Q47 Was any of the stolen money or property recovered? PROBE AS NECESSARY 49 □

Yes - all............1
Yes - some...........2
No - none............0

ASK ALL

Q48 (Apart from things that were actually stolen) did the person/people who did it damage or mess up anything (else) that belonged to you or anyone else in your household (including any damage which may have been done getting in or out)? 51 □

Yes.....................	1 ASK Q49
No damage at all........	0 GO TO Q51

Q49 What did they do? PROBE FULLY. What else? Anything else?

WRITE IN ______________________________

____

____

____

53 54 55 56 57

Q50 What was the total value of the damage they did? IF 'DONT KNOW' PROBE FOR AN ESTIMATE. PROMPT WITH CATEGORIES IF NECESSARY.

Nothing.................	1
Under $50...............	2
$50 but under $100......	3
$100 but under $250.....	4
$250 but under $500.....	5
$500 but under $1,000...	6
$1,000 but under $2,000.	7
$2,000..................	8
No idea.................	9

60

Q51 INTERVIEWER: CHECK Q44 and Q48
WAS ANY PROPERTY STOLEN OR ANYTHING DAMAGED?

Yes.......	1 ASK Q52
No........	0 GO TO Q56

62

Q52 Was any of the property which was stolen or damaged covered by an insurance policy?

Yes.......	1 ASK Q53
No........	0 GO TO
Don't know	9 Q56

64

Q53 Did you or anyone else in your household make a claim for the property which was stolen or damaged?

Yes.......	1 GO TO Q55
No........	0 ASK Q54
Don't know	9 GO TO Q56

66

Q54 Why did you not make an insurance claim?

____

____

GO TO Q56

68 69 70 71 72

Q55 Did the insurance company give you all the money you claimed, some of it or none of it?

All....................	1	
Some....................	2	
None....................	3	74
Still waiting for settlement.............	4	
Don't know.............	9	

ASK ALL

Q56 At the time it happened, did you or anyone else know what was happening?

Yes..........	1 ASK Q57	
No...........	0 GO TO Q61	76
Don't know..	9	

IF ANYONE KNEW

Q57 Who knew about it? CODE ALL THAT APPLY

Respondent	1	78
Other household member...	2	79
Other (WRITE IN)		80

____

____

1	2
0	7

CARD NO.

Q58 Did the person/any of the people who did it hit anyone or use violence against them?

Yes..........	1 ASK Q59	
No...........	0 GO TO	4
Don't know...	9 Q61	

Q59 How many people did they do 'this to?

WRITE NUMBER HERE
IF DON'T KNOW CODE 99

5	6

Q60 Was anyone bruised, scratched, cut or injured in any way?
IF YES ASK: In what way? RECORD BELOW

CODE ALL THAT APPLY

No - not injured.......... 0
Yes - bruises/black eyes... 1
Yes - scratches........... 2
Yes - cuts................ 3
Yes - broken bones......... 4
Yes - other (WRITE IN)

____

____

9 □ 10 □ 12 □ 14 □

ASK ALL

Q61 Because of what happened, did anyone lose time from their job at any stage?

Yes...................... 1
No....................... 0
Don't know............... 9

16 □

Q62 What sorts of practical difficulties or hassles did it cause you or your household? What else? Anything else? PROBE FULLY.

____

____

____

18 19 20 21 22

Q63 Did it cause any other sorts of problems for you or anyone else in your household? PROBE IF NECESSARY: for example, lack of sleep or anxiety? Anything else?

____

____

____

24 25 26 27 28

Q64 After what happened, was there any kind of help or advice, like the things on this card (SHOWCARD B), which you needed but did not get? Was there any other sort of help you needed?

CODE ALL THAT APPLY
(use any pair of boxes)

30 31

Information or advice about: insurance.................01

immediate financial help..02

33 34

compensation.............03

repairs..................04

36 37

legal advice.............05

crime prevention..........06

39 40

Someone to talk to about the crime and your feelings.......07

Advice about effects on children...........................08

41 42

Information on the progress of the case...................09

None of these..00

Other (WRITE IN) ____________________

____

Q65 Going back to the crime itself, did the police get to know about it?

Yes....	1 GO TO Q67
No.....	0 ASK Q66

55

IF NO AT Q65

Q66 Why not? PROBE FULLY.

____

____

____

____

57 58 59 60 61

GO TO Q73

IF YES AT Q65

Q67 How did they get to know about it?

Police told by respondent..................	1	ASK Q68
Police told by other person on respondent's behalf............................	2	
Police told by other person................	3	GO TO Q69
Police were there..........................	4	
Police found out another way (WRITE IN) ______________________________		
Don't know................................	9	

67 □

Q68 People sometimes don't tell the Police about crimes which are committed.
Why did you decide to report this crime? Any other reasons?

____

____

____

____

ASK Q69

69 70 71 72 73

IF POLICE GOT TO KNOW ABOUT MATTER (YES AT Q65)

Q69 Did you tell the Police everything you knew or suspected about the incident?

Yes......	1 GO TO Q71
No.......	0 ASK Q70

75 □

Q70 Why not? PROBE FULLY.

____

____

____

____

76 77 78 79 80

1 2

0 8

CARD NO.

Q71 Overall were you satisfied or dissatisfied with the way the Police dealt with the matter?

IF SATISFIED ASK: Very satisfied or just fairly satisfied?

IF DISSATISFIED ASK: A bit dissatisfied or very dissatisfied?

Very satisfied.........	1	
Fairly satisfied.......	2	GO TO Q73
A bit dissatisfied.....	3	ASK Q72
Very dissatisfied......	4	
Don't know/can't say...	9	GO TO Q73

4

Q72 Why were you dissatisfied? What other reason? Any other reason?

CODE ALL THAT APPLY
(use any pair of boxes)

6 7 / 9 10 / 12 13 / 15 16 / 18 19 / 21 22 / 24 25

They were slow to arrive when sent for......	01
They did not come when sent for............	02
They didn't investigate matter/did not do enough to investigate matter............	03
They seemed uninterested..................	04
They made mistakes/handled matter badly....	05
They didn't recover property...............	06
They didn't apprehend the offenders........	07
They failed to keep the respondent informed of progress of investigation......	08
They were impolite/unpleasant..............	09
They didn't believe me/they accused me......	10
Other (WRITE IN)	
Don't know....................................	99

Q73 INTERVIEWER RECORD IF ANYONE ELSE WAS IN THE ROOM DURING VICTIM FORM. CODE ALL THAT APPLY

No-one else present...............	1	
Yes - child/ children under 16....	2	NOW GO TO MAIN QUESTIONNAIRE (PART 2)
Yes - adult/adults................	3	ASK Q74

27

Q74	SPECIFY RELATIONSHIP OF ALL ADULTS PRESENT TO RESPONDENT CODE ALL THAT APPLY	Wife/Husband/Partner.....1	29 □
		Mother/Father............2	
		Daughter/Son.............3	31 □
		Sister/Brother...........4	
		Other relative(s)........5	33 □
		Friend(s)................6	
		Workmate(s)..............7	35 □
		Other....................8 ____________________	

NOW GO TO MAIN QUESTIONNAIRE (PART 2)

1	2
0	9

CARD NO.

4	5	6	7	8

ID NO.

1986 COMMUNITY SURVEY

MAIN QUESTIONNAIRE : PART 2

THIS QUESTIONNAIRE IS USED WITH ALL RESPONDENTS

ASK ALL

Q75 Apart from your own household, do you know anyone in the Wellington area who has been the victim of a crime in the last 12 months?

Yes...........	1	ASK Q76
No............	0	GO TO Q78
Don't know....	9	

11

Q76 How many people?

WRITE NUMBER HERE □□

13 14

Q77 What sorts of offences were they the victims of?

CODE ALL THAT APPLY (use any box)

Burglary..................	1
Vehicle theft.............	2
Theft from vehicle........	3
Theft from person.........	4
Vandalism.................	5
Attack and robbery........	6
Assault by a stranger.....	7
Sexual assault............	8
Other (WRITE IN)	

____

____

16 □

18 □

20 □

22 □

24 □

26 □

28 □

ASK ALL

Q78 I would now like to ask you about the way in which crime problems are handled in this area. First, do you think the Police are doing enough about the crime problems in this area?

Yes......................1
No.......................0
No crime around here......2
Don't know................9

30 □

Q79 Apart from what the Police are doing, what sort of things are being done by anyone else in your area to prevent crime? PROBE FULLY.

____

____

____

____

32 33 34 35

Don't know................9

Q80 Do you think the community is doing enough about the crime problems in this area?

Yes....................	1	GO TO Q82
No.....................	0	ASK Q81
No crime around here....	2	GO TO Q82
Not a community responsibility	3	
Don't know ,	9	

37 □

Q81 What more do you think the community could do?

____

____

____

____

32 33 34 35

Don't know............. 9

ASK ALL

Q82 Do you think the Police talk enough with the community about problems which local people are concerned about?

44

Yes.................... 1
No..................... 0
Don't know............. 9

Q83 Which of these statements (SHOWCARD C) is closest to the way you feel about the service the Police provide around here?

46

Very satisfied......... 1
Satisfied.............. 2
Neither satisfied nor
dissatisfied.... 3
Dissatisfied........... 4
Very dissatisfied...... 5
Don't know............. 9
No response 8

Q84 Various suggestions have been made about improving police services in New Zealand. Do you think the police should try any of the following in your area? Anything else? PROBE FULLY.

