COMPLEX SAMPLING DESIGN BASED INFERENCE ON FAMILIAL MODELS FOR COUNT DATA

LAUREN IRENE GRANTER

Complex Sampling Design Based Inference on Familial Models for Count Data

by

©Lauren Irene Granter

A practicum submitted to the School of Graduate Studies
in partial fulfillment of the requirement for the Degree
of
Master of Applied Statistics

Department of Mathematics and Statistics
Memorial University of Newfoundland

St. John's Newfoundland, Canada

August 2007

Abstract

Consistent and efficient estimation of the parameters of generalized linear mixed models (GLMMs) has proven to be difficult in the infinite population setup. This estimation issue becomes more complex in the infinite population setup where the estimation is done based on a sample of a small number of clusters chosen from a finite population with a large number of unequally sized clusters. This practicum examines the role of the sampling designs on the estimation of the parameters of the GLMM based super-population for clustered count data.

Acknowledgements

First and foremost, I thank the Lord for giving me the courage and strength to pursue this opportunity. Truely, without His guidance, this would never have been possible.

I would also like to extend the greatest thanks to my supervisor, Dr. Sutradhar, for his continued support, both academically and financially. I am very grateful for his commitment throughout the entire process. It was a privilege to work under his supervision. I cannot forget to thank my co-supervisor, Dr. Alwell Oyet, as well.

To my parents, whose love and encouragement helped overcome the many obstacles that were thrown my way during these three years. To my partner, Sheldon, I don't know how I would have gotten through the last few months without you. I also express the greatest gratitude to Elizabeth Jones, who was my employer outside the department for the duration of this practicum. Thank you for being so flexible and understanding of my hectic schedule.

Lastly, I would like to thank the department of Mathematics and Statistics for their financial support and all the faculty and staff for their help. Indeed, you have all made the process much more enjoyable and manageable.

Contents

List of Tables

Chapter 1

Introduction

1.1 Motivation for Complex Sampling Based Inferences

There exists a vast literature in cluster sampling (Cochran (1977, Chapter 9)) where it may be of interest to estimate the finite population total. In this cluster sampling setup, one deals with a suitable sample of clusters chosen from the finite population consisting of a large number of unbalanced clusters. Consequently, to reflect the unequal cluster sizes, traditionally, a probability proportional to size (PPS) sampling scheme is used to choose the sample, and the totals of the clusters of that sample, along with the PPS weights, are exploited for the unbiased, and hence, consistent estimation of the finite population total. For example, the number of visits to a physician paid by a member of a family can be considered as a discrete random variable and the health department may be interested in estimating the total number of visits paid by the family members in a city. For this purpose, a sample of families may be chosen using PPS sampling and the total number of visits for each family in the sample, along with their family sizes, may be exploited.

It may, however, be the case that, as opposed to the number of visits, the researcher/health department may be interested in knowing the effects of certain covariates on the number of visits paid by the individual members. For example, the number of visits may be affected by gender, age, and the number of chronic conditions of the individual. One may be interested in estimating the effects of these covariates in a finite population setup. Since the visits by the family members are correlated due to a common family effect, the estimation of the effects of the covariates becomes complicated. In fact, there does not exist adequate discussion on such a clustered regression problem in the finite population setup.

We must, however, note that as opposed to the finite population setup, there exist some studies to deal with the clustered regression problem in the infinite population setup. Breslow and Clayton (1993), for example, discuss a penalized quasilikelihood (PQL) method of estimation for a generalized linear mixed model (GLMM). This PQL approach may, however, not produce consistent estimates for the variance of the cluster effects. See, for example, Sutradhar and Qu (1998) and Jiang (1998). Jiang (1998) proposes a method of simulated moments (SMM) to obtain consistent estimates for such a variance parameter under a GLMM setup. However, this estimate, along with the regression effects, may be inefficient. Sutradhar (2004) provides an improvement over this method of moments using an exact generalized quasilikelihood (GQL) approach. Nevertheless, when one deals with this regression problem in a finite population setup, it becomes time and cost effective to estimate the parameters of the model, based on a suitable sample of clusters. This motivated us to consider a sample of K clusters/families from a population with a large number, N, of unbalanced families, and use the GQL idea of Su-

tradhar (2004) in such a finite population setup. To be a little more specific, a sample chosen based on the PPS sampling scheme would be used for the estimation of the parameters of the clustered count data model in the finite population setup.

1.2 Objective of the Practicum

In Chapter 2, we give a brief overview of the Simple Random Sampling (SRS) and Probability Proportional to Size (PPS) sampling schemes. These two sampling schemes are compared using a simulation study to estimate the finite population totals in a non-regression setup. Chapter 3 deals with a traditional clustered regression model, or a familial mixed model for count data, for the infinite population setup. For the estimation of the parameters, namely, the regression effects and variance component of the model, the performance of a recent generalized quasilikelihood (GQL) technique is examined through a simulation study. In Chapter 4, we continue to deal with a clustered regression model for count data, but unlike Chapter 3, we consider a finite population setup. The effects of the two sampling schemes, SRS and PPS, on the GQL estimation of the parameters of a familial mixed model are examined using another set of simulation studies. We conclude the practicum with a summary and suggestions for future studies and improvements in Chapter 5.

Chapter 2

Estimation of Finite Population Total

There is a vast literature in single stage cluster sampling, where the population total is consistently estimated by using a suitable weighted average, where the weights are chosen based on the well-known sampling scheme Probability Proportional to Size (PPS). We refer to Cochran (1977, chapter 9) and the references therein for such an estimation.

Let there be N clusters in the population. Let y_i $(i = 1, \ldots, N)$ denote the ith cluster total and m_i $(i = 1, \ldots, N)$ denote the corresponding cluster size. Suppose that we are interested in estimating the finite population total, $Y = \sum_{i=1}^{N} y_i$, based on a sample of size n. Also suppose that s_r^* is a set of responding units containing n responses. Note that under the clustered population, it is customary to choose this set s_r^* using the Probability Proportional to Size (PPS) technique. This is because the PPS technique uses proper weights based on the cluster size in selecting the population unit. Furthermore, it is well-known that the application of the Simple Random Sampling (SRS) technique would produce an unbiased estimate, but with poor precision, in the clustered population case. Nevertheless, in this section, we conduct a simulation study to examine the relative performance of

the PPS and SRS sampling techniques where the clustered data are generated from a suitable distribution for counts such as the Poisson distribution. This will primarily be done to understand the small sample relative performance of the estimation, which is not adequately addressed in the literature.

For the purpose of the simulation-based comparisons, we provide the formulae for SRS and PPS scheme-based estimation of the totals.

2.1 SRS versus PPS Estimation

SRS Estimation Our population of interest consists of N pairs of information $(m_1, y_1), \ldots, (m_i, y_i), \ldots, (m_N, y_N)$. Here, the m_i's $(i = 1, \ldots, N)$ are assumed to be known, m_i being the size of the ith cluster. However, the y_i's are unknown. It is of interest to estimate $Y = \sum_i^N y_i$. m_i's are considered to be the same for all i, or when m_i's are ignored, one may use the SRS technique to estimate the population total, $\hat{Y} = \sum_{i \in s_r^*} y_i$. To be specific, Y is estimated based on the SRS technique using the formula

$$\hat{Y}_{SRS} = \sum_{i \in s_r^*} w_i y_i, \tag{2.1}$$

where $w_i = N/n$ is the constant weight for the population to be included in the sample, and $y_1, \ldots, y_i, \ldots, y_n$ are known in the sample, s_r^*. This estimator has the variance given by

$$V(\hat{Y}) = \frac{N^2}{n}(1 - f)S_r^2, \tag{2.2}$$

with $f = n/N$ and $S_r^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N-1}$. This variance may be estimated by

$$v(\hat{Y}) = \frac{N^2}{n}(1-f)s_r^2, \tag{2.3}$$

as given in Cochran (1977, Chapter 9), where $s_r^2 = \sum_{i \in s_r^*}(y_i - \bar{y})^2/n - 1$.

We remark that the estimator of Y given in (2.1) is a simple estimator. The estimate $\hat{Y}$, however, will often be of poor precision, as pointed out by Cochran (1977, Chapter 9). This lack of precision will occur when the means per element, $\bar{y}_i = \sum_{j=1}^{m_i} y_{ij}/m_i = y_i/m_i$ do not vary much from cluster to cluster, while the m_i vary greatly. Thus, y_i in (2.1) will vary more from cluster to cluster, and so, the variance given in (2.2) will be large, causing $\hat{Y}$ in (2.1) to be inefficient.

PPS Estimation In an effort to remedy the precision deficiency of the SRS estimator, there exists an alternative sampling technique known as PPS sampling, where the y_i in the sample of size n is chosen with replacement from the population based on weights proportional to the cluster size. For $i = 1, \ldots, N$, we define $z_i = m_i/m_o$, where $m_o = \sum_{i=1}^{N} m_i$. We now let s_p^* denote the sample of size n based on the PPS scheme. Under this PPS scheme, $\sum_{i=1}^{N} y_i$ is estimated by

$$\hat{Y}_{ppz} = \frac{1}{n}\sum_{i \in s_p^*}\frac{y_i}{z_i}, \tag{2.4}$$

which has the variance

$$V(\hat{Y}_{ppz}) = \frac{1}{n}\sum_{i=1}^{N} z_i\left(\frac{y_i}{z_i} - Y\right)^2. \tag{2.5}$$

Furthermore, the variance in (2.5) is usually estimated by

$$v(\hat{Y}_{ppz}) = \sum_{i \in s_p^*} (\frac{y_i}{z_i} - \hat{Y}_{ppz})^2 / n(n-1). \tag{2.6}$$

2.2 A Simulation Study

In this simulation study, we consider populations with N clusters ($N = 100, 200, 300$). Suppose that we assign m_i as follows:

$$m_i = \begin{cases} 4 & \text{for } i = 1, \ldots, \frac{N}{4} \\ 6 & \text{for } i = \frac{N}{4} + 1, \ldots, \frac{3N}{4} \\ 5 & \text{for } i = \frac{3N}{4} + 1, \ldots, N \end{cases} \tag{2.7}$$

