SOCIAL AND ECONOMIC STUDIES USING SURVEY DATA:

Accounting for Sample Design

Milorad S. Kovacevic

Data Analysis Research and
Data Analysis Resource Centre
Statistics Canada
kovamil@statcan.ca

Seminar for CIQSS, Montreal

February 8, 2002

Acknowledgment:

This lecture highly benefitted from contributions and earlier seminars given by:

Georgia Roberts (DARC, Statistics Canada),
and
Owen Phillips (DARC, Statistics Canada),
Prof. Roland Thomas (Carleton University, Ottawa)

OUTLINE:

PART I:

I) Introduction

- Complex sample designs
  - -Stratification, Clustering, Weighting
- Statistical inference
- Descriptive vs. analytic studies
- Software for analysis of survey data:
  - General properties

II) Descriptive use of survey data

- Design-based inference
- Point estimation
  - -Design unbiasedness, Consistency, Efficiency
- Interval estimates
- Variance estimation
- Software for analysis of survey data:
  - -Capability for variance estimation

III) Analytic use of survey data I

- -A simple example: test of significance
- Inference for Regression
  - -Example
- Summary
- Software for analysis of survey data:
  - -Analytical methods accommodated

PART II:

IV) Analytic use of survey data II

- Categorical data analysis
- Fitting a logistic regression model
- Analysing duration data

PART III:

- Selected References

Complex Sample Designs

The sample design usually reflects (at least partially) the structure of the population.

A finite population of units with non-stochastic (fixed) values $\{Y_1, \ldots, Y_N\}$

Components of complexity

- Stratification
- Clustering
- Unequal probabilities of selection- weights

- Random adjustments to the weights
  - non-response adjustment
  - post-stratification/calibration

- Imputation

- **Stratification**

– need information from a frame
– partition of the finite population before sample selection
– mutually exclusive and exhaustive subpopulations
– independent samples, different designs, sampling schemes, even estimation

– can reduce variance of estimates (relative to SRS) if chosen efficiently

Example:

LFS, SCF, SLID, NLSCY:
provinces, geographic and/or socio-economic strata:
EIER (53), Urban (954), Rural (210), Apartment (68)

- **Clustering**
  Ultimate units are nested within larger units - clusters

  Reasons: lack of relevant frame; to reduce costs of data collection

  Results:

– multi-stage sampling
– unequal selection probabilities
– **increased variance** (relative to SRS of the same size) due to a positive intracluster correlation
– larger confidence intervals
– reduced number of DF for variance estimation
– reject null hypothesis less often

Ignoring clustering*:*

– variances are incorrect,
– tests are incorrect (reject more often)
– undercoverage of confidence intervals

Example:

LFS, SCF, SLID, NLSCY:
enumeration areas (EA),
groups of EA's

- **Survey Weights**

Inclusion probability $\pi_i = \sum_{s \in S} Prob\{s \ni i\}$

Basic sampling weight: $w_i = \pi_i^{-1}$

$$E\left(\sum_{i \in s} w_i\right) = N$$

Adjustments made to basic weight: (non-response, post-stratification, calibration, etc.)

Remarks: 1. Weight adjustments contribute to reduce estimation biases due to imbalances in the sample.

2. Stratification, clustering and unequal selection probabilities yield samples whose data are **neither independent nor identically distributed**

Statistical Inference

"The development of **generalization** from **sample data,** usually with **calculated** degrees of **uncertainty**"
(J.M. Last (ed.) A Dictionary of Epidemiology, 1998, Wiley)

- The **sample data** are the basis for any statement

- The concept of "**generalization**" implies that there is an object to the inference - A POPULATION (of some form) to which the statistical statements from the sample apply

- Some mechanism (framework, approach, method) is needed to accomplish the task of "generalization"

  Randomness comes from:
  - the sample design (**design-based** approach), or
  - the underlying stochastic process/model assumed to generate the population values (**model-based** approach)

- "**Statistical uncertainty"** accompanies any statement made about the population from a sample. A chosen mechanism should be able to quantify the level of uncertainty.

Descriptive vs. Analytic Studies

Descriptive studies focus on summary measures of the finite population - UNAMBIGUOUS PARAMETERS such as totals, means, frequencies, percentiles, etc.

- Principal inference is design-based with
  - survey weighted estimators and
  - estimators of standard errors based on the sample design

Analytic studies go beyond description, and look for more COMPLEX RELATIONSHIPS (e.g. association, causality), among observed variables.

- Definition of the parameters of interest is no longer unambiguous
  - model parameters such as regression coefficients are parameters of some assumed distribution or superpopulation
  - model parameters are distinct from "parameters" defined on a finite population

- Inference can be *model based* (subject to assumptions about a super-population) or *design based* (with analogues of model parameters defined on the finite population)

- Descriptive / Analytic distinction is useful but is not a barrier.

Software For Analysis Of Survey Data

Desirable features for an analyst:
[Assumption: Analyst can have access to confidential microdata]

- "relatively" easy to use for people familiar with SAS / SPSS / BMDP etc.
- "relatively" easy to combine with software doing data manipulations
- good documentation
- can accommodate the range of analyses of interest to the analyst including the descriptive studies as well
- can do Canada-level analyses (i.e. can handle a file the size of the Canada file)
- can accommodate the design information for all surveys that the analyst uses
- correctly uses survey weights when producing estimates
- gives standard error estimates that adequately account for the survey design
- gives useful test statistics (for frequently used hypotheses)
- gives covariance estimates needed for (less frequently used) hypothesis tests.

	SAS SVY	SUDAAN 7.5/8	WesVar 3.0/4.0	Stata 6
Easy to use	Yes	Yes	Yes	Yes
Documentation	Extensive	Improved	Incomplete	Extensive
Range of analyses	Limited	Wide	Not as much as SUDAAN	Wide
Correctly weighted estimates	Yes	Yes	Yes	Yes
Design-based SE's	Yes(- -)	Yes (-)	Yes(-)	Yes(-)
Useful test statistics	Yes	Yes	Yes	Yes
Design-based covariance	Yes	Yes (-)	For some procedures	?
Canada-level analysis	Yes	Yes	Yes	Yes

II) DESCRIPTIVE STUDIES

The goal: To infer about the **finite population parameters** which are functions of the population values of the variable(s) of interest, typically $\theta = g(Y_1, ..., Y_N)$, using a sample statistic denoted by $\hat{\theta} = h(y_1, ..., y_n)$

Design-Based Approach:

Interest is in finding estimators with **good** properties with respect to their (sampling) distribution over repeated samples $\boldsymbol{s}$ generated by the sample design $p(s)$

<u>Point estimation</u>

Example: Population total $\theta = Y = \sum_i Y_i$

Estimator: $\hat{\theta} = \hat{Y} = \sum_{i \in s} y_i / \pi_i = \sum_{i \in s} w_i y_i$

– **Unbiasedness:** $B(\hat{\theta}) = E_p(\hat{\theta}) - \theta = 0$

where $E_p(\hat{\theta}) = \sum_s p(s)\, \hat{\theta}_s$

– **Consistency:** $\hat{\theta}_s - \theta \to 0$ as the sample increases.