(SHOWCARD D) (use any pair of boxes)

Having more police officers overall.........1

Having more detectives.....................2

Having more police car patrols.............3

Having more police foot patrols............4

Encouraging community crime prevention......5

Having more community constables...........6

Getting more involved themselves with the community (eg. through schools, sports groups etc)..........................7

Other (WRITE IN)

____

____

____

Don't know..................................99

48 49 □□

51 52 □□

54 55 □□

57 58 □□

60 61 □□

63 64 □□

66 67 □□

Q85 In the last 12 months have you:
(Use adjacent box)

	Yes	No	
Reported a crime to the police, including things we have already talked about......................	1	0	69 □
Reported to the police that some property has been lost............	1	0	71 □
Reported a missing person to the police	1	0	73 □
Reported any other sort of problem, difficulty or disturbance to the police............................	1	0	75 □
Asked the police for advice on crime prevention..................	1	0	77 □
Asked a police officer for directions........................	1	0	79 □
Asked the police for any other sort of advice, help or information.....	1	0	80 □

INTERVIEWER:

CHECK ANSWERS TO Q85. IF NO TO ALL, SKIP TO Q90. IF YES TO ANY, ASK Q86.

IF ORIGINATED CONTACT WITH POLICE (YES TO ANY PART OF Q85)

1	2
1	0

CARD NO.

Q86 When you have wanted help from the police, have you generally found them helpful or unhelpful or is your experience mixed?

IF HELPFUL Very helpful or only fairly helpful?

IF UNHELPFUL Very unhelpful or only a bit unhelpful?

Very helpful.........	1	GO TO
Fairly helpful.......	2	Q88
Mixed experience.....	3	ASK
A bit unhelpful......	4	Q87
Very unhelpful.......	5	

4
□

IF MIXED FEELINGS OR FELT IT TO BE UNHELPFUL

Q87 Why do you say that?

____

____

____

ASK Q88

5	6	7	8

Q88 And when you have wanted help from the police, have you found them generally pleasant in the way they talked to you or generally unpleasant or is your experience mixed?

IF PLEASANT Very pleasant or only fairly pleasant?

IF UNPLEASANT Very unpleasant or only a bit unpleasant?

Very pleasant........	1	GO TO
Fairly pleasant......	2	Q90
Mixed experience.....	3	
A bit unpleasant.....	4	ASK
Very unpleasant......	5	Q89

10
□

IF MIXED EXPERIENCE OR FOUND THE POLICE UNPLEASANT

Q89 Why do you say that?

____

____

____

12 13 14 15 16

ASK Q90

ASK ALL

Q90 Have you been approached or stopped by the police for any reason at all in the last 12 months?

Yes.....	1 ASK Q91
No......	0 GO TO Q95

18

IF APPROACHED BY POLICE (YES AT Q90)

Q91 How many times?

WRITE NUMBER HERE
IF DON'T KNOW CODE 99 □□

20 21

Q92 Why were you approached/stopped (on each of these occasions)?

____

____

____

____

23 24 25 26 27

Q93 When you were approached by the police, did you find themgenerally polite or generally impolite or is your experience mixed?

IF POLITE Very polite or only fairly polite?

IF IMPOLITE Very impolite or only a bit impolite?

Very polite.........	1	GO TO
Fairly polite.......	2	Q95
Mixed experience....	3	ASK Q94
A bit impolite......	4	
Very impolite.......	5	

32

IF MIXED EXPERIENCE OR FOUND THE POLICE IMPOLITE

Q94 Why do you say that?

____

____

34 35 36 37 38

GO TO Q95

ASK ALL

Q95 Do you have any relatives, neighbours, friends or people you know through work who are in the police?

Yes........	1	GO TO Q97
No..........	0	ASK Q96
Don't know..	9	
No response.	8	

40 □

Q96 Do you know any police officers well enough to talk to?

Yes.........	1
No..........	0
Don't know..	9
No response.	8

42 □

ASK ALL

Q97 In the last 12 months have you had anything else to do with the police at all which we have not talked about already.

Yes.......	1	ASK Q98
No.........	0	GO TO Q99
Don't know.	9	
No response	8	

44 □

Q98 What was that about?

____

____

____

46	47	48	49	50
⧅	⧅	⧅	⧅	⧅

INTERVIEWER: CHECK ANSWERS TO QUESTIONS 85, 90 AND 97. IF "YES" TO ANY, ASK Q99. OTHERWISE GO TO Q112.

Q99 In the last 12 months have you been really annoyed about the way a police officer has behaved or about the way the police have handled a matter in which you have been involved?

Yes.......	1	ASK Q100
No........	0	GO TO Q109
No response	8	

55 □

IF ANNOYED IN PAST 12 MONTHS (YES AT Q99)

Q100 How often in the past 12 months? WRITE IN: □□ TIMES

58 59

If don't know Code 99

Q101 (Last time you were really annoyed) what happened that annoyed you? PROBE FOR OUTLINE DETAILS OF TYPE OF INCIDENT AND CAUSE OF ANNOYANCE. RECORD VERBATIM.

____

____

____

59 60 61 62 63

Q102 In the last 12 months when you were annoyed about police behaviour, did you ever feel like complaining to someone about it?

Yes..........	1	ASK Q103
No...........	0	GO TO Q109

70

Q103 IF YES AT Q102.Did you actually make an official complaint?

Yes..........	1	ASK Q104
No...........	0	GO TO Q108

72

0	1
1	1

CARD NO.

Q104 IF YES AT Q103 Who did you make the complaint to?

CODE ALL THAT APPLY (use any pair of boxes)

Local police station..........	1
Senior police officer.........	2
Commissioner of Police........	3
Minister of Police............	4
Member of Parliament..........	5
Governor General..............	6
Ombudsman.....................	7
Lawyer........................	8
Other (WRITE IN)	

____

____

4 5

7 8

10 11

13 14

Q105 How did you make the complaint? <u>CODE ALL THAT APPLY</u>

By phone.............. 1 — 16 □

By letter............. 2 — 18 □

In person............. 3

Other (WRITE IN) — 20 □

____

____

Q106 What was the result?

____

____

22	23	24	25	26
⧵	⧵	⧵	⧵	⧵

Q107 How did you feel about the way in which your complaint was handled? (SHOWCARD E)

Very satisfied.......	1	GO TO Q109	28 □
Satisfied............	2		
Mixed feelings.......	3		
Dissatisfied.........	4		
Very dissatisfied....	5		

Q108 <u>IF NO AT Q103</u> Why not? PROBE FULLY. RECORD VERBATIM.

____

____

____

30	31	32	33	34
⧵	⧵	⧵	⧵	⧵

Q109 In the last 12 months have you been really pleased about the way a police officer behaved towards you or someone you know <u>or</u> about the way the police handled a matter in which you were involved?

Yes........	1	ASK Q110	36 □
No.........	0	GO TO	
No response.	8	Q112	

IF PLEASED IN PAST 12 MONTHS (YES AT Q108)

Q110 How often in the past 12 months? WRITE IN: □□ 37 38

Q111 (Last time you were really pleased) what happened that pleased you? PROBE FOR OUTLINE DETAILS OF INCIDENT AND CAUSE OF PLEASURE. RECORD VERBATIM.

____

____

____

39 40 41 42 43

ASK ALL

Q112 The next questions are about some things people might do simply to protect themselves against crime when they go out after dark. When you are out after dark, how often simply to be careful do you .... try not to walk near certain sorts of people? RECORD BELOW
REPEAT FOR EACH ITEM. SHOWCARD F

How often simply to be careful do you.....

	Always	Usually	Some times	Rarely	Never	Not Applicable	Don't know	
a) Try not to walk near certain sorts of people	..1..	..2..	..3..	..4..	..5..	..6..	..9..	45 □
b) Stay away from certain streets or areas	..1..	..2..	..3..	..4..	..5..	..6..	..9..	47 □
c) Go out with someone else rather than by yourself	..1..	..2..	..3..	..4..	..5..	..6..	..9..	49 □
d) Try not to use buses or trains	..1..	..2..	..3..	..4..	..5..	..6..	..9..	51 □
e) Use a car rather than walk	..1..	..2..	..3..	..4..	..5..	..6..	..9..	53 □

Q113 Are there any places which you avoid going to because you are worried about crime or violence?

Yes..........	1	ASK Q114
Sometimes...	2	
No..........	0	GO TO Q115

54

Q114 Which places? PROBE Why do you say that?

____

____

____

56 57 58 59 60

ASK ALL

1 2

CARD NO.

Q115 Have you taken any of the following measures to prevent crime?

SHOWCARD G

CODE ALL THAT APPLY
Use any pair of boxes

Put more/better locks (on doors and windows)............	01
Put in stronger doors..............	02
Arranged to have mail collected when going away..........	03
Left lights/radio on when going out or going away...........	04
Put a burglar alarm on house.......	05
Put a burglar alarm in car.........	06
Got a guard dog	07
Taken a self-defence course........	08
Got a project identification kit...	09
Joined a Neighbourhood Watch or Neighbourhood Support Group........	10

Other (WRITE IN)

____

____

None.............................. 00

Q116 INTERVIEWER: IF RESPONDENT BELONGS TO A NEIGHBOURHOOD WATCH OR NEIGHBOURHOOD SUPPORT GROUP, ASK WHICH IT IS CALLED AND CODE BELOW.

Neighbourhood Watch.....	1	ASK
Neighbourhood Support...	2	Q117
Neither................	0	GO TO Q121

39 □

Q117 How is Neighbourhood Watch/Support organised in your area?

41 42 43 44

____

____

____

____

Q118 Why did you join Neighbourhood Watch/Support? PROBE FULLY. (If more than three reasons, probe for three principal reasons).

46 47 48 49

____

____

____

____

Q119 Do you think it works?

Yes...........1
No............0
Don't know....9

50 □

Q120 Why do you think that? PROBE FULLY.

52 53 54 55

____

____

____

____

GO TO Q128

Q121 What do you think Neighbourhood Watch or Neighbourhood Support Groups are?

____

____

____

____

57 58

GO TO Q122

No idea........ GO TO Q129

Q122 Are there any groups like this in your area?

Yes............	1	ASK Q123
No.............	0	GO TO
Don't know......	9	Q124

59

IF YES

Q123 What are some of your reasons for not belonging to them?
(IF MORE THAN 3 REASONS, PROBE FOR THREE PRINCIPAL REASONS).