Next, suppose that y_{ij} denotes the count response for the jth member ($j = 1, \ldots, m_i$) of the ith cluster, and $y_i = \sum_{j=1}^{m_i} y_{ij}$ denotes the cluster total so that $y_1, \ldots, y_i, \ldots, y_N$ denotes N totals in the finite population. To generate y_i, we consider that $y_{ij} \sim Poi(\mu_i)$, where we choose μ_i, for example, as

$$\mu_i = \begin{cases} 2 & \text{for } i = 1, \ldots, \frac{N}{4} \\ 4 & \text{for } i = \frac{N}{4} + 1, \ldots, \frac{3N}{4} \\ 3 & \text{for } i = \frac{3N}{4} + 1, \ldots, N \end{cases} \tag{2.8}$$

It is clear from (2.7) and (2.8) that the count responses will have the infinite population total $Y_I = \sum_{i=1}^{N} m_i \mu_i$, yielding $Y_I = 1775, 3550, 5324$ for $N = 100, 200, 300$, respectively. For the construction of the population under the simulation study, let y_i be generated from $Poi(\mu_i)$. Consequently, in the simulation study, $y_1, \ldots, y_N$ are known, whereas, in practice, they are unknown. For convenience, we have computed $Y = \sum_{i=1}^{N} y_i$ (finite population total) from the simulated observations. It was found that the values of Y are very close to the infinite population total Y_I, for all $N = 100, 200, 300$, as ex-

pected. It is of interest to estimate the total, Y, based on samples chosen by SRS and PPS techniques. As far as the sample size is concerned, we consider $n = 10, 20, 30, 50$ when $N = 100$. Similarly, we choose $n = 10, 20, 30, 50, 100$ for $N = 200$, and $n = 10, 20, 30, 50, 100, 150$ when $N = 300$.

To obtain SRS and PPS estimates of Y, we have written a Fortran program to generate the sample of size n under each of these sampling schemes and compare the performance of these two techniques in estimating the parameters of the finite population. This program containing both SRS and PPS schemes is given in the appendix.

Next, using formulas (2.1) and (2.4), we have estimated the population total based on SRS and PPS designs, where we denote these estimates by $\hat{Y}_{SRS}$ and $\hat{Y}_{PPS}$ for SRS and PPS, respectively. The simulation means (SM) of these estimates based on 2000 simulations, along with the simulated standard errors (SSE) are reported in Tables 2.1 and 2.2 for two different selections of the values of μ_i $(i = 1, \ldots, N)$. Specifically, the simulation results with μ_i defined as in (2.8) are reported in Table 2.1, whereas the results with $\mu_i = \mu = 3$ for all $i = 1, \ldots, N$ are exhibited in Table 2.2. Note that in Tables 2.1 and 2.2 we also report the simulated bias computed by

$$Bias = \frac{1}{2000} \sum_{s=1}^{2000} |\hat{Y}_s - Y_s|, \tag{2.9}$$

under both sampling schemes, as well as estimated standard error (ESE) computed by (2.3) and (2.6) for SRS and PPS based designs, respectively.

The results in Table 2.1 show that PPS sampling performs much better than SRS in estimating the finite population total. This holds for all n and N considered in the simulation. For example, for the population size $N = 100$, the finite population total was found to be $Y = 1774$ (which is close to $Y_I = 1775$) based on 2000 simulations. When this total Y was estimated by using a sample of size $n = 10$, SRS produces the estimate 1756, while the PPS sampling based estimate was found to be 1765. Clearly, the PPS estimate is much closer to the population total Y. When the corresponding standard errors are compared, the PPS design produces estimates with smaller standard errors. Also, the biases are smaller for PPS sampling as compared to SRS. We may note that the estimated standard errors for both SRS and PPS sampling appear to work well, as they are found to be close to the corresponding simulated standard errors. Similar interpretation holds for all other results of this table.

It is clear from Tables 2.1 and 2.2 that as n increases, standard errors decrease under both SRS and PPS designs, with similar standard errors under the PPS design. We also observe that the bias appears to be decreasing more under PPS as n increases, as compared to the results for SRS. We also note that, as expected, when the relative sample size n is much smaller as compared to the population size N, both SRS and PPS sampling perform poorly, with PPS performing somewhat better than SRS.

Table 2.1: Simulated Mean (SM), Simulated Standard Errors (SSE), Estimated Standard Errors (ESE) and Simulated Bias (SB) for the estimates of finite population total (Y) using SRS and PPS sampling design (with variable cluster size m_i as in (2.7)) based on 2000 simulations for selected sample size (n) drawn from the population size (N) with unequal means as in (2.8).

N		100		200		300	
$Y_{I,N}$		1775		3550		5324	
n	Quantity	$\hat{Y}_{SRS}$	$\hat{Y}_{PPS}$	$\hat{Y}_{SRS}$	$\hat{Y}_{PPS}$	$\hat{Y}_{SRS}$	$\hat{Y}_{PPS}$
10	SM	1755.57	1765.06	3171.05	3548.39	3521.66	5338.29
	SSE	259.76	189.38	590.00	376.94	743.45	561.51
	ESE	248.53	181.81	557.66	359.53	692.90	538.71
	SB	165.10	97.68	806.70	316.53	1825.60	456.10
20	SM	1767.08	1776.95	3149.24	3548.52	3531.16	5316.78
	SSE	186.52	137.97	406.86	266.83	507.78	394.31
	ESE	177.08	128.55	396.81	257.85	506.06	386.94
	SB	186.25	126.09	505.05	292.97	1726.35	269.68
30	SM	1758.62	1771.26	3167.57	3545.91	3544.23	5323.64
	SSE	151.49	115.25	337.18	222.06	444.32	332.57
	ESE	146.20	105.43	325.57	211.76	415.54	316.08
	SB	124.74	64.11	439.07	184.84	1756.35	216.00
50	SM	1765.70	1769.62	3158.17	3549.59	3516.09	5318.42
	SSE	122.38	92.14	272.74	180.02	336.49	259.90
	ESE	112.56	81.75	252.95	164.96	321.84	246.04
	SB	95.46	70.76	400.72	132.87	1854.88	180.79
100	SM			3155.62	3554.28	3541.45	5318.42
	SSE			201.92	130.65	252.34	197.14
	ESE			178.82	116.30	229.70	174.85
	SB			374.58	102.31	1819.25	136.63
150	SM					5322.63	5328.17
	SSE					215.95	163.32
	ESE					185.93	142.39
	SB					1802.49	111.49

As is the case with Table 2.1, Table 2.2 shows that PPS performs much better than SRS in the estimation of the finite population total under equal means. This happens because the PPS design chooses more clusters of larger size in the sample. For example, for $N = 200$, we have an infinite population total of $Y_I = 3150$. For $n = 30$, the PPS scheme gives the closer estimate of 3149, while SRS gives an estimate of 2980. However, when n increases, the estimates do not seem to get much better under either of the two sampling schemes as was the case for Table 2.1. This shows that small sample size works reasonably well, but when the standard errors are compared, they decrease as n increases. This shows that a larger sample size is needed for better performance, as expected. Next, the estimating formulas for the standard errors appear to work well as the estimated standard errors are seen to be close to the simulated standard errors.

In summary, it is clear from both tables that the PPS technique is far superior to SRS in estimating the finite population total. This is because the PPS design is constructed to utilize more information as compared to the SRS design.

Table 2.2: Simulated Mean (SM), Simulated Standard Errors (SSE), Estimated Standard Errors (ESE) and Simulated Bias (SB) for the estimates of finite population total (Y) using SRS and PPS sampling design (with variable cluster size m_i as in (2.7)) based on 2000 simulations for selected sample size (n) drawn from the population size (N) with equal means $\mu_i = \mu = 3$.

N		100		200		300	
$Y_{I,N}$		1573		3150		4725	
n	Quantity	$\hat{Y}_{SRS}$	$\hat{Y}_{PPS}$	$\hat{Y}_{SRS}$	$\hat{Y}_{PPS}$	$\hat{Y}_{SRS}$	$\hat{Y}_{PPS}$
10	SM	1557.85	1574.20	2980.04	3148.83	3985.56	4734.30
	SSE	157.98	130.79	336.44	254.58	450.22	394.96
	ESE	147.51	121.92	304.81	240.58	410.32	361.92
	SB	123.90	51.98	403.60	173.30	730.00	348.76
20	SM	1563.51	1575.70	2964.19	3148.82	3981.54	4729.23
	SSE	115.45	98.82	235.73	183.30	321.63	271.72
	ESE	105.23	87.00	219.42	174.15	302.28	263.43
	SB	104.60	70.16	257.60	177.71	694.00	180.65
30	SM	1561.16	1573.85	2972.46	3147.25	4002.74	4729.23
	SSE	95.53	83.36	198.47	154.40	284.02	226.49
	ESE	86.59	71.09	180.45	143.50	244.13	214.94
	SB	79.17	51.15	201.17	86.60	674.33	160.32
50	SM	1563.82	1573.26	2968.78	3149.22	3983.75	4718.73
	SSE	79.78	69.02	204.84	96.51	217.59	186.01
	ESE	67.34	55.49	140.62	111.89	191.45	165.69
	SB	54.50	45.58	204.84	96.51	774.44	133.29
100	SM			2965.89	3153.09	4001.52	4720.92
	SSE			128.00	97.08	174.09	143.35
	ESE			99.64	79.00	136.24	118.47
	SB			173.16	63.40	725.52	100.05
150	SM					3990.17	4722.68
	SSE					163.16	122.33
	ESE					110.54	96.58
	SB					729.05	69.98

Chapter 3

Estimation of Parameters of the Poisson Mixed Model: Infinite Population Setup

In the previous chapter, we estimated the population total for a finite number of clusters, where cluster elements were assumed to follow the Possion distribution. But, in practice, we may be interested in regression effects of certain covariates on the count responses recorded under a large number of clusters. For example, we may be interested in the effects of the covariates age, gender, and education level on the count responses such as the number of visits to the physician by the members of a large number of independent families.

Our purpose is to make inferences for these types of regression effects in the finite population setup. However, there exist many examples for this type of inference in the infinite population setup. For convenience, in this chapter, we briefly review the non-linear regression analysis that other researchers have done for familial count data.

