

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

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

	PAGE
ACKNOWLEDGEMENTS	ii
ABSTRACT	iii
CONTENTS	v
LIST OF FIGURES AND TABLES	ix
PREFACE	x
CHAPTER ONE	
Introduction	1
1.1 Survey Sampling	1
1.2 Sampling Designs	4
1.2.1 Simple random sampling	4
1.2.2 Stratified random sampling	6
1.2.3 Cluster sampling	8
1.2.4 Systematic sampling	10
1.2.5 Multistage sampling	11
1.3 Confidence Intervals	12
1.4 Some Basic Concepts	13
1.5 Design Effects	14
1.5.1 Multivariate design effects	16
1.6 The Problem In The Analysis Of Survey Data	17
1.6.1 Design and model based approaches	19
1.6.2 The superpopulation	19
1.6.3 Design and model expectations	20
CHAPTER TWO	
Survey Data	22
2.1 1986 Wellington Community Questionnaire	23
2.2 1996 New Zealand National Survey of Crime Victims	24
2.2.1 Coding	25

CHAPTER THREE

Regression	29
3.1 The Regression Model	30
3.1.1 The simple linear model	31
3.1.2 Parameter estimation	33
3.2 Multiple Linear Regression	36
3.2.1 Parameter estimation	37
3.3 Weighted Least Squares	38
3.4 Subset Selection	39
3.4.1 Mallows C_p statistic	41
3.5 Non-Linear Regression	42
3.5.1 A special class of non-linear models	43
3.6 Regression Analysis On Complex Surveys	44
3.6.1 Estimation of β	45
3.7 The Age Of Computers	52
3.8 Summary	54

CHAPTER FOUR

Contingency Tables	55
4.1 One Way Tables	56
4.1.1 Goodness-of-fit tests	57
4.2 Two Way Tables	59
4.2.1 Independence test	59
4.2.2 Homogeneity test	60
4.3 Multiway Tables	61
4.4 Generalized Linear Models	62
4.4.1 Log-linear models	63
4.4.2 Logistic regression	66
4.5 Contingency Tables On Complex Surveys	68
4.5.1 Wald tests	71
4.5.2 Log-linear test statistics	73
4.5.3 Logistic regression test statistics	76
4.6 Summary	78

CHAPTER FIVE

Multivariate Analysis	80
5.1 The Multivariate Normal Distribution	82
5.1.1 Likelihood function	83
5.1.2 Maximum likelihood estimators of μ and Σ	84
5.2 Methods Of Multivariate Analysis	85
5.2.1 Principal components	85
Derivation of principal components	86
Graphical representation of principal components	87
5.2.2 Factor analysis	89
5.3 Multivariate Analysis On Complex Surveys	91
5.3.1 Principal components	96
5.3.2 Factor analysis	100
5.4 Summary	102

CHAPTER SIX

6.1 Computing	103
6.1.1 SAS	104
6.1.2 PC CARP	105
6.1.3 SUDAAN	106
6.1.4 Other packages	108
6.2 Application To 'Real' Data	109
6.2.1 Example one: Contingency tables	109
6.2.2 Example two: Regression	111
6.2.3 Example three: Logistic regression	113
6.2.4 General Conclusions	121
6.3 Summary	123

CHAPTER SEVEN

Summary And Conclusions	125
--------------------------------	------------

APPENDIX

A.1	1986 Wellington Community Questionnaire	130
A.2	1996 New Zealand National Survey of Crime Victims	177
A.3	SAS Program For Contingency Table Analyses	199
A.4	Programs For Regression	200
A.4.1	SAS program	200
A.4.2	SUDAAN program	200
A.5	Programs For Logistic Regression	201
A.5.1	SAS program modelling prevalence of burglary	201
A.5.2	SUDAAN program modelling prevalence of burglary	203
A.5.3	SAS program modelling prevalence of violence	204
A.5.4	SUDAAN program modelling prevalence of violence	205

REFERENCES

206

LIST OF FIGURES AND TABLES

	PAGE
Table 2.1. Codes used.	26-28
Figure 3.1. Scatter plot of the mean number of individual offences and age (in years)	31
Figure 3.2. Regression line fitted onto the scatter plot of the mean number of individual offences and age (in years)	32
Figure 5.1. Plot of mean incidence of individual offences and age.	87
Figure 5.2. Plot showing the first Principal Component, Y_1 .	88
Figure 5.3. Plot showing both Principal Components, Y_1 and Y_2 .	88
Table 5.1. Conditional expectations of the covariance estimators with respect to the superpopulation model.	95
Table 6.1. Two way table of Gender by 'Reported a crime to police'.	110
Table 6.2. Two way table of Gender by 'Asked police for directions'.	111
Table 6.3. Linear regression estimates from SAS and SUDAAN.	113
Table 6.4. Logistic regression estimates from SAS, modelling burglary.	115
Table 6.5. Logistic regression estimates from SUDAAN, modelling burglary.	116-117
Table 6.6. Logistic regression estimates from SAS, modelling violence.	118
Table 6.7. Logistic regression estimates from SUDAAN, modelling violence.	119

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.