– **Efficiency** expresses the stability of estimates obtained from the samples that the design $p(s)$ can produce.

Variance of $\hat{\theta}$:

$$V_p(\hat{\theta}) = \sum_s p(s)\left[\hat{\theta}_s - E_p(\hat{\theta}_s)\right]^2$$

- defines precision of estimators
- determined by the sample size and various selection strategies (e.g. stratification, clustering)

- **Sampling error**: $\hat{\theta} - \theta$
- **Mean squared error of $\hat{\theta}$:**

$$MSE_P(\hat{\theta}) = E_p(\hat{\theta} - \theta)^2 = V_p(\hat{\theta}) + [B_p(\hat{\theta})]^2$$

  - indicates the amount of uncertainty that accompanies the estimate

- Both fundamental statistical qualities of the sample strategy
  - validity
  - efficiency

  are implicit in MSE.

Interval Estimates for the parameter θ with approximate $1-\alpha$ confidence level

$$\hat{\theta} \pm t_{d,1-\alpha/2} [\hat{V}(\hat{\theta})]^{-1/2}$$

$t_{d,1-\alpha/2}$ is $1-\alpha/2$ quantile of a ***t***-distribution with *d* DF

$\hat{V}(\hat{\theta})$ is an estimator for $V_p(\hat{\theta})$

Remark: For large samples (large #DF) use $z_{1-\alpha/2}$, $1-\alpha/2$ quantile of the $N(0,1)$-distribution

- **Number of DF** for the multistage design is typically approximated by:

  #(sampled PSUs)-#(strata) (for a stratified sample of PSUs)

  Finer approximations:

  Satterthwaite (1946) approximation:

$$d = 2\frac{[\hat{V}(\hat{\theta})]^2}{\hat{V}[\hat{V}(\hat{\theta})]}$$

  Kenward-Roger (1997) approximation

Variance Estimation

- For a stratified multi-stage sample
- Assumptions:

(*) PSU's are sampled with replacement in all strata

(**) the total Y_h for stratum h can be approximately, independently and unbiasedly estimated from each PSU selected from stratum h.

(*) and (**) are standard assumptions made in most software packages

- A conservative estimator of the design based variance of $\hat{Y}$ (or any **linear function** of the observations) is

$$(1) \qquad \hat{V}(\hat{Y}) = \sum_{h=1}^{H} \frac{m_h}{(m_h - 1)} \sum_{c=1}^{m_h} (u_{hc} - \bar{u}_h)^2,$$

where $u_{hc} = \sum_{i=1}^{m_{hc}} w_{hci} y_{hci}$, and $\bar{u}_h = m_h^{-1} \sum_c u_{hc}$.

- *Advantage:* the sub-sampling within PSU's can be ignored for variance estimation purposes. Only estimates at the PSU level need to be computed.

If (*) doesn't hold a separate estimation of the between-cluster and within-cluster variance components is needed.

– It requires knowledge of the joint inclusion probabilities $\pi_{hcc'}$

– Yates-Grundy-Sen estimator:

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H} \sum_{c=1}^{m_h} \sum_{c>c'}^{m_h} \frac{\pi_{hc}\pi_{hc'} - \pi_{hcc'}}{\pi_{hcc'}} (u_{hc} - u_{hc'})^2 + \sum_{h=1}^{H} \sum_{c=1}^{m_h} \pi_{hc}(1 - f_{hc}) \frac{m_{hc}}{m_{hc} - 1} \sum_{i=1}^{m_{hc}} (u_{hci} - \bar{u}_{hc})^2 \quad (2)$$

where $u_{hci} = w_{hci} y_{hci}$, $\bar{u}_{hc} = m_{hc}^{-1} \sum_i u_{hci}$ and
$f_{hc} = m_{hc} / M_{hc}$ (the second stage sampling rate.)

– Handling the nonresponse adjustments and post-stratification further complicates the estimation

Variance Estimation for Non-Linear Estimators

Example: When N is not known, to estimate the mean $\bar{Y}$, use

$$\hat{Y}/\hat{N} = \frac{\sum_{h=1}^{H}\sum_{c=1}^{m_h}\sum_{i=1}^{m_{hc}} w_{hci}y_{hci}}{\sum_{h=1}^{H}\sum_{c=1}^{m_h}\sum_{i=1}^{m_{hc}} w_{hci}}.$$

Design-based inference depends critically on the availability of methods for estimating the variances of non-linear estimators. Several methods are available:

- Linearization
- Replication methods:
  - Jackknife,
  - Bootstrap,
  - Balanced Half-Sample Replication.

- **Linearization approach**

– uses Taylor series expansions to approximate non-linear statistics as linear combinations of estimated means and totals.

– The variance estimators (1) or (2) can be applied.

– The approach can be applied to a wide variety of statistics - it will be described later in the context of analytic use of survey data.

Disadvantages:

– Estimator specific
– Complexity forces ignoring of impact of some weight adjustments

Replication methods (re-sampling methods):

All involve calculating the non-linear estimator $\hat{\theta}$ on a set of different subsets (replicates) of the data.

– The variance estimator is constructed using sums of squares based on replicate estimates i.e.,

(3) $$\hat{V}(\hat{\theta}) = \sum_r A_r (\hat{\theta}_r - \hat{\theta})^2,$$

where r denotes a particular replicate of the sample, and A_r is a constant that depends on the replication method. **(see Rust and Rao, 1996)**:

- **Jackknife:**
  – drop the c-th PSU from the h-th stratum, upweight the sample weights of the other PSU's in the h-th stratum, do all weight adjustments, and calculate $\hat{\theta}_{(hc)}$, for $h = 1,..,H;\ c = 1,\ldots,m_h$.

  $$A_{hc} = (m_h - 1) / m_h$$

- **Bootstrap:**
  - generate a large number B of sample replicates by drawing a simple random sample of $m_h - 1$ PSU's with replacement from the m_h sampled PSU's, independently for each stratum h;
  - weights are adjusted so that the sampled PSUs represent the population;
  - all other weight adjustments are done too;
  - $\hat{\theta}_r$ is an estimate of θ based on the rth replicate $r = 1,...,B$.

