Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data for Social and Economic Research

DESIGN-BASED AND MODEL-BASED METHODS IN ANALYZING COMPLEX HEALTH SURVEY DATA: A CASE STUDY

Risto Lehtonen[1,2], Kari Djerf [2], Tommi Härkänen[3] and Johanna Laiho[2]

ABSTRACT

In this paper, we discuss the analysis of complex health survey data by using multivariate modeling techniques. Our main interests are in design-based and model-based methods that aim at accounting for clustering, stratification and weighting effects. Our main interests are in clustering effects. Methods considered include generalized linear modeling with on pseudo-likelihood and generalized estimating equations, linear mixed models estimated by restricted maximum likelihood, and hierarchical Bayes techniques using Markov Chain Monte Carlo (MCMC) methods. The methods will be compared empirically, using data from a health interview and examination survey conducted in Finland in 2000 (Health-2000 Study). The data of the Health-2000 Study were collected using personal interviews, questionnaires and clinical examinations. A stratified two-stage cluster sampling design was used. The sampling design involved positive intra-cluster correlation for several study variables. We selected the study variables systolic blood pressure and chronic morbidity for a closer investigation. In many cases, the different methods produced similar numerical results and supported similar statistical conclusions. Methods that failed to account for the design complexities sometimes led to conflicting conclusions. We also discuss the application of the methods by using standard statistical software packages.

KEY WORDS: Multivariate survey analysis; Complex sampling design; Generalized estimating equations; Mixed models; Bayesian techniques.

1. INTRODUCTION

Various sources of complexity are often encountered in the survey design of a health interview and examination study. The sampling design can involve multi-stage sampling, possibly causing positive intra-cluster correlation for a given study variable. Varying inclusion probabilities may be used, and re-weighting may be needed to adjust for non-ignorable unit non-response. Item non-response may be treated by imputation techniques. To obtain reliable results, the analyst should make efforts to account for the various complexities in the analysis phase. Statistical tools implemented in standard statistical software packages can be used for this purpose. Problems may however arise in choosing appropriate methods and tools for a given analysis setting. The aim of this paper is to discuss the available methods, based on our experiences in the context of multivariate analysis of complex health survey data.

We consider design-based and model-based methods that aim at accounting for clustering, stratification and weighting effects. Our main concern is in accounting for the clustering effects. Methods include generalized linear modeling with pseudo-likelihood and generalized estimating equations, linear mixed models estimated by restricted maximum likelihood (REML), and hierarchical Bayes techniques using Markov Chain Monte Carlo (MCMC) methods. We compare empirically the results obtained by these methods using data from an extensive health interview and examination survey conducted in Finland in 2000 (Health-2000 Study). The survey design of the Health-2000 Study, and the study variables selected for this investigation, are summarized in Section 2. In Section 3 we outline the modeling methods used. Results are presented in Section 4 and discussion is in Section 5.

1 University of Jyväskylä
2 Statistics Finland
3 National Public Health Institute

2. SURVEY DATA

2.1 Sampling Design and Data Collection

The Health-2000 Study (Aromaa and Koskinen 2002) was carried out in the year 2000 by a consortium lead by the Public Health Institute, Finland. The target population of the main phase of the study covered the resident population in Finland aged 30 years or over. The size of the target population was 3.3 million persons. The data were collected from a sample of persons using face-to-face interviews, carried out at respondents' homes, and clinical examinations in local Health Centers and by some self-administered questionnaires. Due to clinical examinations the sampling design had to include regional clustering. A stratified two-stage sampling design was used with local Health Center Districts (comprising one or several municipalities) as the first-stage sampling units (i.e. regional clusters). There were a total of 249 regional clusters in the population. A total of 15 certainty strata (the 15 largest towns) were first formed as clusters with probability of one. The remaining 234 clusters were then divided into five regional strata, covering the whole (mainland) Finland. A total of 65 clusters were drawn from these strata by systematic PPS sampling with inclusion probabilities proportional to the size of the target population in a cluster. Thus, the total number of strata and first-stage sample clusters was 20 and 80, respectively.