Familial Mixed Model for Count Data Let there be K independent clusters/families randomly selected from an infinite population. Note that in

the previous chapter, for convenience, we used "n" instead of K and n clusters were chosen from N finite clusters. Recall that y_{ij} is the count response for the jth ($j = 1, \ldots, m_i$) member of the ith family. We now consider $x_{ij} = (x_{ij1}, \ldots, x_{ijm_i})^T$ as the p-dimensional regression vector corresponding to y_{ij}. Also suppose that $\beta = (\beta_1, \ldots, \beta_n, \ldots, \beta_p)^T$ be the p-dimensional effects of the variable x_{ij} on y_{ij}. In this setup, it is common to use the Poisson distribution y_{ij} for the random counts, that is,

$$f(y_{ij}) = \frac{e^{-\mu^*_{ij}} \mu^{*y_{ij}}_{ij}}{y_{ij}!}, \tag{3.1}$$

where $\mu^*_{ij} = exp(x^T_{ij}\beta + \gamma^*_i)$ with γ^*_i representing the ith family effect. As far as the distribution of γ^*_i is concerned, it is reasonable to use the Gaussian distribution. See, for example, Breslow and Clayton (1993), Jiang (1998), Sutradhar and Qu (1998), and Sutradhar (2004). Thus, we assume that $\gamma^*_i \overset{iid}{\sim} N(0, \sigma^2_\gamma)$. It is of interest to obtain consistent and efficient estimates for the regression effect β and the variance component of the random effects σ^2_γ.

Note that the responses $y_{i1}, \ldots, y_{im_i}$ in the ith family are correlated. This is because the family members share the common effect γ^*_i. To be specific, the mean, variance, and the correlation structure under this model are given by

$$\begin{aligned} \mu_{ij} = E(Y_{ij}) &= E_{\gamma^*_i} E[Y_{ij}|\gamma^*_i] \\ &= \mathrm{E}_{\gamma^*_i}[\mu^*_{ij}] \\ &= e^{x^T_{ij}\beta + \frac{\sigma^2_\gamma}{2}}, \end{aligned} \tag{3.2}$$

$$\begin{aligned}\sigma_{ijj} = \mathrm{Var}(Y_{ij}) &= E(Y_{ij}^2) - \mu_{ij}^2 \\ &= E_{\gamma_i^*}[\mu_{ij}^{*2} + \mu_{ij}^*] - \mu_{ij}^2 \\ &= \mu_{ij} + (e^{\sigma^2} - 1)\mu_{ij}^2, \end{aligned} \tag{3.3}$$

and

$$\begin{aligned}\sigma_{ijk} = \mathrm{Cov}(Y_{ij}, Y_{ik}) &= E(Y_{ij}Y_{ik}) - \mu_{ij}\mu_{ik} \\ &= \mathrm{E}_{\gamma_i^*}\mathrm{E}[Y_{ij}Y_{ik}|\gamma_i^*] - \mu_{ij}\mu_{ik} \\ &= \mathrm{E}_{\gamma_i^*}[\mu_{ij}^*\mu_{ik}^*] - \mu_{ij}\mu_{ik} \\ &= (e^{\sigma^2} - 1)\mu_{ij}\mu_{ik}, \end{aligned} \tag{3.4}$$

respectively. Note that the formulas in (3.2)-(3.4) were obtained by using the fact that for a positive integer t and for $\gamma_i^* \overset{iid}{\sim} N(0, \sigma_\gamma^2)$, we have

$$E(e^{t\gamma_i^*}) = e^{\frac{1}{2}t^2\sigma_\gamma^2}.$$

3.1 Some Remarks on Likelihood Estimation

With respect to the inference for β and σ_γ^2, one may attempt to use the traditional likelihood approach, which, however, is extremely complicated. This is because, in the present case, the log likelihood may be written as

$$\log L(\beta, \sigma_\gamma) = -\sum_{i-1}^{K}\sum_{j-1}^{m_i} y_{ij}! + \sum_{i=1}^{K}\sum_{j=1}^{m_i} y_{ij} x_{ij}^T \beta + \sum_{i=1}^{K} \log J_i,$$

with

$$J_i = \int_{-\infty}^{\infty} \exp[d_i(\gamma_i)]\phi(\gamma_i) d\gamma_i,$$

where for $\gamma_i = \gamma_i^*/\sigma_\gamma$, $\phi(\gamma_i)$ is the standard normal density, and

$$d_i(\gamma_i) = \sigma_\gamma \gamma_i \sum_{j=1}^{m_i} y_{ij} - \sum_{j=1}^{m_i} \exp(x_{ij}^T \beta + \sigma_\gamma \gamma_i).$$

The estimation of β and σ_γ^2 requires the solution of the likelihood estimating equations

$$\frac{\partial \log L(\beta, \sigma_\gamma^2)}{\partial \beta} = 0 \quad \text{and} \quad \frac{\partial \log L(\beta, \sigma_\gamma^2)}{\partial \sigma_\gamma^2} = 0, \tag{3.5}$$

which appear to be complicated because of the computational difficulties for the second order derivatives. Furthermore, finding the covariance matrix of the estimators will be more complicated because of the difficulty in computing the Fisher Information matrix (Sutradhar and Das (2001, eq.(4.8))). This complexity increases when one considers multi-dimensional random effects.

For these reasons, many authors such as Breslow and Clayton (1993) have avoided the use of the exact likelihood approach. Specifically, these authors

have used a penalized quasi-likelihood (PQL) approach. Jiang (1998) has used a simulated moment approach for the estimation of β and σ_γ^2. Jiang (1998) has adopted this moment approach because of the inconsistency problem encountered by the PQL approach in estimating σ_γ^2. See also Sutradhar and Qu (1998). Later, Sutradhar and Rao (2003) and Sutradhar (2004) have proposed a generalized quasi-likelihood (GQL) approach for the estimation of the parameters. This approach, unlike the moment approach of Jiang (1998) and of Jiang and Zhang (2001), provides both consistent and efficient estimates for β and σ_γ^2. In this chapter, we follow Sutradhar (2004) and verify the perfomance of the GQL approach for such an estimation through a simulation study in the infinite population setup. We do this for the main purpose of using and examining the performance of the GQL approach in the finite population setup which will be provided in the next chapter.

3.2 GQL Estimation for the Regression Effects

In this section, we will follow the Generalized Quasi-Likelihood (GQL) approach (Sutradhar (2004)) for the estimation of the β parameter. This provides a consistent as well as efficient estimate for β, provided one can estimate the σ^2 parameter consistently. For consistent estimation of σ^2, we use the method of moments (Jiang (1998)) and provide the estimation formula in the next section. Note that we also could use the GQL approach to efficiently estimate σ^2, but we have chosen the moment estimation approach instead, as we are much more interested in estimating the main regression parameter, β. Also note that the moment approach is much simpler than GQL.

For the GQL estimation of β, we write the estimating equation as follows:

$$\sum_{i=1}^{K} \frac{\partial \mu_i^T}{\partial \beta} \Sigma_i^{-1} (y_i - \mu_i) = 0, \tag{3.6}$$

where $y_i = (y_{i1}, y_{i2}, \ldots, y_{im_i})^T$, $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{im_i})^T$, and Σ_i is the covariance matrix of y_i. In notation,

$$\Sigma_i = \begin{pmatrix} \sigma_{i11} & \sigma_{i12} & \cdots & \sigma_{i1m_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{ij1} & \sigma_{ij2} & \cdots & \sigma_{ijm_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{im_i1} & \sigma_{im_i2} & \cdots & \sigma_{im_im_i} \end{pmatrix}, \tag{3.7}$$

where the formulas for σ_{ijj} and σ_{ijk} $(j \neq k)$ are given in (3.3) and (3.4), respectively. For compuational convenience, we simplify the formula for the derivative involved in (3.6) as

$$\frac{\partial \mu_i^T}{\partial \beta} = X_i^T A_i, \tag{3.8}$$

where X_i and A_i are defined as

$$X_i = \begin{pmatrix} x_{i11} & x_{i12} & \cdots & x_{i1p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{ij1} & x_{ij2} & \cdots & x_{ijp} \\ \vdots & \vdots & \ddots & \vdots \\ x_{im_i1} & x_{im_i2} & \cdots & x_{im_ip} \end{pmatrix}, \quad A_i = \begin{pmatrix} \mu_{i1} & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & & & \vdots \\ \vdots & & \mu_{ij} & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \mu_{im_i} \end{pmatrix} \tag{3.9}$$

By using (3.8) in (3.6), we obtain the GQL estimating equation for β given by

$$\sum_{i=1}^{K} X_i^T A_i \Sigma_i^{-1} (y_i - \mu_i) = 0. \tag{3.10}$$

We can now solve this estimating equation for β by using the well-known Newton-Raphson iteration technique. The iterative formula is given by

$$\hat{\beta}_{new} = \hat{\beta}_{old} + \left[\sum_{i=1}^{K} X_i^T A_i \Sigma_i^{-1} A_i X_i \right]^{-1} \sum_{i=1}^{K} X_i^T A_i \Sigma_i^{-1} (y_i - \mu_i). \tag{3.11}$$

3.3 Moment Estimation for the Variance Component

Following Jiang (1998) [see also Sutradhar (2004)], we use the second order responses to construct a moment estimating formula for σ^2. To be specific, we consider

$$S = \sum_{i=1}^{K} \sum_{j=1}^{m_i} (y_{ij} - \mu_{ij})^2 + \sum_{i=1}^{K} \sum_{j<u}^{m_i} (y_{ij} - \mu_{ij})(y_{iu} - \mu_{iu}) \tag{3.12}$$

as a basic statistic and find its expectation as

$$\begin{aligned} \mathrm{E}(S) &= \sum_{i=1}^{K}\sum_{j=1}^{m_i}\mathrm{E}(y_{ij}-\mu_{ij})^2 + \sum_{i=1}^{K}\sum_{j<u}^{m_i}\mathrm{E}\{(y_{ij}-\mu_{ij})(y_{iu}-\mu_{iu})\} \\ &= \sum_{i=1}^{K}\sum_{j=1}^{m_i}\mu_{ij} + (e^{\sigma^2}-1)\sum_{i=1}^{K}\sum_{j=1}^{m_i}\mu_{ij}^2 + \sum_{i=1}^{K}\sum_{j<u}^{m_i}e^{\sigma^2}\mu_{ij}\mu_{iu}, \end{aligned} \tag{3.13}$$

by (3.3) and (3.4). Now, by solving $S - \mathrm{E}(S) = 0$ we obtain

$$e^{\sigma^2} g_1(\beta) = S - g_2(\beta), \tag{3.14}$$

where

$$\begin{aligned} g_1(\beta) &= \sum_{i=1}^{K}\sum_{j=1}^{m_i}\mu_{ij}^2 + \sum_{i=1}^{K}\sum_{j<u}^{m_i}\mu_{ij}\mu_{iu} \\ g_2(\beta) &= \sum_{i=1}^{K}\sum_{j=1}^{m_i}\mu_{ij} - \sum_{i=1}^{K}\sum_{j=1}^{m_i}\mu_{ij}^2. \end{aligned} \tag{3.15}$$

It then follows from (3.15) that the moment estimating formula for σ^2 is given by

$$\hat{\sigma}^2 = \log\frac{S - g_2(\beta)}{g_1(\beta)} \tag{3.16}$$

In the next section, we examine the performance of the GQL estimate of β and the moment estimate of σ^2 obtained by (3.11) and (3.16), respectively, through a simulation study.