  - $A_r = 1/B$ (Constant for all replicates)

- **BRR**: (used for designs with 2 PSU's per stratum.)
  - Each balanced half-sample (i.e., replicate) is a half sample selected according to an orthogonal array of 1's and 0's, of dimension T, related to H.
  - Then replicates are $\hat{\theta}_r$, $r = 1,...,T$.

    $A_r = 1/T$ (Constant for all replicates)

- **Using A Set of the Ready-to-use Replicate Weights**:
  - Calculate the replicate estimates $\hat{\theta}_r$
  - Apply formula (3) with the appropriate constants A_r.

Related issues

- Number of degrees of freedom

  Usually approximated by

  min {#replicates, #PSU's - #strata}

- Estimation for domains
  -For linearization and JK method:
  Elimination of individuals outside the subpopulation from the data set before analysis can lead to incorrect variance estimators

  -For replicate methods
  May use only the observations from the domain

- Computationally intensity

  - New modifications:
  One step jackknife
  Linearized bootstrap

Software For Analysis Of Survey Data

VARIANCE ESTIMATION FEATURES - I

	SAS SVY	Stata 6	SUDAAN 7.5/ 8	WesVar 3.0/4
1st stage design: STRWR	Y	Y	Y	Y
Taylor linearization (TL)	Y	Y	Y	
TL - fpc at 1st stage	Y	Y	Y	
TL- fpc at subsequent stages			Y	
TL - 1st stage: STRWOR, unequal probs Subsequent stages: equal probs, WR or WOR			Y	
TL - Post-stratification adjustment			Partly (-)	
Warnings regarding suitable design		324U	3.4/4.8	pp 3, 138

VARIANCE ESTIMATION FEATURES - II

	SAS SVY	Stata 6	SUDAAN 7.5/ 8	WesVar 3.0 / 4
Jackknife - create replicate weights			Y(-)	Y
Jackknife - supply replicate weights			Y	Y
BRR - create replicate weights			N	Y
BRR - supply replicate weights			Y(-)	Y
Bootstrap - create replicate weights			N	N
Bootstrap - supply replicate weights			Y	Y
fpc at first stage			N	Y
Create replicate weights with post-strat adj			N	Y
Create replicate weights with raking ratio adj				Y
Create replicate weights with other weight adj			N	N/Y

III) ANALYTIC USE OF SURVEY DATA I

A simple example:

(Example motivated from "The Upward Mobility of Low Paid Canadians" by Drolet and Morissette (1998))

The test of significance of the difference in the percentages of the 'low weekly earners' in Canadian large firms (500+ employees) for males and females in 1993

$$H_0: \quad p_M - p_F = 0 \text{ against } H_A: \quad p_M - p_F \neq 0.$$

Remark: Some statisticians argue that, such a test is meaningless for the finite population because if a complete census was taken, the two percenteges would not be equal (only in exceptional cases) **(Korn and Graubard, 1999)**

A model-based approach:

- View the male sample observations as **independent random realizations** from an infinite population of values each taking value 1 with probability p_M, and value 0 with probability 1-p_M. Thus,

$$E_\xi(I_i)=p_M, \qquad V_\xi(I_i)=p_M(1-p_M)$$

Hence, the observations are **identically distributed** as well.

And similarly, the female sample observations are identically distributed independent random realizations from an infinite population of values each taking value 1 with probability p_F, and value 0 with probability 1-p_F.

These two distributions are assumed independent.

- Then the inference is made about the infinite population's (superpopulation) parameters: p_M, p_F.

- The number of the men (women) in the sample who are the employees in the large firms and who are the low income earners is then binomially distributed:

$$\sum_{i=1}^{n_M} I_i \sim binomial(p_M, n_M), \quad \sum_{i=1}^{n_F} I_i \sim binomial(p_F, n_F)$$

The test statistic is

$$T = \frac{\hat{p}_M - \hat{p}_F}{\sqrt{\frac{\hat{p}_M(1-\hat{p}_M)}{n_M} + \frac{\hat{p}_F(1-\hat{p}_F)}{n_F}}},$$

where $\hat{p}_M = \sum_{i=1}^{n_M} I_i / n_M$ and $\hat{p}_F = \sum_{i=1}^{n_F} I_i / n_F$,

and if H_0 is true, $T \sim N(0,1)$

where $\hat{p}_M = \sum_{i=1}^{n_M} I_i / n_M$ and $\hat{p}_F = \sum_{i=1}^{n_F} I_i / n_F$

***Potential Problems*:**

- Misspecification of the model
  - The distributions for men and women may not be independent binomial distributions
  - The sample observations may not be identically distributed (e.g. heteroscedastic)

- Sample doesn't represent the distribution
  - The sample observations may be correlated instead of independently distributed (e.g. clustering)
  - There is a selection bias (more low income earning men in the sample)

Consequences:

- The model parameters may no longer have a substantive interpretation.
- The estimated parameters are biased and their standard errors are incorrect.

Design-based approach:

- In this approach, the SLID sample can be viewed as a two-phase sample:
  - The first-phase sample - produces the finite population,
  - The second-phase sample produces the survey sample

  View the finite population as a **random realization** of a process that generates the population, or as a random sample from a **superpopulation**, ξ.

- As before, the superpopulation model can be described as ***independent binomial sampling*** since observations on Y {indicator of the low weekly earnings} are independently binomially distributed for male employees ***bin***(π_M, N_M) and female employees ***bin***(π_F, N_F)**.**

  - The inference is made about the superpopulation parameters: π_M, π_F.

- The design-based estimate of the difference

$$\hat{D}=\hat{p}_M-\hat{p}_F$$

is approximately **design-unbiased** for the finite population difference $D=p_M-p_F$, (i.e. $E_p(\hat{D}) \approx D$)

- The finite population difference D is **model-unbiased** estimate of the superpopulation parameter $\Delta=\pi_M-\pi_F$, i.e.

$E_\xi(D) = \Delta$

- This implies that the design-based estimate $\hat{D}$ is unbiased for Δ as well, since

$$E(\hat{D})=E_\xi E_p(\hat{D}) = E_\xi(D) = \Delta$$

- The variance of $\hat{D}$ has also to account for both 'phases':

$$V(\hat{D}) = E_\xi V_p(\hat{D}) + V_\xi E_p(\hat{D}) = E_\xi V_p(\hat{D}) + V_\xi(D)$$

- If $\hat{V}_p(\hat{D})$ is an unbiased estimator for $V_p(\hat{D})$ it is also unbiased for $E_\xi V_p(\hat{D})$ (anticipated variance).