____

____

____

____

60 61 62

GO TO Q128

IF NO

Q124 Do you think such groups would work in this area?

Yes............	1	GO TO Q126
No.............	0	
Maybe..........	2	ASK Q125
Don't know......	9	

64

Q125 Why do you think they would/might not work?

____

____

____

____

66 67 68 69

GO TO Q128

ASK ALL WHO THINK A SCHEME WOULD WORK IN THIS AREA

Q126 Would you personally be prepared to join one of these schemes?

Yes, would be prepared...:	1	GO TO Q128	71 □
No, would not be prepared..	2	ASK Q127	
Don't know................	9		

Q127 Why do you say that?

72 73 74

GO TO Q128

Q128 What sort of role do you think the police should play in such schemes? PROBE FULLY.

75 76 77

ASK ALL

Q129 Before we do the last short section, is there anything else you would like to say about crime or the police in your area? RECORD VERBATIM.

78 79 80

NOW GO TO DEMOGRAPHIC QUESTIONNAIRE.

1	2
1	3

CARD NO.

4	5	6	7	8

ID NO.

1986 COMMUNITY SURVEY

DEMOGRAPHIC QUESTIONNAIRE

COMPLETE THIS QUESTIONNAIRE FOR ALL RESPONDENTS

Q130 I would like to finish by asking some questions about you and your household to help us work out the results.

Firstly, is there more than one flat or occupied dwelling on this property?

Yes.....	1 ASK Q131
No......	0 GO TO Q132

11

Q131 How many flats/ occupied dwellings in total, including this one?

WRITE NUMBER □□

13 14

Q132(a) Are there any other houses in this street with the same street number as you (eg. 24A, 24B etc).

No........	0 GO TO Q133
Yes.......	1 ASK Q132(b)

16

(b) How many?

WRITE NUMBER
IF DONT KNOW CODE 99 □□

18 19

N.B. INTERVIEWER TO CHECK WITH FIRST PAGE OF QUESTIONNAIRE.

ASK ALL

Q133 Which of the following best describes this household?
(SHOWCARD H)

One person living alone...........	1
Solo parent with child/children...	2
Couple without children...........	3
Couple with children	4
Extended family/whanau...........	5
Flatmates.........................	6
Other (WRITE IN)	

____

21

Q134 Can I check some details about the members of your household?

First, how many people in total are there in your household, INCLUDING YOU?

WRITE NUMBER □□

23 24
□□

Q135 How many of the people living in this household are aged 16 and above?

WRITE NUMBER □□

27 28
□□

Q136 I would now like to record some details about each person living in this household.

RECORD DETAILS OF EACH HOUSEHOLD MEMBER BELOW

	SEX	AGE	RELATIONSHIP TO RESPONDENT
	Male....1 Female..2	Don't know = 99 Refused = 98	Spouse/cohabitee...........1 Parent/parent-in-law ......2 Brother/sister (or in-law).3 Son/daughter (or in-law)...4 Other relative.............5 Non-relative...............6
RESPONDENT	30 □	32 33 □□	
PERSON 2	35 □	37 38 □□	40 □
PERSON 3	42 □	44 45 □□	46 □
PERSON 4	49 □	51 52 □□	54 □
PERSON 5	56 □	58 59 □□	61 □
PERSON 6	63 □	65 66 □□	68 □

CODE Q136 FROM MAIN RECORDING SHEET.

1	2
1	4

CARD NO.

PERSON 7	4 □	6 7 □□	9 □
PERSON 8	11 □	12 14 □□	16 □
PERSON 9	18 □	20 21 □□	22 □
PERSON 10	25 □	27 28 □□	30 □
PERSON 11	32 □	34 35 □□	37 □
PERSON 12	39 □	41 42 □□	44 □
PERSON 13	46 □	48 49 □□	51 □
PERSON 14	53 □	55 56 □□	58 □
PERSON 15	60 □	62 63 □□	65 □

Q137 To which of the following groups do you belong? READ OUT

1	2
1	5

CARD NO.

NZ Maori..............	1
Cook Island Maori......	2
European/Pakeha.......	3
Pacific Islander......	4
Other................. (WRITE IN) ____________	
No response..........	8

4
□

Q138 (a) Do you rent this house/flat or own it?

Owned (including with mortgage)....	1 GO TO Q139
Rented............	2 ASK Q138(b)

9 □

(b) Who do you rent from?

Private owner........	1
Local authority......	2
Government...........	3

10 □

Q139 Which of the following best describes your work situation: (SHOWCARD I)

Working in paid employment:

(a) less than 10 hours per week......	1	ASK Q140
(b) between 10 and 30 hours per week.	2	
(c) more than 30 hours per week......	3	
Full-time household/parenting responsibilities.....................	4	GO TO Q142
Retired..........................	5	
Unemployed..........................	6	
Sick or disabled and unable to work..	7	
In full-time education..............	8	

12 □

IF WORKING IN PAID EMPLOYMENT

Q140 What is your job - BE SPECIFIC (eg. Primary School teacher, self employed building contractor etc).

____

14 ☒

Q141 Are you the main income earner in this house?

Yes..........	1 CLOSE INTERVIEW HERE
No..........	0 ASK Q142
No main income earner........ (eg. flat, dual career households etc).	2 ASK Q143

16 □

Q142 What is the job of the main income earner in this house? BE SPECIFIC.

18 ⧅

____

CLOSE INTERVIEW HERE.

No main income earner.......... 0 ASK Q143

Q143 IF NO MAIN INCOME EARNER

What is the main source of income for this house?

(eg. superanuation, unemployment benefit etc.)?

20 ⧅

____

THANK YOU VERY MUCH FOR TAKING PART IN THIS SURVEY.

A.2 1996 NEW ZEALAND NATIONAL SURVEY OF CRIME VICTIMS

Note that the following copy of the 1996 New Zealand National Survey of Crime Victims has added annotations.

Q → "W" for write.

AGB McNair

Job No: 101208

Household Number: □□□ HH

Address : ______________________________

Date: □□ □□ 9 6 DATE

Day **Month**

Area Name : □

Area Unit No : □□□□ AU

Calls to Obtain: □ [illegible] 67

Interviewer No : □□□□ 61

Start Time of Interview : □□□□

1. INITIAL INTRODUCTION

I am from AGB McNair, the survey company. We are doing a New Zealand-wide survey on behalf of the New Zealand Police/Victoria University and a number of other government departments, and its purpose is to get an accurate measure of how much crime there is in the community and what affect it has by asking people what sort of things have happened to them or other people in their household.

You may have seen or heard some publicity about the survey recently. This is a very important survey because its results will help government departments to plan how to deal with crime more effectively.

Read the following if necessary.

I have here a sheet which gives you more information about the survey. You can read it now if you like, bu I am happy to leave it with you at the end of the interview.

When it is clear that the initial contact is co-operative, proceed to respondent selection.

2. RESPONDENT SELECTION

It is very important that we interview a representative selection of New Zealanders for this survey, so I have to ask you a few questions to help me select who in your household is the right person for me to talk to.

Can you please tell me the first name and birth month of everyone aged 15 and over who lives here at the present time.

OCCUPANTS OF HOUSEHOLD (AGED 15+)

PERSON NO.	FIRST NAME	MONTH OF BIRTH (1-12)	TICK WHEN SELECTED
1			
2			
3			
4			
5			
6			
7			

Probe to ensure all residents 15+ are included. The interview is conducted with the person who has the next birthday.

....(NAME)... is the person I need to interview. Is s/he available now?

If not available now, establish when and make appointment.

3. RESPONDENT INTRODUCTION

Repeat initial introduction if respondent and initial contact are different people.

Hand information sheet and consent form to respondent

Before we get underway with the interview, I have to get you to read this sheet and sign this consent form. This is a formality we have to go through with everyone who takes part in the survey, and it just confirms that you understand what the survey is about, and that you have agreed to participate in it.

Consent form signed
Yes.....................1
Terminate ← No.....................2

The only other thing I would like to mention at this point is that if you would like to talk to a Helping Organisation about anything you talk to me about, I can give you the telephone number of an organisation to contact.

Q1 *In general, what kind of neighbourhood would you say you live in. Would you say it is a neighbourhood in which people do things together and try to help each other or one in which people mostly go their own way?*

W1

Help each other..............1
Go own way....................2
Mixture...........................3
Don't know.....................9

Q2 *How long have you lived in this neighbourhood?*

W2

Less than 1 year..............................1
1 year but less than 18 months........2
18 months but less than 3 years 3
3 years but less than 5 years............4
5 years but less than 10 years..........5
10 years or more.............................6

(Explain - only if asked: "This neighbourhood" = the streets around them, or for rural people this is their district.)