3.4 Performance of GQL Estimation: A Simulation Study

For this simulation study, we consider an infinite population and use a sample of K families/clusters ($K = 100, 200$) and with size m_i for the ith cluster, where $i = 1, \ldots, N$. To be specific, we consider

$$m_i = \begin{cases} 2 & \text{for } i = 1, \ldots, \frac{K}{4} \\ 4 & \text{for } i = \frac{K}{4} + 1, \ldots, \frac{3K}{4} \\ 3 & \text{for } i = \frac{3K}{4} + 1, \ldots, K \end{cases} \tag{3.17}$$

As far as the covariates x_{ij} ($i = 1, \ldots, K$, $j = 1, \ldots, m_i$) are concerned, we consider $p = 2$-dimensional covariates as:

$$x_{ij1} = \begin{cases} 0 & \text{for } j = 1 \text{ , } i = 1, \ldots, \frac{K}{4} \\ 1 & \text{for } j = 2 \text{ , } i = 1, \ldots, \frac{K}{4} \end{cases}$$

$$x_{ij1} = \begin{cases} -1 & \text{for } j = 1 \text{ , } i = \frac{K}{4} + 1, \ldots, \frac{3K}{4} \\ -1 & \text{for } j = 2 \text{ , } i = \frac{K}{4} + 1, \ldots, \frac{3K}{4} \\ 1 & \text{for } j = 3 \text{ , } i = \frac{K}{4} + 1, \ldots, \frac{3K}{4} \\ 1 & \text{for } j = 4 \text{ , } i = \frac{K}{4} + 1, \ldots, \frac{3K}{4} \end{cases}$$

$$x_{ij1} = \begin{cases} -1 & \text{for } j = 1 \text{ , } i = \frac{3K}{4} + 1, \ldots, K \\ 0 & \text{for } j = 2 \text{ , } i = \frac{3K}{4} + 1, \ldots, K \\ 1 & \text{for } j = 3 \text{ , } i = \frac{3K}{4} + 1, \ldots, K \end{cases}$$

for the first covariate. The second covariate, x_{ij2}, was randomly generated from the binary distribution for selected i and j as follows:

$$x_{ij2} \sim b(0.3) \text{ for } j = 1, \; i = 1, \ldots, \frac{K}{4}$$

$$x_{ij2} \sim b(0.5) \text{ for } j = 2,\ i = 1, \ldots, \tfrac{K}{4}$$

$$\begin{aligned}
x_{ij2} &\sim b(0.1) \text{ for } j = 1,\ i = \tfrac{K}{4} + 1, \ldots, \tfrac{3K}{4} \\
x_{ij2} &\sim b(0.3) \text{ for } j = 2,\ i = \tfrac{K}{4} + 1, \ldots, \tfrac{3K}{4} \\
x_{ij2} &\sim b(0.5) \text{ for } j = 3,\ i = \tfrac{K}{4} + 1, \ldots, \tfrac{3K}{4} \\
x_{ij2} &\sim b(0.7) \text{ for } j = 4,\ i = \tfrac{K}{4} + 1, \ldots, \tfrac{3K}{4}
\end{aligned}$$

$$\begin{aligned}
x_{ij2} &\sim b(0.2) \text{ for } j = 1,\ i = \tfrac{3K}{4} + 1, \ldots, K \\
x_{ij2} &\sim b(0.5) \text{ for } j = 2,\ i = \tfrac{3K}{4} + 1, \ldots, K \\
x_{ij2} &\sim b(0.8) \text{ for } j = 3,\ i = \tfrac{3K}{4} + 1, \ldots, K
\end{aligned}$$

Thus, we consider $\beta = (\beta_1, \beta_2)^T$. Next, with regard to the random family effects, we consider $\gamma_i^* \overset{iid}{\sim} N(0, \sigma^2)$ with selected values of $\sigma^2 = 0.09, 0.25, 0.5625, 1.00, 1.44$. To generate the data, we consider $\beta_1 = 1.0$ and $\beta_2 = 0.5$. The above values of the covariates, the generated values for γ_i^* and β_1 and β_2, lead to the numerical values for $\mu_{ij}^* = exp(x_{ij}^T\beta + \gamma_i^*)$ as given in (3.1). Next, we generate y_{ij} from the Poisson distribution with mean μ_{ij}^*. That is, y_{ij} conditional on γ_i^* is generated following

$$y_{ij} \overset{iid}{\sim} Poi(\mu_{ij}^*). \tag{3.18}$$

It then follows that, unconditionally, these responses, y_{ij}, for $j = 1, \ldots, m_i$ under the ith family will be correlated following the correlation structure given in (3.4)

All together, we consider 2000 simulations. Under each simulation, we estimate $\beta = (\beta_1, \beta_2)^T$ by using the GQL estimating equation (3.10). To be specific, we solved (3.10) by using (3.11). This has been done for a given value of σ^2. Note, however, that this σ^2 has also to be estimated, as it is

unknown. With the first step values of β_1 and β_2, we now use the moment estimating formula (3.16) to obtain an estimate of σ^2. With this value of σ^2, we go back to obtain an improved estimate for β by using (3.11). These improved values of β_1 and β_2 are in turn used in (3.16) to obtain an improved estimate of σ^2. This constitutes a cycle of iteration. This cycle continues until convergence. Finally, the converged estimates for β_1, β_2, and σ^2 under all 2000 simulations are used to compute the simulated means, which are referred to as the simulated estimates. The simulated standard errors are also computed. These simulated means and standard errors are reported in Table 3.1 for all 5 selected values of σ^2 as well as for $N = 100, 200$.

The results in Table 3.1 show that for both $K = 100$ and $K = 200$, the GQL approach yielded estimates for β_1 that are very close to 1.0 with small standard errors that are quite small. For β_2, however, the GQL approach gave estimates close to 0.5 with slightly larger standard errors. For example, for $\sigma^2 = 0.09$, we have $\beta_1 = 1.0012$, $\beta_2 = 0.4961$. The results are comparable for $K = 200$. However, we do find that as the true value of σ^2 increases, the estimates for σ^2 become less accurate. For example, for $\sigma^2 = 1.44$, we get the following results: $\beta_1 = 0.9940(0.1048)$, $\beta_2 = 0.4815(0.2086)$, and $\sigma^2 = 1.2044(0.5286)$. We see that although the values for β are similar in magnitude as those for $\sigma^2 = 0.09$, the value for the simulated σ^2 is much lower than we would expect to get. The results for $K = 200$ are slightly better with $\sigma^2 = 1.2903(0.4689)$.

Overall, we do see that the estimates for β seem reasonable and seem to be consistent. The estimates for σ^2, although somewhat disappointing for the larger values, are also consistent.

Note that we have examined the performance of the GQL approach for

Table 3.1: Simulated means (SM) and simulated standard errors (SSE) for the estimates of β and σ^2 based on cluster size K (drawn from an infinite population) under the familial regression Poisson mixed model (3.1)-(3.4) using 2000 simulations.

			Estimate		
K	True σ^2	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}^2$
100	0.09	SM	1.0012	0.4961	0.0867
		SSE	0.0640	0.0791	0.0375
	0.25	SM	0.9989	0.4937	0.2400
		SSE	0.0661	0.0889	0.0761
	0.5625	SM	0.9972	0.4876	0.5381
		SSE	0.0677	0.0997	0.1982
	1.0000	SM	0.9962	0.4848	0.9125
		SSE	0.0696	0.1147	0.3728
	1.4400	SM	0.9940	0.4815	1.2044
		SSE	0.1048	0.2086	0.5286
200	0.09	SM	1.0008	0.4978	0.0878
		SSE	0.0461	0.0567	0.0258
	0.25	SM	1.0006	0.4957	0.2459
		SSE	0.0477	0.0636	0.0555
	0.5625	SM	1.0001	0.4921	0.5504
		SSE	0.0484	0.0704	0.1439
	1.0000	SM	0.9973	0.4910	0.9410
		SSE	0.0501	0.0769	0.3123
	1.4400	SM	0.9969	0.4871	1.2903
		SSE	0.0515	0.0818	0.4689

estimates of the parameters for the Poisson mixed model. As pointed out, this approach appears to work well in estimating the parameters involved. This gives us some confidence to also use this approach under the finite population setup. We discuss this in Chapter 4.

Chapter 4

Finite Sampling Based Inference on Poisson Mixed Models

In the previous chapter, we dealt with a Poisson mixed model under infinite population setup. To be specific, a large sample of clusters with unequal sizes was chosen from an infinite population. There, the underlying sample was thought to be selected based on the well known simple random sampling (SRS) technique. Thus, the inference made in the last chapter was completely infinite population-based. In a sample survey setup, however, we deal with a finite population of N clusters, similar to that of Chapter 2. We then make the inferences based on a smaller sample of n clusters chosen by using a suitable sampling technique. In this chapter, we follow this finite population concept, but, as opposed to the non-regression problem in Chapter 2, we now deal with an inference problem in the regression setup.

4.1 Weighted GQL (WGQL) PPS Design Based Estimation

4.1.1 Estimation of β

Note that at the finite population level, the responses under the clusters are practically unknown. If, however, they were known, we could follow (3.6) and use the estimating equation

$$\sum_{i=1}^{N} \frac{\partial \mu_i^T}{\partial \beta} \Sigma_i^{-1} (y_i - \mu_i) = 0, \tag{4.1}$$

to estimate the regression parameter β. We also note that in (4.1), N is the number of clusters in the finite population, whereas in (3.1), K was a number of clusters in a sample chosen from the infinite population.