The second-stage sample (about 8,000 people aged 30 years or over) was allocated proportionally to the strata. People aged 80 or over were over-sampled with a double inclusion probability relative to the younger age groups. Finally, individual persons were selected from each stratum with systematic sampling from an implicitly stratified frame register. About 88% of the sample persons were interviewed, 80% attended a comprehensive health examination and 5% attended a condensed examination at home. The most essential information on health and functional capacity was obtained from 93% of the subjects.

We used a data set with minimum editing and imputation and other confounding factors. The data set constructed for this study consists of 5,954 observations. There were a total of 2,495 PSU:s in our setting, out of which 65 are in the five two-stage strata, and the rest 2,430 in the 15 certainty strata (where each individual constituted a "cluster"). Thus, there were 2,475 degrees of freedom available for design-based variance estimation in most cases considered. For weight construction we also made a simplified assumption of ignorable unit non-response. The analysis weights were rescaled such that the mean of the weights was equal to one. The coefficient of variation of the weights was quite small taking the over-sampling into account, 15 per cent. Sampling design, first-stage sampling, conducting health interviews, and the development of weighting procedures were on the responsibility of Statistics Finland. The second-stage sample of persons was drawn by the Social Insurance Institution of Finland. National Public Health Institute was responsible of the collection of clinical data and the overall co-ordination of the study. Note that results presented in this paper may differ from other publications of the Health 2000 Study.

2.2 Design effects

The variables selected from the health examination phase were the continuous study variables systolic blood pressure and diastolic blood pressure, and from the health interview phase the binary study variables chronic illness (1 yes, 0 otherwise) and self-perceived health status good or very good (1 yes, 0 otherwise). In addition, a count variable selected from the health interview phase was the number of physician visits due to illness during the preceding year.
The study variables appeared to be positively intra-cluster correlated (Table 1). In the table, the first design effect estimate accounts for all the design complexities (weighting, stratification, clustering), and the second deff estimate accounts for clustering and stratification. Design effect estimates were largest for means of the continuous study variables. A design effect estimate close to one was obtained for the proportion estimator of the binary variable indicating self-perceived health status. The contribution of weighting to design effect estimates was small.

We selected two study variables, systolic blood pressure and chronic illness, for our final analyses. The motivation for this selection was that both data collection modes, health interview and health examination, were covered. The variables are of different types (continuous, binary) and suggest the application of logistic models in addition to linear models. Furthermore, both variables indicated a relatively strong clustering effect.

Table 1. Design effect estimates of means or proportions of selected study variables.

Variable	Mean or proportion	Overall design effect	Design effect due to clustering and stratification
Diastolic blood pressure	80.87	2.86	2.79
Systolic blood pressure	133.14	2.41	2.36
Chronic illness	0.52	1.75	1.71
Visiting a physician due to illness during the preceding year	4.50	1.45	1.42
Self-perceived health status good or very good	0.62	1.23	1.20

3. MULTIVARIATE MODELLING METHODS

3.1 Predictors

Models were fitted involving main effects and interactions of the available predictors. Our substance matter interests are in predictors relating to physical conditions and socio-economic characteristics. Sex-age adjusted figures were calculated in most cases. Our initial set of predictor variables was the following. Demographic variables: Age (in years) and gender (1 males, 2 females). Socio-economic variable: Education (in years spent for completed education). Variable related to physical conditions: Circum waist (in centimeters). For illustrative purposes we also constructed new variables by grouping the observations into three nearly equal-sized groups (three-parts) according to the values of a given predictor. This was applied for the variables education and circum waist.