Q3 *Do you think there is a crime problem in this neighbourhood?*

W3

Yes.............................1
No..............................2
Don't know................9

No / Don't know ← Q5

Q4 *What sort of crime problems do you think ther are in this neighbourhood?*

Probe to No
Code all mentions below

W4A1
W4A2

Burglary, break-ins 01
Vandalism...............................02
Graffiti 03
Street attacks...........................04
Petty thefts..............................05
Assault 06
Domestic Violence 07
Sexual Crimes.........................08
Car theft..................................09
Theft from cars 10
Damage to cars 11
Drink driving 12
Prowlers..................................13 — W4A13 14-16 ?
Other (Specify)
_________________________37 — W4A 37 38 ?
Don't know.............................39 — W4A39

Q5 *Do you think that in the last 12 months there has been more or less crime in your neighbourhood than before, or has it stayed about the same?*

Probe: *Is that a lot or a little more/less?*

W5

A lot more crime.................1
A little more crime..............2
About the same 3
A little less crime................4
A lot less crime....................5
No crime around here 6
Don't know 9

Showcard A

Q6 *Using the categories on Page A, can you tell me how much of a problem you think the following things are in your neighbourhood?*

Read and code from coding box below

Very big problem 1
Fairly big problem............................... 2
Not a very big problem 3
Not a problem at all............................ 4
Don't know 9

a1
a. .. Rubbish and litter lying about on streets or empty sections .... _____
b. .. Broken windows in shops and public buildings, and other vandalism or deliberate damage to property. .... _____
c. .. Graffiti on walls, schools, shops, churches etc. .... _____
d. .. Uncontrolled dogs roaming the neighbourhood............................. _____
e. .. Teenagers hanging around on the streets.................................... _____
f. .. Gang members living or hanging around in the neighbourhood........ _____
7
g. .. Drunks, glue sniffers or people high on drugs on the streets... _____

Showcard B

Q7 *Looking at Page B, can you tell me about how often you go out at night, that is, after dark?*

W7
Never .... 1✱
At least once a week 2
At least once a fortnight.............. 3
At least once a month 4
Less often than once a month 5
Don't know 9

Q9 ← (codes 2–9)

Q8 *Are there any particular reasons why you never go out at night?*

Probe to no.
Code all mentions below

W8A1 Age/disability 01
Don't want to/nowhere interesting to go 02
/ Have to look after children, sick relatives or other dependents 03
✓ Lack of transport...04
Might be attacked/robbed 05
Might have things stolen from me 06
Might be burgled..07
General fear that something might happen 08
No money/too expensive 09
Busy working at home 10
Content to stay in/watch TV etc. 11
W8A12 Don't like night driving 12
13. Other (specify)

38. ______________________________________37
W8A39 Don't know 39

Now go to Q11

Q9 *Do you ever walk alone in your neighbourhood after dark?*

W9
Yes 1
Q11 ← No 2

Q10 *How safe do you feel walking alone in your neighbourhood after dark? Would you say you feel...*

Read out

W10
..Very safe............................... 1
..Fairly safe 2
..A bit unsafe............................ 3
..Very unsafe............................ 4
(Do Not Read) Don't know... 9

Now go to Q12

Q10/Q11 in combination
= Q10_11

Q11 *If you were to walk alone in your neighbourhood after dark, how safe would you feel? Would you say that you would feel...*

W11

Read out

..Very safe	1
..Fairly safe	2
..A bit unsafe	3
..Very unsafe	4
(Do Not Read) Don't know	9

Showcard C

Q12 *Some people worry about being the victim of a crime. I am going to read out some types of crime and I would like you to tell me for each one, how worried you are about being a victim of it, using the categories on Page C.*

Read and code from coding box below

Very worried	1
Fairly worried	2
Not very worried	3
Not at all worried	4
Not applicable	6
Don't know	9

W12a1

a. .. Having your house burgled .... _____
b. .. Having your car stolen .... _____
c. .. Having some of your belongings stolen .... _____
d. .. Being attacked and robbed .... _____
e. .. Having your home or property damaged by vandals .... _____
f. .. Having your car deliberately damaged or broken into .... _____
g. .. Being assaulted by strangers .... _____
h. .. Being assaulted by people you know .... _____
i .. Being assaulted because of your race .... _____
j .. Being in a traffic accident caused by a drunk driver .... _____

Ask female respondents only

k. .. Being sexually assaulted or raped .. _____
W12a11

Showcard D

Q13 *Using one of the phrases on Page D, could y tell me how worried you are about any of the following things happening to you?*

Read and code from coding box belo

Very worried	1
Fairly worried	2
Not very worried	3
Not at all worried	4
Not applicable	6
Don't know	9

W13A1

a. .. Becoming seriously ill.................... _____
b. .. Someone else in your household becoming seriously ill .... _____
c. .. Being unable to cope with debts .... _____
d. .. You or someone in your household losing their job .... _____
e. .. You or someone in your household having a road accident............ _____
f. .. You or someone in your household being seriously injured in an accident in your home .... _____

W13A6

Q14 *Again using one of the phrases on Page D, could you tell me how worried you are abou being racially harassed by people on the street?*

W14

Very worried	1
Fairly worried	2
Not very worried	3
Not at all worried	4
Don't know	9

I would now like to ask you a few questions about the sorts of things people do to protect themselves against crime.

Q.15 Check back to Q7, if 'never' (code 1✱) coded → Q16, all others continue

Showcard E

When some people go out at night they are careful to avoid situations that might put themselves at risk. Using Page E, could you tell me how often when you are out at night you do any of the following things in order to protect yourself against crime.

Read and code from coding box below

Always	1
Mostly	2
Sometimes	3
Rarely	4
Never	5
Not applicable	6
Don't know	9

5a1

a. .. Try not to walk near certain sorts of people ______

b. .. Stay away from certain streets, areas or activities ______

c. .. Go out with someone else rather than by yourself ______

d. .. Try not to use buses or trains ______

e. .. Use a car or taxi rather than walk... ______

f. .. Carry a weapon or something you could use as a weapon ______

g. .. Carry a personal alarm of some sort ______

5a7

Q16 *Over the last couple of years, would you say that you have changed where you go or what you do outside your home, in order to protect yourself against crime? Or would you say you still do much the same things?*

W16

Have changed 1
Still do much the same things 2
Don't know 9

Q17a *Have you ever learned self-defence or martial arts?*

W17a

Yes 1
Q18 ← No 2

Q.17b *Would you say your reason for learning self defence or martial arts was solely to protect yourself against crime, partly to protect yourself against crime, or did crime protection have nothing to do with it?*

W17b

Solely protection 1
Partly protection 2
Nothing to do with it 3

Q18 *Thinking of the various things which people can do to protect their homes from burglary - like having burglar alarms or better locks on doors. Do you think things like this make homes....*

W18

Read out

..A lot safer 1
..A little safer 2
..No safer 3
(Do not read) Don't know 9

Q19 *I am now going to read out some security measures that people can have, and I would like you to tell me which, if any, you have at your house? Do you have .....?*

Read out and code all mentioned

W19a1 ..Burglar alarm on premises......................01
..Doors with double locks or dead locks...02
..Security chain on doors 03
..Security bolts on doors 04
..Security screen on doors..........................05
..Windows with keys to open them...........06
..Bars or grilles on windows 07
..Safety latch to prevent window
opening fully..08
..A guard dog 09
..Lights, radio or television on a
timer switch 10
..Outside lights on a sensor switch 11
..Security markings on household
property..12
W19a13 ..Surveillance by security firm..................13
14-22 ..Any other security measures (specify)
19=37 ________________________________ 37
None 38
19=39 **(Do not read)** Don't know......................39

Q19TOT = [illegible]

Q20 *In order to prevent burglary, some people leave lights or radio or television on in their house when they are going out. How often, if ever, do you do this? Do you....*

Read out

W20 ..Always 1
..Mostly..2
..Sometimes 3
..Rarely..4
..Never 5
(Do not read) Don't know.....................9

Q21 *Taking everything into account, how difficult do you think it would be for a burglar to get into your home. Do you think it would be...*

W21 **Read out**

..Very easy 1
..Fairly easy..............................2
..Fairly difficult.........................3
Q23 ← ..Very difficult 4
Don't know 9

Q22 *Is there any particular reason why your household hasn't done more to protect your home from possible burglary?*

Code all mentions

W22A1 Can't afford to 01
Nothing more/don't know what more
can be done 02
Wouldn't work/wouldn't be effective 03
Haven't got around to it/can't be
bothered 04
Because it is a rented property..................05
Neighbourwatch/ neighbours home
all the time 06
Area safe/not much crime.........................07
W22A8 Someone home all the time 08
9-14 Other (specify)
________________________________ 37
Don't know..39

Q23 *Does your household tell a neighbour when everyone in the house is going to be away - f example on holiday or for a trip?*

W23
Yes, always................1
Yes, sometimes..........2
Never 3
Don't know................9

Q24 *Do any of your next door neighbours know the times when your home is regularly unoccupied each day?*

W24
Yes 1
No..................................2
Never/rarely left empty ..3
Don't Know....................9

Q25 *Do you currently have any car, van or truck which you drive regularly?*

w25

Yes .....1 ✱
Q28 ← No.......2

Q26 *How often do you lock the vehicle when it is parked away from your property?*

Read out

w26

Q28 ← ..Always 1
..Mostly 2
..Sometimes...........3
..Rarely..................4
..Never...................5

Q27 *Is there any particular reason why you don't (always) lock it?*

Code all mentions

Forget to...1
Only away from car for short time..............2
7a1 Don't think anything is going to happen to it...3
Don't think locking it reduces likelihood of crime...4
Door(s) don't lock 5
7a7 Other (specify)
8?. ______________________________________7 ←
7a9 Don't know..9

add'l codes cited

Q28 *People who have been the victim of a crime sometimes need help or assistance. Do you know of any community services apart from the police, which would be available to you if you were a victim of crime? Which ones?*

Probe to no.

w28a1 Victim Support Group.....................01
Rape Crisis.......................................02
Women's Refuge 03
HELP/Sexual Abuse Centre............04
Citizens Advice Bureau 05
Iwi or other Maori organisation 06
Hospital...07
w28a8 Samaritans..08
9-22?
w28a37 Other (specify)
38? ______________________________________37
w28a38 Don't know of any 39

EXPERIENCE AS A VICTIM

The next few questions are about things that might have happened to you or your household over the last .... months, that is since the beginning of last year, in which you may have been the victim of a crime or offence. I only want to know about things which have happened to you personally, or to other people in your household **(Explain: "household" means people living with you.)** *I also only want to know about things that have happened in New Zealand.*

I don't just want to know about serious things - I want to know about small things too. It is often difficult to remember exactly when things happen, so take what time you need.

Q29 **Check back to Q.25, if yes (code 1✱) coded, code Q.29 as Yes (code 1)**

W29

First, has anyone in this household owned or had the regular use of a car, van or truck at any time since 1st January 1995?