Because of the unavailability of the responses under each of the N clusters, we now choose a sample of K clusters based on a suitable sampling technique. Here, it is reasonable to use a PPS sampling technique because of the fact that the sizes for all N clusters are known and they may vary from cluster to cluster. To be more specific, the probability of selection of a cluster should depend on its size for efficient information. For the purpose, similar to Chapter 2, we use $z_i = m_i/m_o$ as the probability of the selection of the ith cluster in the sample of size K, and write the estimating equation as

$$\frac{1}{K} \sum_{i \in s_p^*} \frac{1}{z_i} \frac{\partial \mu_i^T}{\partial \beta} \Sigma_i^{-1} (y_i - \mu_i) = 0, \tag{4.2}$$

where s_p^* denotes the sample of size K chosen from the population of size

N clusters. Note that an estimating equation similar to (4.2) was suggested by Binder (1983) for the estimation of the regression effects in the finite population setup for longitudinal clustered data, whereas (4.2) is written for familial clustered data. For some discussion on the importance of using a sampling scheme that incorporates sampling weights for non-clustered data, see, for example, Pfeffermann (1993). Further note that we may also express (4.2) as

$$\frac{1}{K}\sum_{i=1}^{K}\frac{1}{z_i}\frac{\partial\mu_i^T}{\partial\beta}\Sigma_i^{-1}(y_i-\mu_i)=0.$$

The function on the left hand side of (4.2) may estimate the finite population function given in (4.1) unbiasedly, provided we use PPS sampling with replacement. This is because under the "with replacement" scheme, we may express (4.2) as

$$\frac{1}{K}\sum_{i=1}^{N}\frac{t_i}{z_i}\frac{\partial\mu_i^T}{\partial\beta}\Sigma_i^{-1}(y_i-\mu_i)=0, \tag{4.3}$$

where $t_i=0,\ldots,K$ and follows the joint multinomial distribution

$$P(t_1,t_2,\ldots,t_N)=\frac{K!}{t_1!t_2!\cdots t_N!}z_1^{t_1}z_2^{t_2}\cdots z_N^{t_N}, \tag{4.4}$$

and it follows that $E(t_i)=Kz_i$. Now, if we take the expectation of the left hand side of (4.3) over t_i, we arrive at the expression given on the left hand side of (4.1) as illustrated below

$$E(\frac{1}{K}\sum_{i=1}^{N}\frac{t_i}{z_i}\frac{\partial\mu_i^T}{\partial\beta}\Sigma_i^{-1}(y_i-\mu_i)) = \sum_{i=1}^{N}\frac{\partial\mu_i^T}{\partial\beta}\Sigma_i^{-1}(y_i-\mu_i) \tag{4.5}$$

It is, therefore, clear that to estimate β based on the PPS sampling technique, we can use the estimating equation (4.2). In the simulation study in section 4.3, we will examine the performance of the PPS sampling based WGQL estimator for β to be obtained from (4.2).

4.1.2 Weighted MM (WMM) for σ^2

We can follow the same methodology as given previously in section 3.3 to find the weighted moment estimate for σ^2. This estimate will be denoted as $\hat{\sigma^2_w}$. We start with the equation given in (3.12), and modify it by accommodating the sampling weights similar to that of the last subsection. So, we write

$$S_w = \sum_{i\in s_p^*} w_i \sum_{j=1}^{m_i}(y_{ij}-\mu_{ij})^2 + \sum_{i\in s_p^*} w_i \sum_{j<u}^{m_i}(y_{ij}-\mu_{ij})(y_{iu}-\mu_{iu}), \tag{4.6}$$

where under PPS sampling with replacement, we have $w_i = 1/z_i$ with $z_i = m_i/m_o$ as given in Chapter 2. Following (3.3) and (3.4), we obtain the model-based expectation as

$$\mathrm{E}(S_w) = \sum_{i=1}^{K} w_i \sum_{j=1}^{m_i}\mu_{ij} + (e^{\sigma^2}-1)\sum_{i=1}^{K} w_i \sum_{j=1}^{m_i}\mu_{ij}^2 + \sum_{i=1}^{K} w_i \sum_{j<u}^{m_i} e^{\sigma^2}\mu_{ij}\mu_{iu}. \tag{4.7}$$

By solving $S_w - E(S_w) = 0$ we obtain

$$e^{\sigma^2} g_{1_w}(\beta) = S_w - g_{2_w}(\beta), \tag{4.8}$$

where

$$\begin{aligned} g_{1_w}(\beta) &= \sum_{i=1}^{K} w_i \sum_{j=1}^{m_i} \mu_{ij}^2 + \sum_{i=1}^{K} w_i \sum_{j<u}^{m_i} \mu_{ij}\mu_{iu} \\ g_{2_w}(\beta) &= \sum_{i=1}^{K} w_i \sum_{j=1}^{m_i} \mu_{ij} - \sum_{i=1}^{K} w_i \sum_{j=1}^{m_i} \mu_{ij}^2. \end{aligned} \tag{4.9}$$

This will lead us to the weighted moment estimating formula for σ^2 analogous to that of (3.16). To be specific, the weighted moment estimator is given by

$$\hat{\sigma_w}^2 = \log \frac{S_w - g_{2_w}(\beta)}{g_{1_w}(\beta)}. \tag{4.10}$$

Note that this weighted moment estimator in (4.1) is a consistent estimator. This is because it was obtained by solving the unbiased estimating equation $S_w - E(S_w) = 0$. This estimator, unlike the WGQL estimator of β, may not, however, be highly efficient. By exploiting the WGQL approach for σ^2 estimation, we may obtain a highly efficient estimator which will be more complicated and has been omitted from the present discussion. In Section 4.3, we will also examine the performance of the weighted method of moments (WMM) estimator of σ^2 through a simulation study.

4.2 SRS Design Based Estimation

Under a PPS sampling design, we selected a sample of K clusters from the finite population with N clusters based on the probability for the selection of the ith cluster with $z_i = m_i/m_o$. Now, to choose a simple random sample of size K for the same finite population, we need to use $m_i = m$, say, for all $i = 1, \ldots, N$ so that $z_i = m_i / \sum_{i=1}^{N} m_i = 1/N$. Consequently, when the SRS design is used to choose a sample from an unbalanced clustered population, the weights do not contribute towards the estimation of the parameter. Thus, we can simply use the GQL estimating equation

$$\sum_{i=1}^{K} \frac{\partial \mu_i^T}{\partial \beta} \Sigma_i^{-1} (y_i - \mu_i) = 0$$

to solve for β, which is the same equation as (3.6) in Chapter 3. Similarly, by solving (3.16) we may obtain the SRS estimator for $\hat{\sigma}^2$. Note, however, that there is a big difference between the samples considered here and those of Chapter 3. This is because in Chapter 3, the random sample was taken from an infinite population consisting of clusters with equal size, whereas in this section, we are dealing with SRS from the finite population of N clusters with unequal sizes.

4.3 Relative Performance of the Approaches: A Simulation Study

Our objective is to compare the estimation performance of the GQL estimate for β obtained by (4.2) using the PPS sampling scheme and its counterpart obtained by using SRS. For the comparison of the estimates of σ^2, we

use equation (4.10) under PPS sampling and its counterpart SRS sampling as discussed in subsection 4.2.

We are using the same p=2-dimensional covariates as we did in section 3.4, which are given as

$$x_{ij1} = \begin{cases} 0 & \text{for } j = 1 \text{ , } i = 1, \ldots, \frac{K}{4} \\ 1 & \text{for } j = 2 \text{ , } i = 1, \ldots, \frac{K}{4} \end{cases}$$

$$x_{ij1} = \begin{cases} -1 & \text{for } j = 1 \text{ , } i = \frac{K}{4} + 1, \ldots, \frac{3K}{4} \\ -1 & \text{for } j = 2 \text{ , } i = \frac{K}{4} + 1, \ldots, \frac{3K}{4} \\ 1 & \text{for } j = 3 \text{ , } i = \frac{K}{4} + 1, \ldots, \frac{3K}{4} \\ 1 & \text{for } j = 4 \text{ , } i = \frac{K}{4} + 1, \ldots, \frac{3K}{4} \end{cases}$$

$$x_{ij1} = \begin{cases} -1 & \text{for } j = 1 \text{ , } i = \frac{3K}{4} + 1, \ldots, K \\ 0 & \text{for } j = 2 \text{ , } i = \frac{3K}{4} + 1, \ldots, K \\ 1 & \text{for } j = 3 \text{ , } i = \frac{3K}{4} + 1, \ldots, K \end{cases}$$

for the first covariate. The second covariate was randomly generated from the binary distribution and is given as

$$x_{ij2} \sim b(0.3) \text{ for } j = 1, \; i = 1, \ldots, \tfrac{K}{4}$$
$$x_{ij2} \sim b(0.5) \text{ for } j = 2, \; i = 1, \ldots, \tfrac{K}{4}$$

$$x_{ij2} \sim b(0.1) \text{ for } j = 1, \; i = \tfrac{K}{4} + 1, \ldots, \tfrac{3K}{4}$$
$$x_{ij2} \sim b(0.3) \text{ for } j = 2, \; i = \tfrac{K}{4} + 1, \ldots, \tfrac{3K}{4}$$
$$x_{ij2} \sim b(0.5) \text{ for } j = 3, \; i = \tfrac{K}{4} + 1, \ldots, \tfrac{3K}{4}$$
$$x_{ij2} \sim b(0.7) \text{ for } j = 4, \; i = \tfrac{K}{4} + 1, \ldots, \tfrac{3K}{4}$$

$$x_{ij2} \sim b(0.2) \text{ for } j = 1, \; i = \tfrac{3K}{4} + 1, \ldots, K$$
$$x_{ij2} \sim b(0.5) \text{ for } j = 2, \; i = \tfrac{3K}{4} + 1, \ldots, K$$
$$x_{ij2} \sim b(0.8) \text{ for } j = 3, \; i = \tfrac{3K}{4} + 1, \ldots, K$$

We are also considering the same cluster sizes as given in (3.17). For each simulation, we generate these clusters with size m_i, where $i = 1, \ldots, N$, drawn from a finite population of $N = 200$ and 300 under either the SRS or PPS sampling scheme. For the SRS case, the clusters are selected at random, with replacement, irrespective of their sizes. The estimation of β is done by using the GQL estimation equation (3.6), and σ^2 is estimated by the method of moments using (3.16). Under the SRS scheme, the results for the GQL estimation of the β parameter and the MM estimation of the σ^2 parameter are displayed in Tables 4.1 and 4.2 for $N = 200$ and 300, respectively.