3.2 Multivariate Methods

In generalized linear modeling under the so-called nuisance or aggregated approach (Skinner et al. 1989; Lehtonen and Pahkinen 1996), design-consistent estimation and asymptotically valid testing can be obtained. In this approach "pseudo" likelihood (PML) and related estimation techniques, "sandwich" type empirical standard error estimators and design-based Wald test statistics are used. Techniques based on generalized estimating equations (GEE) with multivariate quasi-likelihood (QML) estimation are used as a more advanced alternative (Liang and Zeger 1986; Diggle, Liang and Zeger 1994). In GEE, an assumption of an exchangeable correlation structure of observations in a cluster is applied. In these methods, the main inferential interests are in model coefficients; intra-cluster correlations as such are not necessarily of scientific interest. Multilevel or mixed models (Goldstein 1995; McCullogh and Searle 2001) involving fixed and random effects are used as another alternative. Cluster-specific random effects are incorporated in a model in order to account for the clustering effects. In this approach, the intra-cluster correlations and cluster-level inferences also are often of scientific interest. To complete our exercise, we included hierarchical Bayes techniques for an application to a simple random effects model (Gelman et al. 1995). We compare the methods empirically, by calculating point estimates and their estimated standard errors and design effects, and t-test and similar statistics.

3.3 Analysis Options

To manage a comparison of the selected methods, we formulated a set of analysis options to be referred to in our multivariate analysis exercise (Table 2). With Options 1 and 2, based on a fixed-effects model, we use the generalized estimating equations (GEE) approach. The GEE method with an independent correlation structure (Option 1) relates to the standard PML method where observations are assumed independent within clusters for the estimation of the regression coefficients but are allowed intra-cluster correlated in the estimation of the covariance matrix of the estimated regression coefficients (using a "sandwich", or robust, or empirical, variance estimator; e.g. Lehtonen and Pahkinen 1996 p. 271). In the GEE method assuming an exchangeable correlation structure (Option 2), observations are allowed intra-cluster correlated in the estimation of both the regression coefficients and the covariance matrix of the estimated regression coefficients, and QML estimation is used. With GEE estimation, we

also wanted to examine whether we will end up with closely comparable estimation and testing results with "design-based" software (such as SUDAAN) and software with more a "model-based" orientation (such as SAS). We assumed that this comparison might have some practical relevance for the users of the Health-2000 database. The GEE methods are available in all SUDAAN modeling procedures (Web reference 1) and in the SAS procedure GENMOD (Web reference 2). SUDAAN and SAS also allow the incorporation of element weights in the analysis.

Option 3 uses a mixed model formulation involving both fixed and random effects. For simplicity, we adopted a variance components model where cluster-specific random intercepts are included in the model in addition to the fixed effects. REML (residual, or restricted, ML) estimation was used for the estimation of the variance components. A "sandwich" type robust standard error estimator was again used for the estimated fixed effects. Estimation of the random effects model under Option 4 was carried out by MCMC techniques. We also incorporated element weights in our modeling exercises in Options 3 and 4. These analyses can be carried out with standard statistical software products. We used the SAS procedure MIXED for Option 3 and WinBUGS (Web reference 3) for Option 4 involving hierarchical Bayes analysis. We postulated a linear model for the continuous response variable and a logistic model for the binary response variable.

Table 2. Analysis options used in multivariate modeling exercise.

Option	Model formulation and estimation method	Aiming at accounting for…		
		Weighting	Stratification	Clustering
Option 0	Reference option Fixed-effects model, ML, model-based SE method	No	No	No
Option 1	Fixed-effects model, PML, robust SE method	Yes	Yes	Yes
Option 2 a) and b)	Fixed-effects model, QML, robust SE method a) SUDAAN application, b) SAS application	Yes	a) Yes b) No	Yes
Option 3	Mixed model, REML, robust SE method	Yes	No	Yes
Option 4	Bayesian random effects model, MCMC	Yes	Yes	Yes

Abbreviations:
SE: Standard error
ML: Maximum likelihood estimation
PML: Pseudo maximum likelihood estimation
QML: Quasi maximum likelihood estimation
REML: Residual maximum likelihood estimation
MCMC: Markov Chain Monte Carlo

In Reference Option, serving as a baseline method, a standard fixed-effects model was postulated and simple random sampling with replacement was assumed. This option ignores all the sampling complexities, corresponding to an analysis with a conventional OLS or ML estimation of the regression coefficients and model-based standard error estimators. By definition, design effect statistics for this option are equal to one. The analysis can be carried out by any standard statistical software package (such as SAS and SPSS).