- Under the superpopulation model, $V_\xi(D)$ is equal to

$$V_\xi(D) = V_\xi(p_M) + V_\xi(p_F) = \frac{p_M(1-p_M)}{N_M} + \frac{p_F(1-p_F)}{N_F}$$
$$= O_p(N_M^{-1}) + O_p(N_F^{-1}) \approx 0$$

where N_M and N_F are the respective sizes of male and female groups of employees in Canadian large firms.

- **Thus, $V(\hat{D})$ can be approximately unbiasedly estimated by the design-based variance $\hat{V}_p(\hat{D})$.**

Therefore, inference about the superpopulation parameters can be done using the design-based estimates, $\hat{D}$, $\hat{V}_p(\hat{D})$.

Note that for these estimates no assumption about the superpopulation is needed:

– They are **robust** to model misspecification.

How does SAS estimate the standard errors?

- Procedure UNIVARIATE
  - Unweighted (purely model-based)

$$\hat{p}_{M,u} = \sum_{i=1}^{n_M} I_i / n_M \ ,$$

$$\hat{V}(\hat{p}_{M,u}) = \frac{1}{n_M} \frac{\sum_{i=1}^{n_M} (I_i - \hat{p}_{M,u})^2}{n_M} = \frac{\hat{p}_{M,u}(1-\hat{p}_{M,u})}{n_M}$$

$$s.e.(\hat{p}_{M,u}) = \sqrt{\hat{V}(\hat{p}_{M,u})} \quad \text{(STD MEAN)}$$

- Weighted (still model-based):

$$\hat{p}_M = \sum_{i=1}^{n_M} w_i I_i / \sum_{i=1}^{n_M} w_i \ ,$$

$$\hat{V}(\hat{p}_M) = \frac{1}{\sum w_i} \frac{\sum_{i=1}^{n_M} w_i (I_i - \hat{p}_M)^2}{d} = \frac{\hat{p}_M (1 - \hat{p}_M)}{d}$$

$$s.e.(\hat{p}_M) = \sqrt{\hat{V}(\hat{p}_M)} \quad \text{(STD MEAN)}$$

Note that $\hat{p}_M$ is the same with unscaled and scaled weights.

d is controlled by the VARDEF option.

- If d = $\sum w_i$ (VARDEF=WGT)

– use of unscaled weights produces very small standard errors

- If d = n (default option)

– no difference between the use of scaled and unscaled in the variance estimation
– accounts for the differential weighting but not for other design components

How does SUDAAN estimate the standard errors?

$$(4) \qquad \hat{p}_M = \frac{\sum_{h=1}^{H}\sum_{c=1}^{m_h}\sum_{i=1}^{m_{hc}} w_{hci} I_{hci}(subp)\, I_{hci}(L)}{\sum_{h=1}^{H}\sum_{c=1}^{m_h}\sum_{i=1}^{m_{hc}} w_{hci} I_{hci}(subp)}$$

subp defines the subpopulation (male employees, large firms)

$I_{hci}(L)$ indicates if the person is low weekly earner

The first order Taylor series gives

$$V(\hat{p}_M) \approx Var\left(\sum_h \sum_c \sum_i w^*_{hci}[I_{hci}(L) - p_M]\right)$$

where

$$w^*_{hci} = \frac{w_{hci} I_{hci}(subp)}{\sum_{h=1}^{H}\sum_{c=1}^{m_h}\sum_{i=1}^{m_{hc}} w_{hci} I_{hci}(subp)}$$

Finally, the estimate is

$$\hat{V}(\hat{p}_M) = \sum_{h=1}^{H} \frac{m_h}{(m_h - 1)} \sum_{c=1}^{m_h} (u_{hc} - \bar{u}_h)^2$$

where

$$u_{hc} = \sum_{i=1}^{m_{hc}} w^*_{hci}\left[I_{hci}(L) - \hat{p}_M\right] \text{ and } \bar{u}_h = m_h^{-1}\sum_c u_{hc}.$$

Numerical results:

Estimates of low weekly earners in 1993 in large firms (#employees > 500) by gender

Method	Gender	Estimates					
		%	Standard Errors				Deff
			SAS		SUDAAN		WesVar
					Taylor	Replicate	
Unweighted	Male	27.6	1.87				2.08
	Female	33.4	1.17				2.63
Weighted	Male	30.0	0.06*	1.92**	2.7	2.6	2.7	1.97
	Female	35.9	0.04*	1.19**	1.9	1.9	1.9	2.54

(*) unscaled weights

(**) scaled weights

$$w_i^* = w_i / \bar{w}$$

Test statistic: $$T = \frac{\hat{D}}{\sqrt{\hat{V}(\hat{D})}}$$

where $\hat{V}(\hat{D}) = \hat{V}(\hat{p}_M) + \hat{V}(\hat{p}_F) - 2\hat{Cov}(\hat{p}_M, \hat{p}_F)$

$\hat{D}$=5.9

	SAS			SUDAAN, Taylor
	unweighted	unscaled	scaled	
$\hat{V}(\hat{p}_M)$	3.497	0.0036	3.686	7.129
$\hat{V}(\hat{p}_F)$	1.369	0.0016	1.416	3.610
$\hat{Cov}(\hat{p}_M, \hat{p}_F)$	0, by assumption			0.53
$\sqrt{\hat{V}(\hat{D})}$	2.20	0.07	2.26	3.19
T	2.63	84	2.61	1.85
Interval Estimate (95%)	(1.49,10.11)	(5.76, 6.04)	(1.47, 10.33)	(-0.36, 12.16)

Model Setting

Consider a dependent variable y and p explanatory variables $x_1, \ldots, x_p$ measured on each unit of a finite population.

This is equivalent to assume that:

- the population values y_i and x_{ij}, $j = 1, \ldots, p$ are realizations from a distribution ξ such that the conditional distribution of the y's given the x's satisfies

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i,$$

or in matrix terms,

$$\boldsymbol{y_N} = \boldsymbol{X_N}\boldsymbol{\beta} + \boldsymbol{\varepsilon_N}, \qquad (6)$$

with $E(\varepsilon_N) = \mathbf{0}$ and $Var(\varepsilon_N) = \Sigma_N$.

- If $\Sigma_N = \sigma^2 I_N$ (iid assumption)

- If the model recognizes the natural clusters of population units then Σ_N will be block diagonal.