Yes...............1
\boxed{Q31} ← No................2

Q30 *How many cars, vans or trucks do members of this household own or have regular use of now?*

W30 **Write in (2 digits) None=00** ____ ____

Q31 *Has anyone in this household owned or had the regular use of a motorcycle at any time since 1st January 1995?*

W31

Yes...............1
\boxed{Q33} ← No................2

Q32 *How many motor cycles do members of this household own or have regular use of now?*

W32

Write in (2 digits) None=00 ____ ____

TOTVEH = W30+W32

●Q33 **Check back to Q.29 and Q.31, if no (code 2) coded at both → Q.36**

Since 1st January 1995 have you or anyone else now in your household had their motor
W33 *vehicle or motor cycle stolen or taken away without permission? How many times?*

TheftVI **Write in (2 digits) None=00** ____ ____

◆Q34 *And (apart from this) has anyone had anything stolen from or off their vehicle (suc as vehicle parts or personal possessions)? How many times?*

W34

TheftFVI **Write in (2 digits) None=00** ____ ____

■Q35 *And (apart from this) has anyone had their vehicle tampered with, damaged or vandalised? How many times?*

W35 IntVeh

Write in (2 digits) None=00 ____ ____

◇Q36 *Still thinking back to the period since the beginning of 1995, has anyone tried to or succeeded in getting into your home or gara without permission in order to commit some sort of offence? This includes a holiday hon*
W36 *if you have one. How many times?*

Write in (2 digits) None=00 ____ ____

Q37 *And in that time has anyone stolen anything from outside your home or holiday home (su as from your front gate or garden or shed) which was worth less than $10 (eg milk bott*
W37 *newspapers) How many times?*

Write in (2 digits) None=00 ____ ____

➢**Q38** *And in that time has anyone stolen anything from outside your home or holiday home which was worth more than $10? How many times?*

Write in (2 digits) ____ ____
None=00

✵**Q39** *And (apart from this) since 1st January 1995 has anything been stolen from inside your home or garage by someone who was allowed to be there (eg tradesperson or invited guest)? This includes a holiday home if you have one. How many times?*

Write in (2 digits) ____ ____
None=00

*The next few questions are about things that may have happened to **you personally** - **not** the other people in your household, in the .... months since the 1st January 1995. This is about **anything** that happened to you - at home, in the street, at work, in a shop, on public transport, or anywhere else.*

♣**Q40** *Apart from things you have mentioned already, since the beginning of 1995 has anyone stolen or tried to steal **anything you were carrying**, either out of your hands, or from your pocket or from a bag or case? How many times?*

Write in (2 digits) ____ ____
None=00

⌘**Q41** *(Apart from that) in that time has anyone stolen or tried to steal anything else that belonged **to you**? - such as from an office or anywhere else? How many times?*

Write in (2 digits) ____ ____
None=00

⊙**Q42** *And (apart from this) in that time has anyone tampered with or damaged any of your things on purpose? How many times?*

Write in (2 digits) ____ ____
None=00

♥**Q43** *And again (apart from any incidents you have mentioned already) since 1st January 1995 has any stranger or person you do not know well hit you, kicked you or used force or violence on you in any other way? How many times?* W43

Write in (2 digits) ____ ____
None=00

✪**Q44** *And (apart from this) has any stranger or person you do not know well **threatened** to use force or violence on you or **threatened** to damage things of yours in any way that actually frightened you?* W44

Write in (2 digits) ____ ____
None=00

Now, if you will just bear with me for a minute, I have to summarise your answers to the last few questions because they determine where we go from here.

Check back to questions Q33 to Q44. Code the total number of incidents for each question in Col 1 on opposite page. If none code '00'. (Note that Q37 is not listed here.)

If no incidents recorded at all in Col 1, Go to Demographics

CRIME TYPE			Col 1 Total No. of Incidents		Col 2 No. of Victim Forms to be completed	
● Q.33	Vehicle Theft	P10A33	□□	P10B33	□	
◆ Q.34	Theft from vehicle		□□		□	
■ Q.35	Damage to vehicle		□□		□	
⊙ Q.42	Other damage		□□		□	
⌘ Q.41	Other Theft Theft from property	√	□□	√	□	
✸ Q.39	Theft from inside home		□□		□	A37 ? B37
household ➢ Q.38	Theft from property		□□		□	
✧ Q.36	Burglary		□□		□	
♣ Q.40	Theft from person		□□		□	
personal ✪Q.44	Threats		□□		□	
♥ Q43	Assault	P10A43	□□	P10B43	□	

Blue Victim forms should be filled out for each incident recorded in Col. 1 above.

If total number of required victim forms is more than 4, work back from the bottom of the list above and complete victim forms for the four incidents you come to first. If this means choosing from more than one incident at a particular question, ask respondent to think about the most recent.

Record in Col 2 above for each crime type, the number of incidents where victim forms will be completed.

Before answering Victim Forms, go to Demographics.

Demographics

HHSIZE

D1 *Can I check some details about the members of your household? Firstly, how many people are there in your household, including you?*

Write in (2 digits) ____ ____

D2 *How many people aged 15 and over are there in your household, including yourself?*

HH15P

Write in (2 digits) ____ ____

D3 *How many children aged under 15 are there in your household ?*

HH15L

Write in (2 digits) ____ ____

Interviewer: check that D2 + D3 = D1

Showcard F

D4 *Which one of the statements on Page F best describes this household?*

One person living alone........... 01
Solo parent with child/children 02
Couple without children 03
Couple with children 04
Extended family/whanau 05
Flatmates................................. 06
Family - other combination 07
Other (specify)____________ 37

D5 Code Sex

Male............. 1
Female 2

D6 *Can you please tell me your age?*

Write in (2 digits) ____ ____

AGE

1 = 15-24
2 = 25-39
3 = 40-59
4 = 60+
5 = NS
7 = r.fused

Showcard G

D7 *Looking at Page G, can you please tell me which of these ethnic groups you belong to?*

Code all mentions below

D7A1 European.. 01
New Zealander of European descent.......... 02
New Zealander of Maori descent................ 03
Cook Island Maori..................................... 04
Samoan 05
Tongan.. 06
Niuean 07
Chinese... 08
Indian.. 09
D7A[illegible] Other (specify) ____________________ 37
3E

D8 *Does your household own this house/flat or rent it?*

Rented 1
D10 ← Owned (including with mortgage) 2
Other 7

D9 *Who does your household rent from?*

Read out

..Private owner.........................1
..Local authority/council..........2
..Housing New Zealand 3
..Other (specify).......................7

____

Showcard H

D10 *Looking at Page H, can you please tell me which of these groups best describes your personal work situation?*

Working in paid employment....................... 1
Home duties (not otherwise employed)........ 2
Retired/Superannuitant................................ 3
Social Welfare Beneficiary/Unemployed 4
Sick or disabled and unable to work............. 5
Unpaid work outside the home..................... 6
Student.. 7
Other (specify) ____________________ 8

D11 *What specifically is the occupation of the main income earner in this house? Please be as specific as possible. (eg Primary School Teacher, self-employed Building Contractor etc)*

If "retired" ask what was the specific occupation of the main income earner before retiring.

____

____

No main income earner.................................8

Showcard I

D12 *Looking at Page I, can you please tell me which of these best describes the main source of income for this house?*

Salary/wages...1
Social Welfare Benefit (unemployment,
disability, domestic purposes etc)...............2
Student Allowance......................................3
Superannuation..4
Other (Specify) ______________________ 7

Showcard J

D13 *Looking at Page J, can you tell me which one of these best describes your current situation?*

Legally Married 1
Defacto relationship.................2
Single/never married................3
Widowed..................................4
Divorced/separated 5

If male→ **D16**

If female continue ...

D14 *Are you presently living in a relationship with a man?*

D16 ← Yes............ 1 *
No............ 2

D15 *Have you been living in a relationship with a man at any time during the last 2 years?*

Yes.............1 *
No.............2

D.16 *Do you have a telephone in the household?*

Yes.... 1
No.... 2

D.17 **If yes:**

Can I write the number down please -It's possible that someone from AGB McNair wil ring you to verify this interview took place

STD □□

□□□□□□□

NOTE:
Now respondent should answer Victim Forms and then self completion questionnaire.

If no Victim forms give respondent self completion questionnaire and read the introduction below:

The next section is concerned with peoples' experience as victims of some other kinds of crime which we have not discussed yet, for example violence between people who know each other well, and sex offences. We have a questionnaire inside this envelope, which we would like you to fill in yourself. Then after you have filled it in, you should put it back in the envelope and seal it using this sticker and I will take it with me.

When envelope is opened: *you fill in the cream coloured form first then work out whether or how many green forms you should fill in based on your answers to the cream form. I can help you work this out when you get to it, if you like.*

This section is to be completed after respondent has sealed self-completion questionnaire in envelope

If male→ **Finish Time**

If female continue ...

D18 **Check back to D.14 and D15, If yes (code 1✱) coded at either continue, all others→** **Finish Time**

Thank you for taking part in our survey - we really appreciate your contribution. We may be conducting some more research on this topic in a few weeks time. Here is a sheet explaining what it is about **(hand over 'Women's Safety' sheet and allow respondent time to read)**. *Would you be prepared to take part?*

Q30

Yes...............1
Finish Time ← No.................2

Hand respondent 'Women's safety' consent form to sign

D19 *If you are selected to take part in the research, would you prefer to be interviewed by telephone or have an interviewer come to your home, or have the interviewer meet you somewhere else?*

Q31

Telephone 1
Home.....................2
Somewhere else 3

D20 *If you are selected someone will give you a ring to arrange a time. Which of the following times would be convenient for us to ring?*

Q32

Read out. Code all mentions

..Weekday - daytime.......1
..Weekday - evenings 2
..Weekend - daytime.......3
..Weekend - evenings 4

Finish time (after self completion) □□□□

Now record interview duration (minutes) □□□

Can I get your name please in case someone from AGB McNair rings you to check this interview has taken place?