To weight or not to weight in a complex analytical survey? There seems not to be any unique solution to this important theoretical and practical problem (see for example Pfeffermann et al. 1998 with discussion). We however took a position close to design-based reasoning and incorporated the weights (scaled to mean one) in all options except Reference Option. We also motivated this choice by design consistency reasons and protection against possible model failure. Cost to be paid was somewhat decreased efficiency. However, the effect was small.

Fitting regression models by GEE methods is discussed for example in Horton and Lipsitz (1999) who give a review of software for GEE estimation, covering options offered by SAS, Stata, SUDAAN and S-Plus software products, and Ziegler et al. (1998) who present an extensive literature review on GEE methodology. Multilevel modeling using SAS procedure MIXED is discussed e.g. in Singer (1998).

3.4 Models

A linear mixed model specification can be written compactly as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zv} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{y}$ denotes the $n\times 1$ vector of response variable measurements, $\mathbf{X}$ is the $n\times p$ design matrix for the fixed part of the model, $\boldsymbol{\beta}$ is the corresponding $p\times 1$ fixed parameters vector, $\mathbf{Z}$ is the $n\times q$ design matrix for the random part of the model, $\mathbf{v}$ is the corresponding $q\times 1$ vector of random effects, and $\boldsymbol{\varepsilon}$ is the $n\times 1$ residual vector. We make the standard assumptions that $\mathbf{v}$ and $\boldsymbol{\varepsilon}$ follow normal distributions with

$$\mathrm{E}\begin{bmatrix}\mathbf{v}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix} \text{ and } \mathrm{Var}\begin{bmatrix}\mathbf{v}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & 0\\ 0 & \mathbf{R}\end{bmatrix}$$

where $\mathbf{G}$ is the $q\times q$ variance-covariance matrix of $\mathbf{v}$ and $\mathbf{R}$ is that for $\boldsymbol{\varepsilon}$. The variance of $\mathbf{y}$ thus is $\mathrm{Var}(\mathbf{y}) = \mathbf{ZGZ}^T + \mathbf{R}$. Now, $\mathrm{Var}(\mathbf{y})$ is modelled by setting up the random-effects design matrix $\mathbf{Z}$ and by specifying covariance structures for $\mathbf{G}$ and $\mathbf{R}$. A simple random effects model is a special case of the general specification with $\mathbf{Z}$ containing dummy variables, $\mathbf{G}$ containing variance components in a diagonal structure, and $\mathbf{R} = \sigma^2\mathbf{I}_n$, where $\mathbf{I}_n$ denotes the $n\times n$ identity matrix. The general linear model $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ is a further special case with $\mathbf{Z} = 0$ and $\mathbf{R} = \sigma^2\mathbf{I}_n$. More details can be found for example in McCullogh and Searle (2001).

In model (1) we are especially interested in the beta parameters of the fixed part of the model. We therefore insert our subject matter predictors in the fixed part design matrix of model (1). Design matrix $\mathbf{Z}$ in the random part of (1) allows the modeling of the clustering effects. Random intercepts will constitute the random part parameters in our simple modeling exercise. Thus, the model (1) simplifies to

$$y_{hik} = \mathbf{x}_{hik}^T\boldsymbol{\beta} + v_{hi} + \varepsilon_{hik}, \tag{2}$$

where $\mathbf{x}_{hik}$ are the x-vectors and *h* refers to strata, *i* refers to clusters within strata, and *k* refers to elements within clusters. In our setting the cluster-specific random effects v_{hi} and residuals ε_{hik} are assumed mutually independent and follow normal distributions with zero means and variances of σ_v^2 and σ_e^2, respectively. For systolic blood pressure (continuous response) we use linear fixed-effects models and linear mixed models. For chronic illness (binary response) we use fixed effects binomial logistic models.