- More general models - **multilevel models**:

  – allow for explanatory variables measured at different levels of a population hierarchy

  Example: -at the level of a child within a class,
  -a class within a school,
  -and a school within a school district.

  – multi-level models incorporate random variation within each level

  (see Goldstein 1995)

Model-based approach

- **assumes** that model (6) applies to every element in the sample, i.e.

$$\boldsymbol{y_n} = \boldsymbol{X_n}\boldsymbol{\beta} + \varepsilon_n, \qquad (7)$$

with $E(\varepsilon_n) = \mathbf{0}$ and $Var(\varepsilon_n) = \Sigma_n$ (a sub-matrix of Σ_N)

This is equivalent to assumption of '**no selection bias**'

- When Σ_n is known, BLUE estimates of β are given by Generalized Least Squares (GLS),

$$\hat{\beta}_{GLS} = [\boldsymbol{X_n}'\Sigma_n^{-1}\boldsymbol{X_n}]^{-1}\boldsymbol{X_n}'\Sigma_n^{-1}\boldsymbol{y_n},$$

with variance $Var(\hat{\beta}_{GLS}) = [\boldsymbol{X_n}'\Sigma_n^{-1}\boldsymbol{X_n}]^{-1}$

- When Σ_n can be consistently estimated, then using $\hat{\Sigma}_n$ in the GLS estimator results in estimates that are asymptotically efficient
(e.g. Fuller (1984) and Rao, Sutradhar and Yue (1991).

Model misspecification: Ignoring the Selection Bias

- Suppose that there is a selection bias so that the random component of the model is not centered at zero anymore

$$E(\varepsilon_n) = \theta_n$$

- The derived GLS estimator $\hat{\beta}_{GLS}$ is biased for β since:

$$E\{\hat{\beta}_{GLS}\} = \beta + [\boldsymbol{X}_n' \Sigma_n^{-1} \boldsymbol{X}_n]^{-1} \boldsymbol{X}_n' \Sigma_n^{-1} \theta_n$$

  Here the expectation is taken according to model (7) that assumes no selection bias.

- Hence, the presence of the selection bias in the data results in biased estimation of regression coefficients.

Model misspecification: Ignoring the Clustering

– Suppose that the model variance is misspecified, for example, assume

$$V(\varepsilon_n) = \sigma^2 \boldsymbol{I}_n$$

when in fact the super-population model (Σ_n) applies.

- Using Ordinary Least Squares (OLS):

$$\hat{\beta}_{OLS} = [\boldsymbol{X}_n'\boldsymbol{X}_n]^{-1}\boldsymbol{X}_n'\boldsymbol{y}_n$$
$$V_{OLS}(\hat{\beta}_{OLS}) = \sigma^2[\boldsymbol{X}_n'\boldsymbol{X}_n]^{-1}$$

– $\hat{\beta}_{OLS}$ is model unbiased for the regression parameter β;

– under the true super-population model, the variance of $\hat{\beta}_{OLS}$ is given by

$$Var(\hat{\beta}_{OLS}) = [\boldsymbol{X}_n'\boldsymbol{X}_n]^{-1}[\boldsymbol{X}_n'\Sigma_n\boldsymbol{X}_n][\boldsymbol{X}_n'\boldsymbol{X}_n]^{-1}$$

– using an inappropriate OLS model can result in underestimates of the variances of estimated regression coefficients (and hence in inflated *t*-statistics). **(Scott and Holt,1982)**

Discussion

- Subject to the **key assumption of no selection bias**, and provided **the structure is not too complicated**, model based analyses of survey data can account for survey features like stratification and clustering, and can be efficient.

- Inappropriate assumptions (e.g. ignoring over-sampling and clustering) can lead to biases and invalid test statistics.

- The assumption of ***no selection bias*** means that under the model (6):

  - the selection probability $p(\boldsymbol{s})$ (sample design) cannot depend on the response variable $\boldsymbol{y}_N$

  - nor can $p(\boldsymbol{s})$ depend on some other variable Z_N, related to $\boldsymbol{y}_N$ (given $\boldsymbol{X}_N$) , that is omitted from the model

  - $p(\boldsymbol{s})$ can depend on $\boldsymbol{X}_N$

- In the presence of selection bias, the distributions of ε_N and ε_n will differ, and model estimates of regression parameters, e.g., $\hat{\beta}_{GLS}$ will be biased.

- In the literature, the terms *informative sample design* and *(non-)ignorable sample design* are frequently used to describe situations where selection bias will occur **(see Binder and Roberts, 2001)**

- For highly complex samples, technical definitions of non-ignorability (leading to selection bias) are very difficult to validate.

- Thus the existence of selection bias becomes an empirical issue, and, model based analyses must be compared to design based analyses that fully account for selection.

Design-Based Inference for Regression

- Again, the survey sample can be viewed as a two-phase sample:
  - A first-phase sample - produces the finite population,
  - A second-phase sample produces the survey sample

- Consider the finite population with N realized sets of values (y_i, x_{ij}) as before .

- The **finite population regression parameter $\boldsymbol{B}$** is

$$\boldsymbol{B} = [\boldsymbol{X}_N' \boldsymbol{X}_N]^{-1} \boldsymbol{X}_N' \boldsymbol{y}_N$$

- This is example of a **"corresponding descriptive population quantity"** (CDPQ) **(Pfefferman, 1993)**

  – $\boldsymbol{B}$ is the parameter that would be obtained by applying a LS estimation rule to the entire finite population

  – $\boldsymbol{B}$ is model-unbiased for the super-population parameter β, **irrespective** of the distribution specification

$$E_\xi\{\boldsymbol{B}\} = [\boldsymbol{X}_N'\boldsymbol{X}_N]^{-1}\boldsymbol{X}_N' E_\xi\{\boldsymbol{y}_N\} = \boldsymbol{\beta}$$

The design based approach:

– estimate the CDPQ using sample weighted estimates. Thus

$$\hat{\boldsymbol{B}} = [\boldsymbol{X}_n'\boldsymbol{W}_n\boldsymbol{X}_n]^{-1}\boldsymbol{X}_n'\boldsymbol{W}_n\boldsymbol{y}_n$$

where $\boldsymbol{W}_n$ is the diagonal matrix of sampling weights for units in the sample.

– $\hat{\boldsymbol{B}}$ is unbiased for the model parameter

$$E(\hat{\boldsymbol{B}}) = E_\xi E_p(\hat{\boldsymbol{B}}) = E_\xi(\boldsymbol{B}) = \boldsymbol{\beta}$$

Design-Based Variance Estimation

- The design-based variance estimator is obtained by linearization as

$$\hat{\boldsymbol{V}}_{\boldsymbol{L}}(\hat{\boldsymbol{B}}) = [\boldsymbol{X}_{\boldsymbol{n}}' \boldsymbol{W}_{\boldsymbol{n}} \boldsymbol{X}_{\boldsymbol{n}}]^{-1} \hat{\Omega}(\hat{\boldsymbol{B}}) [\boldsymbol{X}_{\boldsymbol{n}}' \boldsymbol{W}_{\boldsymbol{n}} \boldsymbol{X}_{\boldsymbol{n}}]^{-1}$$

  where $\Omega(\hat{\boldsymbol{B}})$ is the covariance matrix of an estimated total $\sum_n w_i \varepsilon_i \boldsymbol{x_i}$, where $\varepsilon_i = y_i - \boldsymbol{x_i}\hat{\boldsymbol{B}}, \quad i = 1,\ldots,n$ is a linear model residual.

- To implement the linearization estimator, all that is required is a method for estimating the variance of an estimated total.

- An important property of $\hat{\boldsymbol{V}}_{\boldsymbol{L}}(\hat{\boldsymbol{B}})$ is that it is approximately *model unbiased* for the *model based variance* of $\hat{\boldsymbol{B}}$, under a very general super-population model that allows different model variances in each PSU (**Kott, 1991**).

- When applied to an SRS design, $\hat{\boldsymbol{V}}_{\boldsymbol{L}}(\hat{\boldsymbol{B}})$ reduces to

$$\hat{\boldsymbol{V}}_{L,OLS}(\hat{\boldsymbol{B}}) = \frac{n}{n-1}[\boldsymbol{X}_n'\boldsymbol{X}_n]^{-1}(\boldsymbol{X}_n'\boldsymbol{R}\boldsymbol{R}\boldsymbol{X}_n)[\boldsymbol{X}_n'\boldsymbol{X}_n]^{-1}$$

where $\boldsymbol{R}$ is a diagonal matrix of residuals.

This is White's (1980) heteroscedasticity robust variance estimator (implemented in SAS)

Using Weights in Model-Based Estimation Software

- For design-based inference, the weights are properties of the *sample design*.

- For model based Weighted Least Squares (WLS) regression, the weights are features of the **model** *(precision weights)*

  – In model (7) we assume that

$$\boldsymbol{\varepsilon_n} \sim (0,\ \sigma^2 W_n^{-1}) \qquad (8)$$

- The estimated coefficients are the same,

$$\hat{\beta}=\hat{\boldsymbol{B}} = [\boldsymbol{X_n'W_nX_n}]^{-1}\boldsymbol{X_n'W_ny_n}$$

However, this is coincidental.

- Also, the WLS variance of $\hat{\beta}$ as

$$V_{WLS}(\hat{\beta}) = \sigma^2 [\boldsymbol{X_n'} \boldsymbol{W_n} \boldsymbol{X_n}]^{-1},$$

while the appropriate *model based variance* under general Σ_N super-population model is

$$\boldsymbol{V_{GSP}}(\hat{\boldsymbol{\beta}}) = [\boldsymbol{X_n'} \boldsymbol{W_n} \boldsymbol{X_n}]^{-1} [\boldsymbol{X_n'} \boldsymbol{W_n} \Sigma_n \boldsymbol{W_n} \boldsymbol{X_n}] [\boldsymbol{X_n'} \boldsymbol{W_n} \boldsymbol{X_n}]^{-1}.$$

Design Based or Model Based?

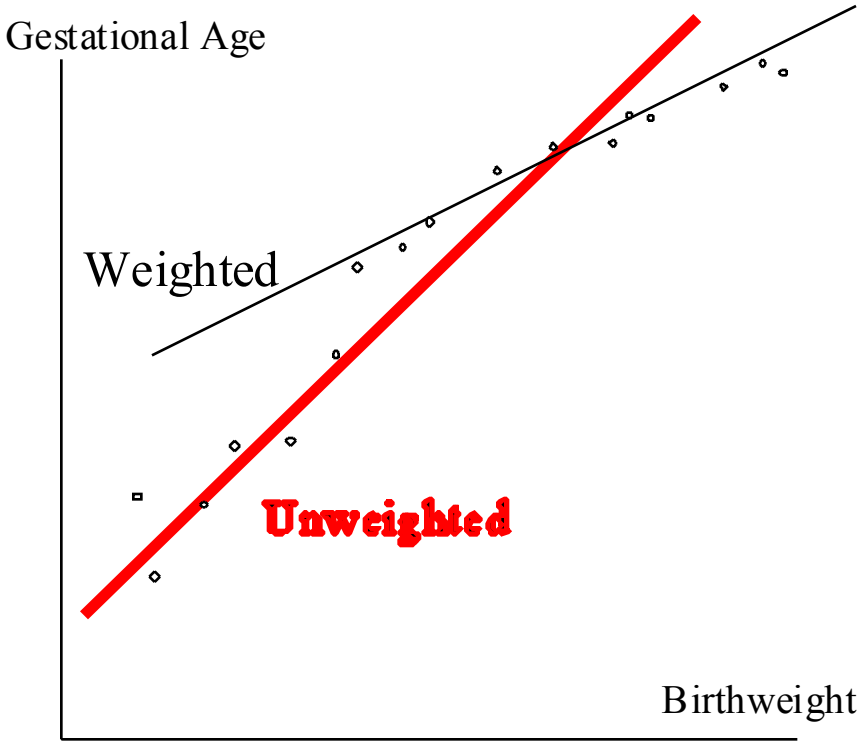
- **This discussion is general and applies beyond the linear regression.**

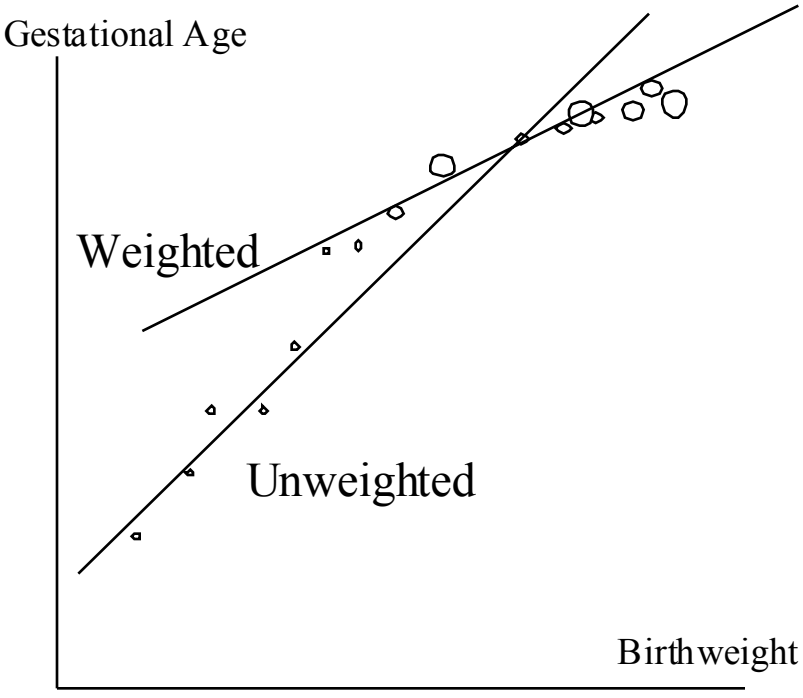
- Ignorability means that $p(\boldsymbol{s})$ does not provide any information *over and above* what is provided by the "design variables" included in the model.

- If $p(\boldsymbol{s})$ depends on the response variable $\boldsymbol{y}$, then the selection scheme is definitely NOT ignorable.

- If sample selection is not obviously dependent on response, the design may still not be ignorable. (The technical conditions for ignorability are very difficult to verify - see, for example, **Sugden and Smith, 1984**).

- Ignoring the weights in this case will result in selection bias. The ***design based inference provides protection.***





(adapted from Korn and Graubard, 1999)

- **Protection by design-based inference**

- Based on the following observations:
  - parameters are defined as CDPQ's
  - design based estimators are *design consistent* for these CDPQ's (e.g., $\hat{\boldsymbol{B}}$ is design consistent for $\boldsymbol{B}$);
  - inference on $\hat{\boldsymbol{B}}$ is based on the randomization distribution induced by the sample design and is completely model-free;
  - *if the model holds in the population*, then as the population size increases, $\boldsymbol{B} \to \boldsymbol{\beta}$:

$$\hat{\boldsymbol{B}} - \boldsymbol{\beta} = (\hat{\boldsymbol{B}} - \boldsymbol{B}) + (\boldsymbol{B} - \beta) = O_p(n^{-1/2}) + O_p(N^{-1/2})$$
$$= O_p(n^{-1/2})$$

[Pfeffermann (1993)]

  - Thus the design based estimate will also converge to the model parameter.

  - If the population model is misspecified, the model parameter β may no longer have a substantive interpretation.

    Nevertheless, $\boldsymbol{B}$ is a real quantity that will still be interpretable, e.g., as the "best linear approximation to the response variable in the least squares sense".

- When the model is correct, the design based variance estimator $\hat{\boldsymbol{V}}_{\boldsymbol{L}}(\hat{\boldsymbol{B}})$ is approximately model unbiased for the model variance of $\hat{\boldsymbol{B}}$, under very general super-population assumptions (Kott, 1991; Pfeffermann, 1993)

- Thus if the model holds, design based inference also consistently estimates the model parameter and provides an approximately model unbiased variance estimate in large samples.

- A possible drawback of design based inference is that it can be less efficient than model based inference when the model is correct.

  – loss of efficiency increases as variation among sample weights increases and as effective sample size decreases
  [see Pfeffermann, 1993; Korn and Graubard, 1999]

  – but, we generally have large sample sizes

Testing on selection bias

– some authors, most recently Korn and Graubard (1999), have recommended investigating selection bias by testing the null hypothesis $\boldsymbol{B} - \boldsymbol{\beta} = \mathbf{0}$, using a statistic of the form

$$(\hat{\beta}_m - \hat{B})'[\hat{\boldsymbol{V}}(\hat{\beta}_m - \hat{B})]^{-1}(\hat{\beta}_m - \hat{B})$$

where $\hat{\beta}_m$ is an unweighted estimate corresponding to a working model for which the design is assumed to be ignorable. The idea is that the (more efficient?) working model will be used unless the hypothesis is rejected.

- This testing strategy has several shortcomings

  – convergence of design based and model based parameter estimates does not guarantee that the working model correctly captures the variance structure;

  – as a result, a working model may appear to be more efficient simply because it is ignoring the clustering;

- to have a reasonable chance of achieving design ignorability, important design variables may be needed in the model, which might include stratum indicators, cluster indicators and measures of size. For stratified multi-stage cluster samples, this is a formidable task which may not be feasible, or which may result in an un-interpretable model;

- in any comprehensive analysis project, many models will need to be estimated. If unweighted analyses are favoured, ignorability tests will need to be carried out in each case. In some cases, unweighted analyses will be deemed acceptable, in others design based analyses will be required. The resulting mix of methodologies will be unappealing, at best.

- finally, accepting a null hypothesis is not a good decision strategy, particularly when efficiency problems are greatest for smaller sample sizes, when test power will be at a minimum.

Example

- **Source:**

  Sharan, Kamal K. (2000). Provincial earnings differences. *Perspectives on Labour and Income*, Catalogue no. 75-001-XPE, Statistics Canada.

- **Research problem:**

  What causes the differences in average earnings among all paid workers in Ontario and Québec?

  Could the differences in the demographic and socio-economic composition of the provinces cause the differences in average earnings?

  Or, the differences come from the different ways workers with similar attributes are compensated?

- **Data**
  SLID 98
  sample size 34,918, for analysis 16,497

Variables:

Total individual earnings in 1998
PROVince of residence =1 if Ontario
2 if Québec,
SEX (1 if male; 2 if female),
EDUCation= 1 if less than high school;
2 if graduated high school;
3 if non-university post secondary certificate;
4 if university degree or certificate;
5 if refusal/do not know
AGE in years
Interaction of PROV4 with each of the other variables