Name: ____________________

Address: ____________________

____

I hereby certify that this interview carried out and recorded by me today is true and accurate, and in accordance with instructions.

Interviewer: ________________

Code:__________ **Date:**______________

Supervisor Check: ____________________

Date: ________________

Supervisor Audit: **Phone**	Yes...... 1 No...... 2
Checking card	Yes...... 1 No...... 2

Office use only

Edited By: ____________________

Punched By: ____________________

Verified By: ____________________

Internal Audit: Yes 1
No 2

3

SELF-COMPLETION QUESTIONNAIRE

CONFIDENTIAL

✷ We promise that your answers are **totally confidential** and will not be seen by the interviewer if you hand back this booklet sealed in the envelope provided.

✷ The person who opens the envelope (and thousands are being collected) will never know who you are and all the answers will be added together by computer.

✷ **Please answer honestly.** It is important that we have a complete picture of what happens to people.

✷ Please ignore the numbers next to the boxes. These are for office use.

The interviewer will now show you how to fill in the questionnaire using the example questions on the next page.

EXAMPLE

Ask your interviewer to demonstrate how you should fill in this questionnaire, by using this example page.

1 *Has anyone in your household* ***deliberately*** *hit or kicked your cat since the 1st January 1995?*

Yes.. □1 → **GO TO 2**
No.. □2 → **GO TO 3**

2 *How many times has this happened since 1st January 1995?*

Please write the number of times in the box □

3 *Has anyone in your household* ***threatened to*** *hit or kick your cat since the 1st January 1995?*

Yes.. □1
No.. □2

*Most of the questions that follow are about things that might have happened to you over the period since 1st January 1995, but some questions relate to earlier periods. We only want to know about things which have happened to **you personally**.*

We don't just want to know about serious things - we want to know about small things too.

If you are unsure how to fill this form in feel free to ask the interviewer to help you.

VIOLENCE BETWEEN PEOPLE WHO KNOW EACH OTHER WELL

1 *Has anyone you know well **deliberately** hit you, kicked you or used force or physical violence on you since 1st January 1995?*

B1

Yes.................................... □ 1 → **GO TO 2**
No..................................... □ 2 → **GO TO 3**

2 *How many times has this happened since 1st January 1995?*

B2 **Please write the number of times in the box** □

3 *Since 1st January 1995 has anyone you know well **threatened** to use force or violence on you, or **threatened** to damage things of yours, in a way that actually frightened you?*

B3

Yes.................................... □ 1 → **GO TO 4**
No..................................... □ 2 → **GO TO 5**

4 *How many times has this happened since 1st January 1995?*

B4 **Please write the number of times in the box** □

Answer Questions 5a) - 5f) if you have <u>ever</u> been in a marital or similar partnership, otherwise go straight to question 6.
These questions apply only to a partner of the opposite sex.

5a *Has any partner ever* ***deliberately*** *destroyed, damaged or harmed something belonging to you, or* ***threatened*** *to do any of these things, in a way that actually frightened you?*

B5a1

Yes.. □1
No... □2

5b *Has any partner ever* ***actually*** *used force or violence on you, such as deliberately hit, kicked, pushed, grabbed or shoved you, or deliberately hit you with something, in a way that could have hurt you?*

B5a2

Yes.. □1
No... □2

5c *Has any partner ever* ***threatened*** *to use force or violence on you, such as threatened to hit, kick, push, grab or shove you, in a way that actually frightened you?*

B5a3

Yes.. □1
No... □2

5d *Has any partner ever used a weapon against you, or threatened to use a weapon against you, such as a knife or a gun or any other weapon?*

B5a4

Yes.. □1
No... □2

5e *Has any partner ever made you carry out any sexual activity when you did not want to, by* ***holding you down or hurting you in some way****?*

B5a5

Yes.. □1
No... □2

5f *Has any partner ever made you carry out any sexual activity when you did not want to, by* ***threatening*** *you in some way?*

B5a6

Yes.. □1
No... □2

SEXUAL VICTIMISATION

6 *During your life, has anyone, either a stranger or anyone you know, ever done any of the following things to you?*

Please tick the boxes that apply to you

B6a1 Had sexual intercourse or attempted intercourse, against your will .... □ 1

B6a2 Had oral sex or attempted oral sex, against your will .... □ 2

B6a3 Had anal sex or attempted anal sex, against your will .... □ 3

B6a4 Penetrated or attempted to penetrate the vagina or anus with fingers or an object, against your will .... □ 4

B6a5 Sexually assaulted you in some other way .... □ 5

B6a8 None of these .... □ 8 → **GO TO 9**

7 *Have any of these things happened since 1st January 1995?*

B7

Yes .... □ 1 → **GO TO 8**
No .... □ 2 → **GO TO 9**

8 *How many times have each of the following things happened since 1st January 1995?*

Please write the number of times in the box

8a1 1. Sexual intercourse or attempted sexual intercourse, against your will............ □

8a2 2. Oral sex or attempted oral sex, against your will .... □

8a3 3. Anal sex or attempted anal sex, against your will... □

8a4 4. Penetration or attempted penetration of the vagina or anus with fingers or an object, against your will .... □

8a5 5. Other sexual assault.. □

9 *Apart from the things you have already mentioned, are there any other offences of these sorts (violence or sexual) which have happened to you since 1 January 1995. Please describe.*

____

____

B9 ____

____

____

NOW TURN TO THE BACK PAGE

Please now read these notes to work out how many green forms you should fill in, if any

1. Write in below, how many offences you have identified in Questions 2, 4, 8 & 9.

Question 2	
Question 4	
Question 8	
Question 9	

2. We would like you to fill in a green form for each of these offences, up to a maximum of four green forms.

3. If you have identified more than four offences, your four green forms should be for the four most recent offences.

4. If any two or more of these offences were committed by the same person **and** they were similar offences, you should fill in only one form for them, and it should be for the offence you think is the most serious of them.

If you are not sure what you should do now, please ask the interviewer.

A.3 SAS PROGRAM FOR CONTINGENCY TABLE ANALYSES

```
/* Analysis of the 1986 Community Questionnaire survey data. */

data dd1;
  filename in1 'c:\ganesh\students\jane\po.dat';
   infile in1;
    input

str clus wgt q8 q11a q11b q11c q11d q11e q11f q11g q11h q32c
q85a q85b q85c q85d q85cd q85e q85f q85g q86 q88 q90 q934 q99
q109 q115self ageg gend ethn idn
q8ind q11aind q32cind q85aind
q86ind q88ind q90ind q93ind q99ind q109ind      ;

/* Inputs the data and variable names. */

if q85a=6 then delete;
if q85f=6 then delete;

/* Deletes the missing observations. */

filename out1 'c:\ganesh\students\jane\po2.dat';
file out1;
put
str clus wgt q8 q11a q11b q11c q11d q11e q11f q11g q11h q32c
q85a q85b q85c q85d q85cd q85e q85f q85g q86 q88 q90 q934 q99
q109 q115self ageg gend ethn idn
q8ind q11aind q32cind q85aind
q86ind q88ind q90ind q93ind q99ind q109ind      ;
run;

/* Creates a new data set with all of the variables excluding the missing
values. */

proc freq ;
tables gend*q85f / chisq;
tables gend*q85a / chisq;
weight wgt;
run;

/* Produces both contingency tables with wgt as the weighting variable. */
```

A.4 PROGRAMS FOR REGRESSION

A.4.1 SAS PROGRAM

```
/* Analysis of the 1996 Victimisation survey data */
/* There are two forms of sample survey weighting, one for household, the
other for individuals. The weighting used depends on the variable. */

libname testlib V604 'c:\ganesh\students\jane';
data testlib.d1;
run;

/* Calling up the data set, d1, which already exists as a SAS data set */

proc reg data=testlib.d1;
 model hhoffi=hhsize d4 d8d9a d8d9b d8d9c d8d9d d11
w1 w6a1 w6a2 w6a3 w6a4 w6a5 w6a6 w6a7 w7;
  weight hhwgtpos; /* poststratified household weight */

run;

/* The regression model, modelling hhoffi (dependent variable) on all of the
other variables (independent variables). */
```

A.4.2 SUDAAN PROGRAM

```
/* Program using the same SAS data set */

proc regress data=/* Data set used*/

[FILETYPE=ASCII|SAS\SUDAAN and other options for the regression
procedure in SUDAAN]

/* nest denotes sample survey nesting
strata = stratum (area based)
au = Macnair area unit

weight hhwgtpos; /* poststratified household weight */
model hhoffi=hhsize d4 d8d9a d8d9b d8d9c d8d9d d8d9e d11 w1 w6a1 w6a2
w6a3 w6a4 w6a5 w6a6 w6a7 w7;
subgroup d4 d8d9a d8d9b d8d9c d8d9d d8d9e d11 w1 w6a1 w6a2 w6a3 w6a4
w6a5 w6a6 w6a7 w7; /* categorical variables */
levels 7 2 2 2 2 2 8 3 4 4 4 4 4 4 4 5; /* number of levels of each categorical
variable */

setenv linesize=80 colwidth=15 rowwidth=15 decwidth=5;
/* options in SUDAAN to control the output */
```