For GEE estimation consider a generalized linear model of the form

$$E_m(F(\mathbf{y})) = \mathbf{X\beta} \tag{3}$$

where *F* refers to the link function. We specify a linear link function for our continuous response variables and a logistic link function for the binary response variables. In GEE we model the covariance structure of observations within clusters by $\mathbf{V}_{hi} = \varphi\mathbf{A}_{hi}^{1/2}\mathbf{R}(\alpha)\mathbf{A}_{hi}^{1/2}$ where $\mathbf{A}_{hi}$ is a diagonal matrix of variance functions, α refers to the "working" correlation of pairs of observations in a cluster, and φ is the scale parameter. For independent correlation structure, α is set to zero, and the "working" correlation matrix $\mathbf{R}(\alpha)$ reduces to an identity matrix. For exchangeable correlation structure, α denotes the off-diagonal elements of $\mathbf{R}(\alpha)$. More details can be found in Liang and Zeger (1986) and Diggle et al. (1994).

Estimation of parameters of models (1) to (3) is carried out as follows. Random effects of (1) are estimated by REML and fixed effects by GLS, respectively, by using the SAS procedure MIXED. Weights can be incorporated in the estimation procedure. An empirical (robust) covariance matrix estimate of the estimated model parameters can be

obtained. GEE models are estimated by multivariate quasi-likelihood (McCullagh and Nelder 1989, Chapter 9) using the SAS procedure GENMOD. In GENMOD, a weight variable can be used as the scale parameter weight such that the scale parameter is divided by the weight variable value for each observation (thus we actually work with weighted GEE:s). Both model-based and empirical (robust) covariance matrix estimates of the estimated model parameters can be obtained. SUDAAN uses Horvitz-Thompson type weighted normal equations for fixed-effects linear models and weighted likelihood equations for non-linear models. In GEE models SUDAAN uses a modified Newton-Raphson algorithm for parameter estimation and calculates the empirical (robust) covariance matrix estimate by implicit Taylor linearization method (Binder 1983). Wald and related test statistics can be calculated by all the procedures (see e.g. Lehtonen and Pahkinen 1996). In SUDAAN we used the procedures REGRESS and LOGISTIC.

Bayesian Models

The full Bayesian approach (Gelman et al. 1995) is applied. The prior distributions are chosen to be vague because no prior knowledge is assumed. The model is a random effects model in which the stratification of 20 strata indexed by *h* needs to be taken into account by their own random effects terms u_h. Also, each cluster *i* has its own random effects term v_{hi} shared by the subjects drawn from that cluster. A linear regression model for a continuous response y_{hik} and covariates $\mathbf{x}_{hik}$ can now be expressed by

$$y_{hik} = \mathbf{x}_{hik}^T \boldsymbol{\beta} + u_h + v_{hi} + \varepsilon_{hik} \tag{4}$$

where ε_{hik} are the individual error terms of the subjects *k*. Random effects u_h and v_{hi}, and error terms ε_{hik} are assumed to be independent and normally distributed with zero means and variances σ_u^2, $\sigma_v{}^2$ and $\sigma_\varepsilon{}^2$, respectively. In the 15 self-representing strata the clusters contain only a single element, in which case the random effects v_{hi} can be omitted by restructuring the likelihood terms

$$p(y_{hik} \mid \mathbf{x}_{hik}, \boldsymbol{\theta}) = \Phi(y_{hik} \mid \mathbf{x}_{hik}^T \boldsymbol{\beta} + u_h + v_{hi}, \sigma_\varepsilon^2)$$

into

$$p(y_{hik} \mid \mathbf{x}_{hik}, \boldsymbol{\theta}) = \Phi(y_{hik} \mid \mathbf{x}_{hik}^T \boldsymbol{\beta} + u_h, \sigma_v^2 + \sigma_\varepsilon^2),$$

where Φ corresponds to the normal density function of the response given the mean and the variance. The benefit of this approach is that the number of random effects in the model remains small, and the estimation by MCMC methods is feasible.