For PPS, we consider two scenarios. Under the first scenario, K clusters are chosen from N clusters based on the PPS sampling scheme but we use equal weights, $z_i = 1/N$ in (4.2) in estimating β, and in (4.10) in estimating σ^2. Under the second scenario, we use weights, $w_i = 1/z_i$, in the estimating equations proportional to the size of the cluster, where $z_i = m_i/m_o$ and $m_o = \sum_{i=1}^{N} m_i$ as described in section 4.1.1 for the estimation of β and section 4.1.2 for the estimation of σ^2. The results for the GQL estimation for β and the MM estimation of σ^2 under equally weighted PPS scheme (first scenario) are shown in Tables 4.3 and 4.4 for $N = 200$ and 300, respectively. Similarly, for the design-based weighted PPS (WPPS) case (second scenario), the results are shown in Tables 4.5 and 4.6, respectively. These results are given for sample clusters $K = 40, 60, 80, 100$ under population $N = 200$ and $K = 60, 80, 100, 140$ for $N = 300$.

Note that for some selected values of the parameters, the iterative technique did not converge. Hence, no estimate was obtained. These are denoted by "–" in Tables 4.1 through 4.6.

Table 4.1: GQL Estimates and associated standard errors for $\beta = (\beta_1, \beta_2)^T$ and the moment estimate of σ^2 and its standard error, using a **Simple Random Sample** of size K clusters from a finite population containing $N = 200$ clusters of unequal sizes based on 1000 simulations.

		Estimate		
K	True σ^2	$\hat{\beta}_1(SSE)$	$\hat{\beta}_2(SSE)$	$\hat{\sigma^2}(SSE)$
40	0.01	0.9919(0.1089)	0.4924(0.1393)	0.0543(0.0870)
	0.09	0.9889(0.1198)	0.4859(0.1572)	0.1179(0.1411)
	0.25	0.9930(0.1242)	0.4720(0.1788)	0.2500(0.2205)
	0.5625	0.9865(0.1369)	0.4530(0.2142)	0.5042(0.3801)
	0.81	—	—	—
60	0.01	0.9955(0.0962)	0.4883(0.1178)	0.0486(0.0784)
	0.09	0.9886(0.1058)	0.4895(0.1376)	0.1092(0.1246)
	0.25	0.9866(0.1105)	0.4764(0.1563)	0.2561(0.2109)
	0.5625	0.9891(0.1210)	0.4489(0.1837)	0.5027(0.3352)
	0.81	—	—	—
80	0.01	0.9910(0.0892)	0.4970(0.1122)	0.0435(0.0675)
	0.09	0.9940(0.0910)	0.4855(0.1200)	0.1043(0.1143)
	0.25	0.9951(0.0983)	0.4768(0.1379)	0.2502(0.1854)
	0.5625	0.9845(0.1114)	0.4778(0.1653)	0.5067(0.3241)
	0.81	—	—	—
100	0.01	0.9933(0.0827)	0.4963(0.1079)	0.0408(0.0598)
	0.09	0.9974(0.0901)	0.4880(0.1211)	0.1093(0.1069)
	0.25	0.9939(0.0897)	0.4750(0.1313)	0.2419(0.1668)
	0.5625	0.9893(0.1065)	0.4732(0.1471)	0.5187(0.3238)
	0.81	0.9860(0.1028)	0.4698(0.1585)	0.7473(0.4640)

Table 4.2: GQL Estimates and associated standard errors for $\beta = (\beta_1, \beta_2)^T$ and the moment estimate of σ^2 and its standard error, using a **Simple Random Sample** of size K clusters from a finite population containing $N = 300$ clusters of unequal sizes based on 1000 simulations.

		Estimate		
K	True σ^2	$\hat{\beta}_1(SSE)$	$\hat{\beta}_2(SSE)$	$\hat{\sigma^2}(SSE)$
60	0.01	0.9900(0.1084)	0.4947(0.1338)	0.0442(0.0703)
	0.09	0.9900(0.1127)	0.4912(0.1475)	0.1079(0.1261)
	0.25	0.9842(0.1283)	0.4800(0.1748)	0.2529(0.2079)
	0.5625	0.9887(0.1360)	0.4521(0.1893)	0.5007(0.3258)
	0.81	—	—	—
80	0.01	0.9978(0.1027)	0.4944(0.1203)	0.0424(0.0661)
	0.09	0.9940(0.1028)	0.4832(0.1385)	0.1097(0.1254)
	0.25	0.9875(0.1165)	0.4831(0.1514)	0.2510(0.1896)
	0.5625	0.9845(0.1228)	0.4689(0.1843)	0.5084(0.3397)
	0.81	—	—	—
100	0.01	0.9913(0.0927)	0.5002(0.1164)	0.0396(0.0580)
	0.09	0.9960(0.0995)	0.4912(0.1289)	0.1038(0.1071)
	0.25	0.9905(0.1091)	0.4818(0.1500)	0.2497(0.1797)
	0.5625	0.9894(0.1088)	0.4721(0.1665)	0.5249(0.3289)
	0.81	0.9819(0.1318)	0.4702(0.1852)	0.7176(0.4504)
140	0.01	0.9965(0.0896)	0.4973(0.1037)	0.0410(0.0609)
	0.09	0.9938(0.0918)	0.4912(0.1209)	0.1018(0.0968)
	0.25	0.9922(0.0989)	0.4875(0.1347)	0.2486(0.1617)
	0.5625	0.9958(0.1080)	0.4711(0.1514)	0.5347(0.2950)
	0.81	0.9868(0.1111)	0.4614(0.1742)	0.7394(0.4192)

Table 4.3: Equal Weights based GQL estimates and associated standard errors for $\beta = (\beta_1, \beta_2)^T$ and the moment estimate of σ^2 and its standard error, using a sample of size K clusters chosen based on the **Probability Proportional to Size** sampling scheme from a finite population containing $N = 200$ clusters of unequal sizes based on 1000 simulations.

		Estimate		
K	True σ^2	$\hat{\beta}_1(SSE)$	$\hat{\beta}_2(SSE)$	$\hat{\sigma^2}(SSE)$
40	0.01	0.9872(0.1039)	0.4939(0.1211)	0.0550(0.0917)
	0.09	0.9908(0.1062)	0.4845(0.1374)	0.1141(0.1264)
	0.25	0.9900(0.1090)	0.4779(0.1511)	0.2463(0.1971)
	0.5625	0.9847(0.1221)	0.4639(0.2133)	0.5075(0.3640)
	0.81	—	—	—
60	0.01	1.0002(0.0874)	0.4891(0.1035)	0.0436(0.0631)
	0.09	0.9954(0.0822)	0.4912(0.1118)	0.1022(0.1076)
	0.25	0.9915(0.0958)	0.4832(0.1286)	0.2533(0.1854)
	0.5625	0.9889(0.0974)	0.4655(0.1439)	0.5182(0.3266)
	0.81	1.0083(0.5774)	0.4598(0.1559)	0.7106(0.4574)
80	0.01	0.9946(0.0794)	0.4959(0.0993)	0.0392(0.0565)
	0.09	0.9978(0.0796)	0.4873(0.1064)	0.1084(0.1017)
	0.25	0.9932(0.0854)	0.4876(0.1180)	0.2526(0.1703)
	0.5625	0.9905(0.0909)	0.4747(0.1310)	0.5128(0.2994)
	0.81	0.9925(0.0941)	0.4626(0.1563)	0.7438(0.4480)
100	0.01	0.9987(0.0711)	0.4945(0.0847)	0.0357(0.0485)
	0.09	0.9995(0.0780)	0.4875(0.0994)	0.1013(0.0870)
	0.25	0.9965(0.0803)	0.4855(0.1108)	0.2505(0.1583)
	0.5625	0.9990(0.0834)	0.4780(0.1177)	0.5177(0.2927)
	0.81	0.9957(0.0830)	0.4777(0.1264)	0.7389(0.3942)

Table 4.4: Equal Weights based GQL estimates and associated standard errors for $\beta = (\beta_1, \beta_2)^T$ and the moment estimate of σ^2 and its standard error, using a sample of size K clusters chosen based on the **Probability Proportional to Size** sampling scheme from a finite population containing $N = 300$ clusters of unequal sizes based on 1000 simulations.

		Estimate		
K	True σ^2	$\hat{\beta}_1(SSE)$	$\hat{\beta}_2(SSE)$	$\hat{\sigma^2}(SSE)$
60	0.01	0.9959(0.0853)	0.4977(0.0978)	0.0371(0.0535)
	0.09	0.9955(0.0857)	0.4891(0.1083)	0.1026(0.1020)
	0.25	0.9946(0.0935)	0.4811(0.1272)	0.2556(0.1701)
	0.5625	0.9923(0.0964)	0.4759(0.1433)	0.5419(0.3381)
	0.81	0.9948(0.0925)	0.4708(0.1529)	0.7225(0.4386)
80	0.01	1.0016(0.0770)	0.4892(0.0876)	0.0342(0.0488)
	0.09	0.9933(0.0762)	0.4922(0.0951)	0.0972(0.0865)
	0.25	0.9980(0.0821)	0.4786(0.1109)	0.2561(0.1564)
	0.5625	0.9909(0.0830)	0.4780(0.1216)	0.5434(0.3009)
	0.81	0.9898(0.0823)	0.4714(0.1321)	0.7316(0.4032)
100	0.01	0.9986(0.0664)	0.4928(0.0761)	0.0318(0.0414)
	0.09	0.9990(0.0696)	0.4892(0.0832)	0.0994(0.0824)
	0.25	0.9931(0.0762)	0.4889(0.0971)	0.2501(0.1452)
	0.5625	0.9936(0.0749)	0.4815(0.1085)	0.5351(0.2630)
	0.81	0.9979(0.0727)	0.4777(0.1168)	0.7353(0.3668)
140	0.01	1.0002(0.0600)	0.4953(0.0709)	0.0331(0.0425)
	0.09	1.0003(0.0619)	0.4897(0.0805)	0.0959(0.0725)
	0.25	0.9973(0.0662)	0.4889(0.0832)	0.2560(0.1326)
	0.5625	0.9978(0.0702)	0.4786(0.1049)	0.5385(0.2483)
	0.81	0.9937(0.0701)	0.4845(0.1133)	0.7527(0.3643)

Table 4.5: Design Weights based GQL estimates and associated standard errors for $\beta = (\beta_1, \beta_2)^T$ and the moment estimate of σ^2 and its standard error, using a sample of size K clusters chosen based on the **Probability Proportional to Size** sampling scheme from a finite population containing $N = 200$ proportionally weighted clusters of unequal sizes based on 1000 simulations.

		Estimate		
K	True σ^2	$\hat{\beta}_1(SSE)$	$\hat{\beta}_2(SSE)$	$\hat{\sigma^2}(SSE)$
40	0.01	0.9862(0.1073)	0.4958(0.1232)	0.0530(0.0881)
	0.09	0.9910(0.1095)	0.4850(0.1401)	0.1103(0.1184)
	0.25	0.9902(0.1122)	0.4800(0.1554)	0.2453(0.1897)
	0.5625	0.9872(0.1227)	0.4592(0.1889)	0.5084(0.3493)
	0.81	—	—	
60	0.01	0.9998(0.0880)	0.4907(0.1037)	0.0427(0.0604)
	0.09	0.9953(0.0902)	0.4929(0.1130)	0.0998(0.1016)
	0.25	0.9927(0.0994)	0.4833(0.1325)	0.2528(0.1793)
	0.5625	0.9889(0.0986)	0.4666(0.1444)	0.5202(0.3172)
	0.81	—	—	—
80	0.01	0.9941(0.0799)	0.4975(0.0998)	0.0374(0.0525)
	0.09	0.9973(0.0805)	0.4882(0.1017)	0.1065(0.0975)
	0.25	0.9946(0.0868)	0.4874(0.1203)	0.2527(0.1625)
	0.5625	0.9904(0.0936)	0.4755(0.1327)	0.5125(0.2910)
	0.81	0.9902(0.0982)	0.4667(0.1516)	0.7471(0.4337)
100	0.01	0.9998(0.0719)	0.4937(0.0863)	0.0348(0.0461)
	0.09	0.9993(0.0786)	0.4878(0.0999)	0.0994(0.0826)
	0.25	0.9975(0.0818)	0.4846(0.1118)	0.2498(0.1520)
	0.5625	0.9898(0.0850)	0.4811(0.1183)	0.5202(0.2813)
	0.81	0.9956(0.0848)	0.4786(0.1265)	0.7420(0.3855)

Table 4.6: Design Weights based GQL estimates and associated standard errors for $\beta = (\beta_1, \beta_2)^T$ and the moment estimate of σ^2 and its standard error, using a sample of size K clusters chosen based on the **Probability Proportional to Size** sampling scheme from a finite population containing $N = 300$ proportionally weighted clusters of unequal sizes based on 1000 simulations.

		Estimate		
K	True σ^2	$\hat{\beta}_1(SSE)$	$\hat{\beta}_2(SSE)$	$\hat{\sigma^2}(SSE)$
60	0.01	0.9955(0.0872)	0.4991(0.0990)	0.0363(0.0516)
	0.09	0.9962(0.0878)	0.4894(0.1112)	0.1016(0.0980)
	0.25	0.9955(0.0947)	0.4822(0.1291)	0.2533(0.1630)
	0.5625	0.9930(0.0981)	0.4766(0.1457)	0.5416(0.3251)
	0.81	0.9952(0.0947)	0.4733(0.1975)	0.7210(0.4079)
80	0.01	1.0011(0.0769)	0.4896(0.0872)	0.0340(0.0472)
	0.09	0.9938(0.0785)	0.4922(0.0976)	0.0963(0.0836)
	0.25	0.9978(0.0831)	0.4811(0.1120)	0.2547(0.1519)
	0.5625	0.9908(0.0840)	0.4782(0.1237)	0.5442(0.2936)
	0.81	0.9898(0.0855)	0.4712(0.1344)	0.7328(0.3888)
100	0.01	0.9972(0.0674)	0.4955(0.0780)	0.0311(0.0391)
	0.09	0.9985(0.0707)	0.4906(0.0856)	0.0984(0.0797)
	0.25	0.9930(0.0774)	0.4906(0.0994)	0.2500(0.1391)
	0.5625	0.9940(0.0769)	0.4829(0.1100)	0.5383(0.2542)
	0.81	0.9976(0.0733)	0.4794(0.1179)	0.7372(0.3582)
140	0.01	1.0003(0.0608)	0.4958(0.0718)	0.0325(0.0406)
	0.09	0.9999(0.0630)	0.4910(0.0815)	0.0942(0.0688)
	0.25	0.9979(0.0681)	0.4886(0.0854)	0.2556(0.1271)
	0.5625	0.9983(0.0715)	0.4793(0.1054)	0.5419(0.2418)
	0.81	0.9935(0.0757)	0.4857(0.1102)	0.7604(0.3569)

4.3.1 Comparison of SRS and Equal Weights based PPS Sampling

The results given in Tables 4.1 and 4.2 for the SRS case and in Tables 4.3 and 4.4 for the equally weighted PPS case show that the later scheme produces much better estimates for the β and σ^2 parameters. For example, when a sample of size $K = 60$ or $K = 100$ is chosen from the finite population with size $N = 200$, the estimates for $\beta_1 = 1.0$ and $\beta_2 = 0.5$ for, say, $\sigma^2 = 0.25$, and the associated standard errors in parentheses, were found as follows. Under SRS when $K = 60$, we have $\hat{\beta}_1 = 0.9866(0.1105)$, $\hat{\beta}_2 = 0.4764(0.1563)$ and $\hat{\sigma}^2 = 0.2561(0.2109)$. Similarly, for $K = 100$, we have $\hat{\beta}_1 = 0.9939(0.0897)$, $\hat{\beta}_2 = 0.4750(0.1313)$ and $\hat{\sigma}^2 = 0.2419(0.1668)$. For the analogous equally weighted PPS cases we have estimates using $K = 60$ as follows: $\hat{\beta}_1 = 0.9915(0.0958)$, $\hat{\beta}_2 = 0.4832(0.1286)$ and $\hat{\sigma}^2 = 0.2533(0.1854)$. Similarly, for $K = 100$, we have $\hat{\beta}_1 = 0.9965(0.0803)$, $\hat{\beta}_2 = 0.4855(0.1108)$ and $\hat{\sigma}^2 = 0.2505(0.1583)$. We can see that within either of the techniques, the estimates as well as their standard errors improve (estimates approach true values and standard errors decrease) as sample size increases from 60 to 100. We also notice that those estimates under equally weighted PPS are better than those under the SRS scheme. As an illustration, for $K = 60$, the $\hat{\beta}_1$ value under SRS is 0.9866(0.1105), while it is 0.9915(0.0958) under equally weighted PPS, a much closer estimate with a smaller standard error. This comparative performace is not surprising. This is because the PPS scheme allows larger sized clusters to be included in the sample with more probability, and hence more information was included in the sample. These estimates also improve with increased population size.

For population $N = 300$, we see similar results as with the $N = 200$ case.

Suppose we compare the results for $K = 80$ and $K = 140$ for $\sigma^2 = 0.09$ from Tables 4.2 and 4.4. For SRS, we report the estimates and their standard errors as follows. For $K = 80$, $\hat{\beta}_1 = 0.9940(0.1028)$, $\hat{\beta}_2 = 0.4832(0.1385)$ and $\hat{\sigma}^2 = 0.1097(0.1254)$. For $K = 140$, $\hat{\beta}_1 = 0.9938(0.0918)$, $\hat{\beta}_2 = 0.4912(0.1209)$ and $\hat{\sigma}^2 = 0.1018(0.0968)$. We again see that with increased sample size, we get improved estimates with smaller errors. For the equally weighted PPS case, for $K = 80$ we have $\hat{\beta}_1 = 0.9933(0.0762)$, $\hat{\beta}_2 = 0.4922(0.0951)$ and $\hat{\sigma}^2 = 0.0972(0.0865)$, and for $K = 140$, $\hat{\beta}_1 = 1.0003(0.0619)$, $\hat{\beta}_2 = 0.4897(0.0805)$ and $\hat{\sigma}^2 = 0.0959(0.0725)$. It is evident that equally weighted PPS outperforms SRS in all cases.

4.3.2 Comparison of SRS and WPPS Sampling

In a similar comparison as given in the previous subsection, we can see that Design-Based Weighted PPS (WPPS) also performs better than SRS. We compare the results for SRS given in the previous subsection to those of WPPS. Specifically, we compare the results for $K = 60$ and $K = 100$ when $N = 200$ from Tables 4.1 and 4.5. For SRS with $K = 60$, we have $\hat{\beta}_1 = 0.9866(0.1105)$, $\hat{\beta}_2 = 0.4764(0.1563)$ and $\hat{\sigma}^2 = 0.2561(0.2109)$. Similarly, for $K = 100$, we have $\hat{\beta}_1 = 0.9939(0.0897)$, $\hat{\beta}_2 = 0.4750(0.1313)$ and $\hat{\sigma}^2 = 0.2419(0.1668)$, as recorded in the previous subsection. For WPPS $K = 60$, we have $\hat{\beta}_1 = 0.9927(0.0994)$, $\hat{\beta}_2 = 0.4833(0.1325)$ using the WGQL estimating technique, and $\hat{\sigma}^2 = 0.2528(0.1793)$ using WMM. For $K = 100$, we have $\hat{\beta}_1 = 0.9975(0.0818)$, $\hat{\beta}_2 = 0.4846(0.1118)$ and $\hat{\sigma}^2 = 0.2498(0.1520)$. In both cases, WPPS gives estimates with smaller biases and standard errors.

Using the same SRS cases where $N = 300$, $K = 80$ and 140, and $\sigma^2 = 0.09$, we can compare with WPPS for the same cases from Table 4.6.