A.5 PROGRAMS FOR LOGISTIC REGRESSION

A.5.1 SAS PROGRAM MODELLING PREVALENCE OF BURGLARY

```
/* Analysis of the 1996 Victimisation survey data */

libname testlib V604 'c:\ganesh\students\jane';
data testlib.d2; set testlib.data27b;

d4_1=0; d4_2=0; d4_3=0; d4_4=0; d4_5=0; d4_6=0;
if d4=1 then d4_1=1;
else if d4=2 then d4_2=1;
else if d4=3 then d4_3=1;
else if d4=4 then d4_4=1;
else if d4=5 then d4_5=1;
else if d4=6 then d4_6=1;
if d4=36 then delete;
if d4=37 then delete;

/* Creates each level of each categorical variable into binary variables and
deletes the missing observations from the analyses */

d8_1=0; d8_2=0;
if d8=1 then d8_1=1;
else if d8=2 then d8_2=1;
if d8=7 then delete;

d9_1=0; d9_2=0;
if d9=1 then d9_1=1;
else if d9=2 then d9_2=1;
if d9=7 then delete;
if d9=8 then delete;

if d8=1 and d9=1 then d8d9a=1; else d8d9a=0;
if d8=1 and d9=2 then d8d9b=1; else d8d9b=0;
if d8=1 and d9=3 then d8d9c=1; else d8d9c=0;
if d8=2 and d9=. then d8d9d=1; else d8d9d=0;
if d8=3 and d9=. then d8d9e=1; else d8d9e=0;

d11_1=0; d11_2=0; d11_3=0; d11_4=0; d11_5=0; d11_6=0; d11_7=0;
if d11=1 then d11_1=1;
else if d11=2 then d11_2=1;
else if d11=3 then d11_3=1;
else if d11=4 then d11_4=1;
else if d11=5 then d11_5=1;
else if d11=6 then d11_6=1;
else if d11=7 then d11_7=1;
```

```
w1_1=0; w1_2=0;
if w1=1 then w1_1=1;
else if w1=2 then w1_2=1;
if w1=8 then delete;
if w1=9 then delete;

w6a1_1=0; w6a1_2=0; w6a1_3=0;
if w6a1=1 then w6a1_1=1;
else if w6a1=2 then w6a1_2=1;
else if w6a1=3 then w6a1_3=1;
if w6a1=8 then delete;
if w6a1=9 then delete;

w6a2_1=0; w6a2_2=0; w6a2_3=0;
if w6a2=1 then w6a2_1=1;
else if w6a2=2 then w6a2_2=1;
else if w6a2=3 then w6a2_3=1;
if w6a2=8 then delete;
if w6a2=9 then delete;

w6a3_1=0; w6a3_2=0; w6a3_3=0;
if w6a3=1 then w6a3_1=1;
else if w6a3=2 then w6a3_2=1;
else if w6a3=3 then w6a3_3=1;
if w6a3=8 then delete;
if w6a3=9 then delete;

w6a4_1=0; w6a4_2=0; w6a4_3=0;
if w6a4=1 then w6a4_1=1;
else if w6a4=2 then w6a4_2=1;
else if w6a4=3 then w6a4_3=1;
if w6a4=8 then delete;
if w6a4=9 then delete;

w6a5_1=0; w6a5_2=0; w6a5_3=0;
if w6a5=1 then w6a5_1=1;
else if w6a5=2 then w6a5_2=1;
else if w6a5=3 then w6a5_3=1;
if w6a5=8 then delete;
if w6a5=9 then delete;

w6a6_1=0; w6a6_2=0; w6a6_3=0;
if w6a6=1 then w6a6_1=1;
else if w6a6=2 then w6a6_2=1;
else if w6a6=3 then w6a6_3=1;
if w6a6=8 then delete;  if w6a6=9 then delete;
w6a7_1=0; w6a7_2=0; w6a7_3=0;
if w6a7=1 then w6a7_1=1;
else if w6a7=2 then w6a7_2=1;
else if w6a7=3 then w6a7_3=1;
if w6a7=8 then delete;
if w6a7=9 then delete;

w7_1=0; w7_2=0; w7_3=0; w7_4=0;
if w7=1 then w7_1=1;
else if w7=2 then w7_2=1;
else if w7=3 then w7_3=1;
else if w7=4 then w7_4=1;
if w7=8 then delete;
if w7=9 then delete;

run;

proc sort data=testlib.d2;
by strata descending burgp;
run;

/* Organises the data to model the prevalence of burglary and sorts the data
into ascending strata groups */

proc logistic data=testlib.d2 order=data;
 model burgp=hhsize d4_1-d4_6 d8d9a d8d9b d8d9c d8d9d d11_1-d11_7
w1_1 w1_2 w6a1_1-w6a1_3 w6a2_1-w6a2_3 w6a3_1-w6a3_3
w6a4_1-w6a4_3 w6a5_1-w6a5_3 w6a6_1-w6a6_3 w6a7_1-w6a7_3
w7_1-w7_4;
  weight hhwgtpos;
run;

/* A logistic regression analysis, hhsize is taken to be continuous. */
```

A.5.2 SUDAAN PROGRAM FOR MODELLING PREVALENCE OF BURGLARY

```
/* Analysis of the 1996 Victimisation survey data */

proc logistic data=/* data set used */

[FILETYPE= ... and other options in SUDAAN to specify the data]

model burgp=hhsize d4 d8d9a d8d9b d8d9c d8d9d d8d9e d11 w1 w6a1 w6a2
w6a3 w6a4 w6a5 w6a6 w6a7 w7;
subgroup d4 d8d9a d8d9b d8d9c d8d9d d8d9e d11 w1 w6a1 w6a2 w6a3 w6a4
w6a5 w6a6 w6a7 w7;
levels 7 2 2 2 2 2 8 3 4 4 4 4 4 4 4 5;
setenv  linesize=80 colwidth=15 rowwidth=15 decwidth=5;
```

A.5.3 SAS PROGRAM FOR MODELLING PREVALENCE OF VIOLENCE

```
/* Analysis of the 1996 Victimisation survey data */

libname testlib V604 'c:\ganesh\students\jane';

data testlib.d3; set testlib.data27b;
age_1=0; age_2=0; age_3=0; age_4=0; age_5=0;
if age=1 then age_1=1;
else if age=2 then age_2=1;
else if age=3 then age_3=1;
else if age=4 then age_4=1;
else if age=5 then age_5=1;

ethnic_1=0; ethnic_2=0; ethnic_3=0; ethnic_4=0;
if ethnic=1 then ethnic_1=1;
else if ethnic=2 then ethnic_2=1;
else if ethnic=3 then ethnic_3=1;
else if ethnic=4 then ethnic_4=1;
if ethnic=. then delete;

d11_1=0; d11_2=0; d11_3=0; d11_4=0; d11_5=0; d11_6=0; d11_7=0;
if d11=1 then d11_1=1;
else if d11=2 then d11_2=1;
else if d11=3 then d11_3=1;
else if d11=4 then d11_4=1;
else if d11=5 then d11_5=1;
else if d11=6 then d11_6=1;
else if d11=7 then d11_7=1;

d10_1=0; d10_2=0; d10_3=0; d10_4=0; d10_5=0; d10_6=0; d10_7=0;
if d10=1 then d10_1=1;
else if d10=2 then d10_2=1;
else if d10=3 then d10_3=1;
else if d10=4 then d10_4=1;
else if d10=5 then d10_5=1;
else if d10=6 then d10_6=1;
else if d10=7 then d10_7=1;
if d10=9 then delete;
if d10=10 then delete;

d13_1=0; d13_2=0; d13_3=0; d13_4=0;
if d13=1 then d13_1=1;
else if d13=2 then d13_2=1;
else if d13=3 then d13_3=1;
else if d13=4 then d13_4=1;
if d13=8 then delete;
if d13=9 then delete;
```

```
w7_1=0; w7_2=0; w7_3=0; w7_4=0;
if w7=1 then w7_1=1;
else if w7=2 then w7_2=1;
else if w7=3 then w7_3=1;
else if w7=4 then w7_4=1;
if w7=8 then delete;
if w7=9 then delete;

totali=griassi+oassi+abdkidi+threati;  totalp=0;
if totali>0 then totalp=1;  else if totali=0 then totalp=0;

/*  The variable totali is the total of the variables griassi, oassi, abdkidi and
threati.  totalp is the binary variable to be modelled which takes the value 1 if
totali is greater than zero, otherwise there are no incidences of violence and
hence takes a value of zero.  */

run;
proc sort data=testlib.d3;
by strata descending totalp;  /*  modelling the prevalence of violence  */
run;

proc logistic data=testlib.d3 order=data ;
 model totalp=d5 age_1-age_5 ethnic_1-ethnic_4 d11_1-d11_7 d10_1-d10_7
d13_1-d13_4 w7_1-w7_4;
  weight inwgtpos;  /*  individual poststratified weight  */
run;
```

A.5.4 SUDAAN PROGRAM FOR MODELLING PREVALENCE OF VIOLENCE

```
/*  Analysis of the 1996 Victimisation survey data  */

proc logistic data=/* data set used */

[FILETYPE= ... and other options in SUDAAN to specify the data]

weight  inwgtpos;  /*  individual poststratified weight  */

model totalp=d5 age ethnic d10 d11 d13 w7;
subgroup d5 age ethnic d10 d11 d13 w7;
levels 2 6 5 8 8 5 5;

setenv  linesize=80 colwidth=15 rowwidth=15 decwidth=5;
```

REFERENCES

Agresti, A. (1990). Categorical Data Analysis. New York: Wiley.

Anderson, S., Auquier, A., Hauck, W.W., Oakes, D., Vandaele, W. and Weisberg, H.I. (1980). Statistical Methods for Comparative Studies. New York: Wiley.

Bebbington, A.C. and Smith, T.M.F. (1977). The Effect of Survey Design on Multivariate Analysis. In: O'Muircheartaigh, C.A. and Payne, C.D. Model Fitting. New York: Wiley.

Bedrick, E.J. (1983). Adjusted Chi-Squared Tests for Cross-Classified Tables of Survey Data. Biometrika, 70, 591-595.

Binder, D., Kovar, J., Kumar, S., Paton, D. and van Baaren, A. (1987). Analytic Uses of Survey Data: A Review. In: MacNeill, I.B. and Umphrey, G.J. (Eds.) Applied Probability, Stochastic Processes, and Sampling Theory. Dordrecht: Reidel.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge: MIT Press.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. New York: Wiley.