A similar construction for weighted binary responses ranging in {0,1} can be applied by considering

$$\mathrm{F}(\Pr[y_{hik} = 1 | \mathbf{x}_{hik}, \theta]) = \mathbf{x}_{hik}^T \boldsymbol{\beta} + u_h + v_{hi}, \tag{5}$$

where the link function *F* is chosen to be logistic. In the self-representing strata the cluster-level random effects are omitted completely, and the probability distribution of y_{hik} is determined by the linear predictor $\mathbf{x}_{hik}^T \boldsymbol{\beta} + u_h$.

Multiple imputation (Tanner and Wong 1987) is used for augmenting missing data values, and therefore weights are not used in the Bayesian analyses. All 8,028 people were used in analyses. Systolic blood pressure, waist circumference, chronic illness and education had 1,692, 1,739, 1,047 and 1,130 missing values respectively. Other variables of the analyses had no missing values. In Bayesian inference, the missing data values and model parameters are treated in the same way. The missing values of the categorical covariates waist circumference and education are imputed from predictive distributions based on multinomial logistic regression models, which contain age, gender, language and health centre district as explanatory variables. The missing values of the response are imputed by using the analysis models (4) and (5) as predictive distributions.

In the estimation 10,000 iterations of MCMC are executed in addition to 1,000 iterations of burn-in. We report the posterior expectations of the model parameters with standard deviations, MC errors and selected quantiles. For a comparison with the other options, we also calculate studentized statistics.

4. RESULTS

4.1 Linear Models for Systolic Blood Pressure

Our first exercise considers modeling of the variation of systolic blood pressure. The main substance matter predictor was the circum waist variable. For illustrative purposes we used it as a three-category variable (RCIRCUM). In the analysis we adjusted for sex and age effects (age was used as a continuous predictor). A preliminary analysis indicates that the mean systolic blood pressure tends to increase with increasing age, and for a given age group, the means increase with increasing circum waist level. The overall mean is higher for males. The structure of the variation in the mean levels of systolic blood pressure was examined by a linear ANCOVA model under the five analysis options, including the main effects of all predictors as well as their pair-wise interaction terms. When ignoring the design complexities, the analysis under the reference option suggests that a reasonable model includes all the main effects and pair-wise interaction effects of circum waist with age and sex. A similar model was also obtained under the more realistic options that account for the design complexities. However, under these options, the significance of the interaction effect of sex with circum waist was weaker than for the reference option (Table 3).

Table 3. Testing the interaction effect of sex with circum waist in modeling systolic blood pressure under the analysis options.

	DF	F value	Prob.
Reference Option	2	5.51	0.0041
Option 1	2	5.24	0.0054
Option 2 a	2	4.69	0.0093
Option 2 b	2	4.65	0.0096
Option 3	2	4.72	0.0090

Let us examine in more detail the interaction of sex with circum waist. Results are summarized in Table 4. Most liberal results were again given by the Reference Option. Options 1 to 3 give closely agreeing point estimates and estimated standard errors. However, for Options 2a, 2b and 3, the t-test statistics were slightly smaller than for Option 1. In Option 1, estimation is based on the PML method where clustering effects are accounted for in the estimation of the standard errors (but not in the estimation of the fixed effects parameters). In Options 2a, 2b and 3, clustering effects are accounted for both in the estimation of the fixed effect parameters and in standard error estimation, either by the GEE method with an exchangeable correlation structure or by fitting a linear mixed model with cluster-specific random intercepts. When compared to the other options, somewhat smaller point estimate and standard error estimate are given by Option 4, using hierarchical Bayesian techniques for a random effects model.