These results for SRS were given in the previous section, and are not repeated here. Under WPPS for $K = 80$, we found $\hat{\beta}_1 = 0.9938(0.0785)$, $\hat{\beta}_2 = 0.4922(0.0976)$ and $\hat{\sigma}^2 = 0.0963(0.0836)$. Likewise, for $K = 140$, we have $\hat{\beta}_1 = 0.9999(0.0630)$, $\hat{\beta}_2 = 0.4910(0.0815)$ and $\hat{\sigma}^2 = 0.0942(0.0688)$. Though the β estimates do not change much, their standard errors are smaller and the estimate of σ^2 improves, giving more accurate results.

4.3.3 Comparison of Equally Weighted PPS and WPPS

Using the same examples as given in the previous two subsections from Tables 4.3 and 4.4, for $N = 200$ and $K = 60$ under equally weighted PPS we found $\hat{\beta}_1 = 0.9915(0.0958)$, $\hat{\beta}_2 = 0.4832(0.1286)$ and $\hat{\sigma}^2 = 0.2533(0.1854)$. Similarly, for $K = 100$, we have $\hat{\beta}_1 = 0.9965(0.0803)$, $\hat{\beta}_2 = 0.4855(0.1108)$ and $\hat{\sigma}^2 = 0.2505(0.1583)$. For the design based weighted case, WPPS, Table 4.5 reports for $K = 60$, $\hat{\beta}_1 = 0.9927(0.0994)$, $\hat{\beta}_2 = 0.4833(0.1325)$ and $\hat{\sigma}^2 = 0.2528(0.1793)$. For $K = 100$, we have $\hat{\beta}_1 = 0.9975(0.0818)$, $\hat{\beta}_2 = 0.4846(0.1118)$ and $\hat{\sigma}^2 = 0.2498(0.1520)$. In each case, WPPS produces slightly better estimates than the equally weighted PPS scheme.

Under the equally weighted PPS scheme for $N = 300$ for $K = 80$, we get $\hat{\beta}_1 = 0.9933(0.0762)$, $\hat{\beta}_2 = 0.4922(0.0951)$ and $\hat{\sigma}^2 = 0.0972(0.0865)$, and for $K = 140$, $\hat{\beta}_1 = 1.0003(0.0619)$, $\hat{\beta}_2 = 0.4897(0.0805)$ and $\hat{\sigma}^2 = 0.0959(0.0725)$ from Table 4.4. Under WPPS, we acquire the estimates for $K = 80$ as $\hat{\beta}_1 = 0.9938(0.0785)$, $\hat{\beta}_2 = 0.4922(0.0976)$ and $\hat{\sigma}^2 = 0.0963(0.0836)$. For $K = 140$, we have $\hat{\beta}_1 = 0.9999(0.0630)$, $\hat{\beta}_2 = 0.4910(0.0815)$ and $\hat{\sigma}^2 = 0.0942(0.0688)$ from Table 4.6.

We find that under both scenarios, PPS performs much better than that of SRS. WPPS seems to give very similar results to that of equally weighted

PPS. Most notably, we can see that the estimate of σ^2 improves slightly with the addition of weights to the estimating equations.

Overall, we can see that the estimates tend to be more precise with a larger sample relative to the population size, and they exhibit smaller values of the variance parameter. The estimates become even more precise using a sampling scheme that takes the sizes of the clusters into account, such as WPPS.

Chapter 5

Concluding Remarks

There does not appear to be adequate discussion on the estimation of the parameters of the clustered regression models for the count data under the finite population setup. For a discussion on the estimation of regression parameters in such a finite population setup, one may refer to Thompson (1997, Chapter 6).

The main contribution of the practicum is to develop an estimation methodology for consistent estimation of both the regression effects and the variance paramenter of the random effects for a clustered regression model under the finite population setup. It was found that when cluster/family sizes were different, the use of the PPS sampling scheme provided better estimates of the parameters, as opposed to the use of the SRS scheme. These results are, in general, in agreement with those discussed by Binder (1983) and Pfeffermann (1993). Note, however, that Binder (1983) discussed longitudinal cluster data analysis under the finite population, whereas we have discussed a Poisson mixed model, that is, a familial clustered regression model for count data, under the finite population. The non-regression cases were also studied

in the beginning of the practicum. This study showed the usefulness of the application of the PPS scheme for the purpose of estimation of the finite population total.

We remark that in the practicum, we have used the GQL approach for regression estimation and the moment approach for the variance parameter estimation under the finite population setup. One may also attempt to use the GQL approach, instead of the method of moments approach to increase the efficiency of the variance estimate, which can be a future project. We also remark that the estimating equations based results of this practicum should be highly useful to statistical agencies such as Statistics Canada for the inferences on familial count data.

Appendix

Fortran Program to Generate a Simple Random Sample

```
integer i,j,k,N,iseed,n1,n2,n3,n
integer m(N),y1(N/4,n1),y2(N/2,n2)
integer y3(N/4,n3)
integer y(N),ir1(n1),ir2(n2),ir3(n3)

parameter(N=100,n1=4,n2=6,n3=5,n=20)

real mu1,mu2,mu3,y(n)
real r(n),z(n)

[comment:  N is the population cluster number and n is the sample clust

external rnpoi,rnset,rnun

open(10,file='srs.out',status='new')

iseed=2345678
call rnset(iseed)

mu1=2
mu2=4
mu3=3

do i=1,N/4
call rnpoi(n1,mu1,ir1)
do j=1,n1
y1(i,j)=ir1(j)
end do
end do
```

```
do i=1,N/2
call rnpoi(n2,mu2,ir2)
do j=1,n2
y2(i,j)=ir2(j)
end do
end do

do i=1,N/4
call rnpoi(n3,mu3,ir3)
do j=1,n3
y3(i,j)=ir3(j)
end do
end do

do i=1,N/4
y(i)=0
do j=1,n1
y(i)=y(i)+y1(i,j)
end do
end do

do i=N/4+1,N/4*3
y(i)=0
do j=1,n2
y(i)=y(i)+y2(i-N/4,j)
end do
end do
do i=N/4*3+1,N
y(i)=0
do j=1,n3
y(i)=y(i)+y3(i-N/4*3,j)
end do
end do

do i=1,N/4
m(i)=n1
end do
do i=N/4+1,N/4*3
m(i)=n2
end do
do i=N/4*3+1,N
m(i)=n3
end do
```

```
    call rnun(n,r)
    do j=1,n
    z(j)=y(int(r(j)*100))
    end do

    close(10)
end
```

Fortran Program to Generate a Sample using Probability Proportional to Size

```
integer i,j,k,N,iseed,n1,n2,n3,n

parameter(N=100,n1=4,n2=6,n3=5,n=20)

integer m(N),y1(N/4,n1),y2(N/2,n2)
integer y3(N/4,n3)
integer y(N),ir1(n1),ir2(n2),ir3(n3)
integer T,c(N),nr(n)
integer yf(n),mf(n)

real mu1,mu2,mu3
real d(N),r1(N),dd(N),rT,r(n)
real yn(n),absdf(N),z(n),rn
real q(n),df(N),sabsdf(N)

external rnpoi,rnset,rnun

open(10,file='pps.out',status='new')

iseed=2345678
call rnset(iseed)

mu1=2
mu2=4
mu3=3

rn=n

do i=1,N/4
call rnpoi(n1,mu1,ir1)
do j=1,n1
y1(i,j)=ir1(j)
end do
end do

do i=1,N/2
call rnpoi(n2,mu2,ir2)
do j=1,n2
y2(i,j)=ir2(j)
```

```
end do
end do

do i=1,N/4
call rnpoi(n3,mu3,ir3)
do j=1,n3
y3(i,j)=ir3(j)
end do
end do

do i=1,N/4
y(i)=0
do j=1,n1
y(i)=y(i)+y1(i,j)
end do
end do

do i=N/4+1,N/4*3
y(i)=0
do j=1,n2
y(i)=y(i)+y2(i-N/4,j)
end do
end do
do i=N/4*3+1,N
y(i)=0
do j=1,n3
y(i)=y(i)+y3(i-N/4*3,j)
end do
end do

do i=1,N/4
m(i)=n1
end do
do i=N/4+1,N/4*3
m(i)=n2
end do
do i=N/4*3+1,N
m(i)=n3
end do

T=N/4*n1+N/2*n2+N/4*n3
rT=T
```

```
c(1)=m(1)
d(1)=c(1)/rT
do i=2,N
c(i)=c(i-1)+m(i)
end do

do i=2,N
d(i)=c(i)/rT
end do

call rnun(n,q)

do j=1,n
do i=1,N
df(i)=q(j)-d(i)
absdf(i)=abs(df(i))
end do
call svrgn(N,absdf,sabsdf)
do i=1,N
if(sabsdf(1).eq.(1*df(i)).or.sabsdf(1).eq.(-1*df(i)))then
nr(j)=i
end if
end do
end do

do j=1,n
yf(j)=y(nr(j))
mf(j)=m(nr(j))
end do

close(10)

end
```

Bibliography

[1] Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*, **51**, 279-292.

[2] Breslow, N.E. and Clayton, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, **88**, 9-25.

[3] Cochran, W.G. (1977). *Sampling Techniques,* Third Edition. Wiley, New York.

[4] Jiang, J. (1998). Consistent Estimators in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, **93**, 720-729.

[5] Jiang, J. and Zhang, W. (2001). Robust Estimation in Generalized Linear Mixed Models. *Biometrika*, **88**, 753-765.

[6] Pfeffermann, D. (1993). The Role of Sampling Weights when Modeling Survey Data. *International Statistical Review*, **61**, 317-337.

[7] Sutradhar, B.C. (2004). On Exact Quazilikelihood Inference in Generalized Linear Mixed Models. *Sankhya: The Indian Journal of Statistics*, **66**, 261-289.

[8] Sutradhar, B.C. and Das, K. (2000). On the Accuracy of Efficiency of Estimating Equation Approach. *Biometrics*, **56**, 622-625.

[9] Sutradhar, B.C. and Qu, Z. (1998). On Approximate Likelihood Inference in Poisson Mixed Model. *Canadian Journal of Statistics*, **26**, 169-186.

[10] Sutradhar, B.C. and Rao, R.P. (2001). On Marginal Quasi-Likelihood Inference in Generalized Linear Mixed Models. *Journal of Multivariate Analysis*, **76**, 1-34.

[11] Thompson, M.E. (1997). *Theory of Sample Surveys.* Chapman and Hall, New York.