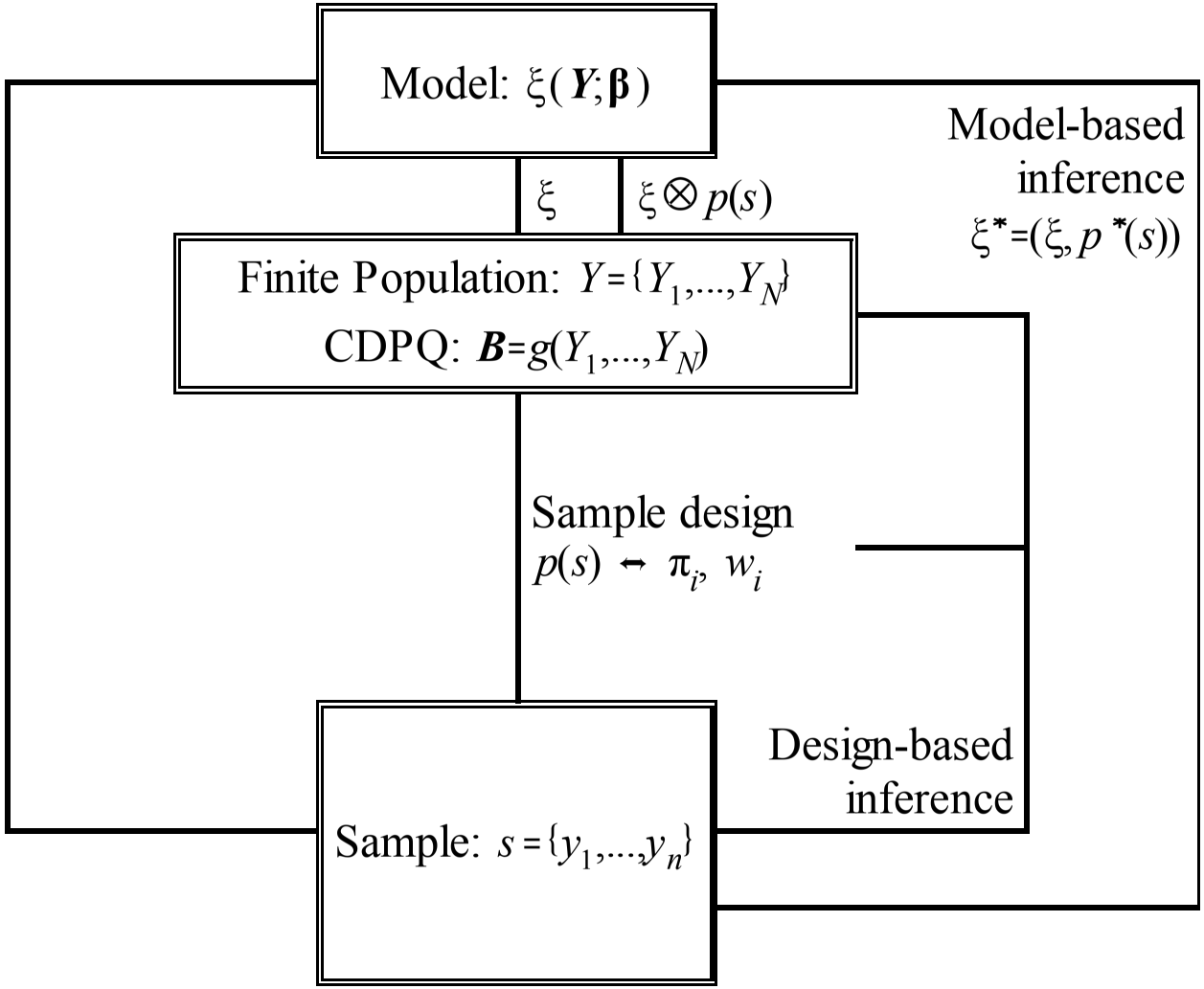
- **Method:**
  A linear regression of log earnings is performed on various demographic and socio-economic predictors

- **Software:**
  - SUDAAN, PROC REGRESS, 1000 bootstrap replicates used for variance estimation
  - SAS, PROC REG, OLS, WLS estimation

	Coefficients	SUDAAN		SAS, WLS		SAS, OLS		
		S.E	p-value	S.E.	p-value	Coefficient	S.E.	p-value
Intercept	**7.9567**	0.0693	0.00	0.03875	<.000	7.9551	0.03888	<.0001
PROV4 (1)								
2	-0.0777	0.1111	0.4848	0.06522	0.2337	0.0034	0.06403	0.9576
GENDER(1)								
2	**0.5017**	0.0300	0.0000	0.01857	<.0001	0.5378	0.01864	<.0001
EDUC(1)								
2	**0.4330**	0.0473	0.0000	0.02916	<.0001	0.4071	0.02920	<.0001
3	**0.7022**	0.0445	0.0000	0.02990	<.0001	0.6760	0.02918	<.0001
4	**1.0349**	0.0500	0.0000	0.03178	<.0001	1.0042	0.03255	<.0001
5	**0.4882**	0.0790	0.0000	0.04298	<.0001	0.3622	0.04681	<.0001
AGE	**0.0316**	0.0013	0.0000	0.00078	<.0001	0.0314	0.00078	<.0001
PROV4, GENDE								
2, 2	-0.0779	0.0441	0.0775	0.02995	**0.009**	-0.0529	0.02987	0.0765
PROV4, EDUC								
2, 2	**-0.1702**	0.0738	0.0213	0.04615	0.0002	-0.1423	0.04565	0.0018
2, 3	**-0.1399**	0.0648	0.0311	0.04614	0.0024	-0.1309	0.04473	0.0034
2, 4	-0.0718	0.0724	0.3213	0.05089	0.1581	-0.0680	0.05113	0.1836
2, 5	0.0135	0.1360	0.9207	0.07906	0.8640	0.0278	0.08295	0.7373
PROV4, AGE								
2,1	0.0023	0.0022	0.2819	0.00129	0.0711	-0.0007	0.00128	0.5652
Number of significant parameters 0.05/0.01			9 7		10 10			9 9

SUMMARY

General framework (**Pffeferman, 1996)**



	Assumed model is valid	**Model is misspecified**
Model-based	- Consistent - Efficient - Valid variance estimates - Valid inference - May be best	- May be inconsistent - Variance estimates may be invalid - Inference may be invalid
Design-based	- Consistent - May be inefficient - Valid variance estimates - Valid inference	- Consistent for model parameter - Valid conditional variance estimates - Valid estimates of total variance - Valid inference

Analytical methods accommodated by software

	SAS SVY	Stata 6	SUDAAN 7.5/8	WesVar 3.0 /4
Means, proportions, totals, ratios; associated se's	Y	Y	Y	Y
Quantiles			Y	Y
Tests of Independence		Y	Y(-)	Y
Contrasts	Y	Y	Y	Y
Linear regression	Y	Y	Y	Y
Logistic regression		Y	Y	Y
Multinomial logistic - ordinal and nominal categories		Y	Y	
Proportional hazards model			Y	
Probit model		Y		
Instrumental variables regression, censored and interval regression, Poisson regression		Y		
Graphical diagnostics	Y(-)	Y(-)		