Chaudhuri, A. and Stenger, H. (1992). Survey Sampling. New York: Marcel Dekker, Inc.

Chaudhuri, A. and Vos, J.W.E. (1988). Unified Theory and Strategies of Survey Sampling. Amsterdam: North-Holland.

Christensen, R. (1990). Log-Linear Models. New York: Springer-Verlag.

Cochran, W.G. (1963). Sampling Techniques, (2nd Ed). New York: Wiley.

Deming, W.E. (1950). Some Theory of Sampling. New York: Wiley.

Everitt, B.S. and Dunn, G. (1992). Applied Multivariate Data Analysis. New York: Oxford University Press.

Fellegi, I.P. (1980). Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples. Journal of the American Statistical Association, 75, 261-268.

Frankel, M.R. (1971). Inference from Survey Samples: An Empirical Investigation. Michigan: Institute for Social Research.

Fuller, W.A. (1975). Regression Analysis for Sample Surveys. Sankhya, C37, 117-132.

Fuller, W.A. (1987). Estimators of the Factor Model for Survey Data. In: MacNeill, I.B. and Umphrey, G.J. (Eds.) Applied Probability, Stochastic Processes and Sampling Theory. Dordrecht: Reidel.

Fuller, W.A. and Isaki, C.T. (1981). Survey Design Under Superpopulation Models. In: Krewski, D., Platek, R. and Rao, J.N.K. (Eds.) Current Topics in Survey Sampling. New York: Academic Press.

Fuller, W.A., Kennedy, W., Schnell, D., Sullivan, G. and Park, H.J. (1986). PC CARP. Ames, Iowa: Statistical Laboratory, Iowa State University.

Ganesalingam, S. (1993). Understanding Multivariate Statistical Methods. Palmerston North: Massey University.

Graybill, F.A. and Iyer, H.K. (1994). Regression Analysis. Belmont: Wadsworth.

Hansen, M.M., Hurwitz, W.N. and Madow, W.G. (1953). Sample Survey Methods and Theory, (in two volumes). New York: Wiley.

Haslett, S.J. (1985). The Linear Non-Homogeneous Estimator in Sample Surveys. Sankhya, B47, 101-117.

Haslett, S.J. (1990). Analysis of the Wellington Community Survey: Porirua and Lower Hutt. Technical Report, Number 2,3. Wellington: Institute of Statistics and Operations Research, Victoria University.

Hastie, T.J. and Pregibon, D. (1992). Generalized Linear Models. In: Chambers, J.M. and Hastie, T.J. (Eds.) Statistical Models in S. California: Wadsworth and Brooks/Cole.

Hidiroglou, M.A. and Paton, D.G. (1987). Some Experiences in Computing Estimates and Their Variances Using Data from Complex Survey Designs. In: MacNeill, I.B. and Umphrey, G.J. (Eds.) Applied Probability, Stochastic Processes and Sampling Theory. Dordrecht: Reidel.

Holt, D. and Scott, A.J. (1981). Regression Analysis Using Survey Data. The Statistician, 30, 169-178.

Holt, D., Scott, A.J. and Ewings, P.D. (1980). Chi-Squared Tests with Survey Data. Journal of the Royal Statistical Society, A143, 303-320.

Holt, D., Smith, T.M.F. and Winter, P.D. (1980). Regression Analysis of Data from Complex Surveys. Journal of the Royal Statistical Society, A143, 474-487

Hyman, H. (1955). Survey Design and Analysis. New York: The Free Press.

Johnson, R.A., and Wichern, D.W. (1992). Applied Multivariate Statistical Analysis, (3rd Edition). New Jersey: Prentice Hall.

Kish, L. (1965). Survey Sampling. New York: Wiley.

Kish, L. and Frankel, M.R. (1974). Inference from Complex Samples (with discussion). Journal of the Royal Statistical Society, B36, 1-37.

Kohler, H. (1985). Statistics for Business and Economics. Illinois: Scott, Foresman and Company.

Larsen, R.J. and Marx, M.L. (1986). An Introduction to Mathematical Statistics and it's applications, (2nd Ed). New Jersey: Prentice-Hall.

Lehtonen, R. and Pahkinen, E.J. (1995). Practical Methods for Design and Analysis of Complex Surveys. Chichester: Wiley.

McCullagh, P. and Nelder, J.A. (1983). Generalized Linear Models. New York: Chapman and Hall.

MacNeill, I.B. and Umphrey, G.J. (Eds.) (1987). Biostatistics. Dordrecht, Holland: Reidel.

Manly, B.F.J. (1986). Multivariate Statistical Methods. London: Chapman and Hall.

Miller, A.J. (1990). Subset Selection in Regression. New York: Chapman and Hall.

Minitab Inc. (1996). MINITAB User's Guide. State College, Pennsylvania: Minitab Inc.

Myers, R.H. (1990). Classical and Modern Regression With Applications, (2nd Ed). Boston: PWS-Kent.

Nathan, G. and Holt, D. (1980). The Effect of Survey Design on Regression Analysis. Journal of the Royal Statistical Society, B42, 377-386.

Nathan, G. and Smith, T.M.F. (1989). The Effect of Selection on Regression Analysis. In: Skinner, C.J., Holt, D. and Smith, T.M.F. (Eds.) Analysis of Complex Surveys. Chichester: Wiley.

O'Muircheartaigh, C.A. and Payne, C.D. (Eds.) (1977). Exploring Data Structures. New York: Wiley.

Payne, C. (1977). The Log-Linear Model for Contingency Tables. In: O'Muircheartaigh, C.A. and Payne, C.D. Model Fitting. New York: Wiley.

Pfeffmann, D. and Nathan, G. (1981). Regression Analysis of Data from Complex Samples. Journal of the American Statistical Association, 76, 681-689.

Rao, C.R. and Toutenburg, H. (1995). Linear Models: Least Squares and Alternatives. New York: Springer-Verlag.

Rao, J.N.K. and Scott, A.J. (1981). The Analysis of Categorical data from Complex Sample Surveys: χ^2 tests for Goodness-of-Fit and Independence in Two Way Tables. Journal of the American Statistical Association, 76, 221-230.

Rao, J.N.K. and Thomas, D.R. (1989). Chi-squared Tests for Contingency Tables. In: Skinner, C.J., Holt, D. and Smith, T.M.F. (Eds.) Analysis of Complex Surveys. Chichester: Wiley.

Read, T.R.C. and Cressie, N.A.C. (1988). Goodness-Of-Fit Statistics for Discrete Multivariate Data. New York: Springer-Verlag.

Roberts, G., Rao, J.N.K. and Kumar, S. (1987). Logistic Regression Analysis of Sample Survey Data. Biometrika, 74, 1-12.

Särndal, C.E., Swensson, B. and Wretman, J.H. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

SAS Institute Inc. (1989). SAS Language and Procedures: Usage, Version 6, (1st edition). Cary, NC: SAS Institute.

Satterthwaite, F.E. (1946). An Approximate Distribution of Estimates of Variance Components. Biometrics, 2, 110-114.

Scott, A.J. and Holt, D. (1982). The Effect of Two Stage Sampling on Ordinary Least Squares Methods. Journal of the American Statistical Association, 77, 848-854.

Scott, A.J. and Rao, J.N.K. (1981). Chi-Squared Tests for Contingency Tables with Proportions Estimated from Survey Data. In: Krewski, D., Platek, R. and Rao, J.N.K. (Eds.) Current Topics in Survey Sampling. New York: Academic Press.

Seber, G.A.F. (1977). Linear Regression Analysis. New York: Wiley.

Shah, B.V., Barnwell, B.G. and Bieler, G.S. (1995). SUDAAN: User's Manual. Research Triangle Park, NC: Research Triangle Institute.

Skinner, C.J. (1986). Regression Estimation and Post-Stratification in Factor Analysis. Psychometrika, 51, 347-356.

Skinner, C.J. (1988). On Conditioning for Model-based Inference in Survey Sampling. Biometrika, 75, 275-286.

Skinner, C.J., Holmes, D.J. and Smith, T.M.F. (1986). The Effect of Sample Design on Principal Component Analysis. Journal of the American Statistical Association, 81, 789-798.

Skinner, C.J., Holt, D. and Smith, T.M.F. (Eds.) (1989). Analysis of Complex Surveys. Chichester: Wiley.

Smith, T.M.F. (1981). Regression Analysis for Complex Surveys. In: Krewski, D., Platek, R. and Rao, J.N.K. (Eds.) Current Topics in Survey Sampling. New York: Academic Press.

Smith, T.M.F. (1984). Present Position and Potential Developments: Some Personal Views (with discussion). Journal of the Royal Statistical Society, A147, 208-221.

Smith, T.M.F. (1989). Introduction to Part B. In: Skinner, C.J., Holt, D. and Smith T.M.F. (Eds.) Analysis of Complex Surveys. Chichester: Wiley.

Smith, T.M.F. and Holmes, D.J. (1989). Multivariate Analysis. In: Skinner, C.J., Holt, D. and Smith T.M.F. (Eds.) Analysis of Complex Surveys. Chichester: Wiley.

Thomas, D.R. and Rao, J.N.K. (1987). Small-sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics under Cluster Sampling. Journal of the American Statistical Association, 82, 630-636.

Thompson, S.K. (1992). Sampling. New York: Wiley.

Thurstone, L.L. (1945). The Effects of Selection in Factor Analysis. Psychometrika, 10, 165-198.

Woodward, J.A., Bonett, D.G. and Brecht, M-L. (1990). Introduction to Linear Models and Experimental Design. Florida: Harcourt Brace Jovanovich.

Young, W., Morris, A., Cameron, N. and Haslett, S. (1997). New Zealand National Survey of Crime Victims, 1996. Wellington: Victimisation Survey Committee.