Table 4. Testing the interaction effect of sex (males) with circum waist (middle class) in modeling systolic blood pressure under the analysis options.

	Estimate	Standard Error	Design effect	t-test (Studentized value)	Prob.
Reference Option					
sex(1).rcircum(2)	0.019	0.0088	1.00	2.14	0.0324
Option 1 (SUDAAN/REGRESS)					
sex(1).rcircum(2)	0.018	0.0091	1.06	2.02	0.0430
Option 2a (SUDAAN/REGRESS)					
sex(1).rcircum(2)	0.017	0.0091	1.05	1.87	0.0614
Option 2b (SAS/GENMOD)					
sex(1).rcircum(2)	0.017	0.0089	1.02	1.90	0.0573
Option 3 (SAS/MIXED)					
sex(1).rcircum(2)	0.017	0.0089	1.02	1.93	0.0532
Option 4 (WinBUGS)					
sex(1).rcircum(2)	0.0093	0.0084	-	1.12	-

4.2 Logistic models for chronic morbidity

We next turn to non-linear multivariate modeling. The response variable, chronic morbidity, is binary and calls for logistic modeling. We used age and sex as predictors in addition to education, which was included as a variable of substance matter interest. Again for illustrative purposes, we used education as a three class predictor (REDUC). Age was incorporated as a continuous predictor in the model, whereas sex and education were categorical. A preliminary analysis indicates that the prevalence of chronic morbidity tends to increase with increasing age, and for a given age group, prevalence decreased with increasing education level. The overall mean was higher for females.

We selected the analysis options 2 and 4, in addition to Reference Option, for a closer inspection. It appeared that in this analysis setting, the most conservative results were obtained with the Reference Option ignoring all the sampling complexities. The design-based GEE options 2a and 2b gave results close to each other, while the Bayesian option 4 produced slightly more conservative results than options 2a and 2b.

All the proper options examined for modeling of chronic morbidity supported a similar model, proposing at least some statistical significance for all the predictors, education, sex and age. Of these, Option 4 appeared to be most conservative. The reference option, however, gave a slight support to a model where the variable sex would be excluded (Table 5).

Table 5. Testing sex effect in modeling chronic morbidity under selected analysis options.

	Estimate	Standard error	Design effect	t-test (Studentized value)	Prob.
Reference Option					
Sex effect	0.1042	0.0570	1.00	1.83	0.0675
Option 2a (SUDAAN/REGRESS)					
Sex effect	0.1125	0.0542	0.91	2.08	0.0380
Option 2b (SAS/GENMOD)					
Sex effect	0.1129	0.0532	0.87	2.12	0.0337
Option 4 (WinBUGS)					
Sex effect	0.1289	0.0546	-	2.36	-

5. DISCUSSION

Health surveys are often complex involving stratification, clustering and unequal inclusion probabilities in the sampling design and imputation and re-weighting schemes in the estimation design. For these surveys, analysis strategies are required which are statistically sound and manageable in practice. Our aim was to investigate the available methodologies and apply the corresponding computational tools in a fairly simple analysis setting for a recent health survey data collected in Finland in the context of the Health-2000 Study. It appeared that to a reasonable extent, the major design complexities can be accounted for by using the appropriate computational tools that are readily available in commonly used statistical software products.

We concentrated on methods which aim at accounting for the most important design complexities. Our special interests were in the clustering effects. We therefore constructed an analysis setting where complicated re-weighting and imputation schemes were excluded. The weight variable was a simple design weight with only minor manipulation. The variation of the weights was small and thus, the contribution of weighting in analysis results was minor. We selected two different study variables with a reasonable set of predictors allowing versatile enough model building. Both study variables indicated relatively strong positive intra-cluster correlation (2.41 for the continuous variable systolic blood pressure and 1.75 for the binary variable chronic morbidity).

The analysis options covered methodologies commonly used in complex surveys. The main differences in the methodologies were in inferential framework, model formulation, estimation strategy and software application. Methods representing both frequentist and Bayesian inference were selected. Within the first group, we applied fixed-effects type modeling and modeling with mixed models, both linear and non-linear. Estimation techniques were used that are constructed to be flexible enough for modeling situations of varying complexity. In addition, we made a choice on computational tools such that the methods are readily available in standard statistical software products. Thus we used software products under headlines of “software for design-based analysis”, “software for model-based analysis” and “software for Bayesian analysis”.

Closely agreeing numerical results, and similar inferential conclusions, were obtained under the different analysis options. This suggests that to a reasonable extent, the clustering effects can be accounted for by any of the methods examined (except the reference method ignoring all the sampling design complexities). The methods using pseudo-likelihood and generalized estimating equations under the design-based approach cover many typical analysis situations for generalized linear modeling. Using mixed or multilevel models under the model-based approach, more complex modeling of covariance structures would be possible.

Analyses based on Bayesian techniques provided similar results with the frequentist analyses. Bayesian inference is flexible in analyzing data with missing values because measurement models for multiple imputation can be easily incorporated in the model.

Acknowledgement: The authors are thankful for the Health 2000 Project for having access to the data for this study.

REFERENCES

Aromaa, A. and Koskinen, S. (eds.) (2002), Health and functional capacity in Finland. Baseline results of the Health 2000 health examination survey, Helsinki, Publications of the National Public Health Institute, B3/2002. (In Finnish with English summary).

Binder, D. (1983), On the variances of asymptotically normal estimators from complex samples, *International Statistical Review*, 51, pp. 279-292.

Diggle, P.J., Liang, K.-Y. and Zeger, S.L. (1994), *Analysis of Longitudinal Data*, Oxford, Oxford University Press.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995), *Bayesian Data Analysis*, London: Chapman & Hall.

Goldstein, H. (1995), *Multilevel Statistical Models, 2nd Edition*, London, Arnold and New York, John Wiley & Sons.

Horton, N.J. and Lipsitz, S.R. (1999), Review of software to fit generalized estimating equation regression models, *The American Statistician*, 53, pp. 160-169.

Lehtonen, R. and Pahkinen, E. (1996), *Practical Methods for Design and Analysis of Complex Surveys, Revised Edition*, Chichester, John Wiley & Sons.

Liang, K.-Y. and Zeger, S.L. (1986), Longitudinal data analysis using generalized linear models, *Biometrika* 73, pp. 13-22.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, Second Edition, London, Chapman and Hall.

McCullogh, C.E. and Searle, S.R. (2001), *Generalized, Linear, and Mixed Models*, New York, John Wiley & Sons.

Pfeffermann, D., Skinner, C.J., Goldstein, H., Holmes, D.J. and Rasbash, J. (1998), Weighting for unequal selection probabilities in multilevel models (With discussion), *Journal of the Royal Statistical Society*, Series B, 60, pp. 23-40.

Singer, J.D. (1998), Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models, *Journal of Educational and Behavioral Statistics*, 24, pp. 323-355.

Skinner, C., Holt, T. and Smith, T.M.F. (eds.) (1989), *Analysis of Complex Surveys*, Chichester, John Wiley & Sons.

Tanner, M. A. and Wong W. H. (1987), The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Society*, 82, pp. 528-550.

Ziegler, A., Kastner, C. and Blettner, M. (1998), The generalized estimating equations: an annotated bibliography. *Biometrical Journal*, 40, pp. 115-139.

Web references

Web reference 1: The SAS web site, http://www.sas.com
Web reference 2: The SUDAAN web site, http://www.rti.org/sudaan
Web reference 3: The WinBUGS web site, http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml