Mimeo Series
No. 2161T May 96
Statistical Estimation Under
Length Biased Distributions
by
Stoffel R. Moeng
Dept. of Biostatistics
Univ. of North Carolina

Name Date

STATISTICAL ESTIMATION UNDER LENGTH BIASED DISTRIBUTIONS

by

Stoffel R. Moeng
Department of Biostatistics
University of North Carolina

Institute of Statistics
Mimeo Series No. 2161T

May 1996

STATISTICAL ESTIMATION UNDER LENGTH BIASED DISTRIBUTIONS

by

Stoffel R. Moeng

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partuak fulfillment of the requirements for the degree of Doctor of Philiphy in the Department of Biostatistics

Chapel Hill

1996

STOFFEL R. MOENG. Statistical Estimation Under Length-Biased Distributions. (Under the joint direction of Drs. P.K. Sen and C.M. Suchindran.)

ABSTRACT

In simple random sampling, there is an equal chance of selection for each unit in the population. However, there are some practical situations where this might not be ideal. We consider the case of length-biased sampling, whereby the chance of selection of a unit is proportional to the length of the unit, by which the length-biased density is defined. Specifically we consider the case of a continuous random variable Y with pdf f(y;$\boldsymbol{\theta}$), where, f(y;$\boldsymbol{\theta}$) is of an exponential family of distributions and $\boldsymbol{\theta}$ is a k-parameter vector to be estimated. We show the bias of the MLE of $\boldsymbol{\theta}$ to be of order n^{-1} and use the jackknife estimation technique to eliminate the leading term in the bias of the MLE of $\boldsymbol{\theta}$. It is shown that the Jackknife estimator has the same normal distribution as the MLE in large samples.

The estimation study is extended to the regression problem where the mean of the sufficient statistics of the exponential family of distributions depends on a set of covariates. Both the bias and asymptotic distribution of the MLE and the Jackknife are provided. The Fisher information from the length-biased distribution and the original distribution are compared through the A-optimalily and D-optimality which are functions of the eigenvalues of the Fisher information matrices.

Data from the Demographic and Health Surveys is analyzed by maximum likelihood methods and Jackknife methods. The response variables is the first order birth interval and the potential covariates are the age at marriage and duration of marriage for the mothers. Simulation studies from the lognormal and gamma distributions indicate a reduction in the bias of the MLE's for the scale parameter of the lognormal distribution. Both the MLE and the Jackknife estimates are approximately normal. The RMSE for the MLE and the Jackknife estimates are given.

ACKNOWLEDGEMENTS

First of all, I would like to thank professor P.K. Sen for his patient guidance and all his advice in the preparation of this dissertation. I would also like to thank Professor C.M. Suchindran for his joint direction and his many helpful encouragement, comments and suggestions on how best to shape this material. I am grateful to my other dissertation committee members, professor M. Symons, Professor G. Koch and Professor R. Bilsborrow for their participation in my research adventure. I owe a particularly deep debt of gratitude to Professor Bilsborrow for his efforts "above and beyond the call of duty" during my studies in helping me realize this goal.

I am grateful for the support of my friends, the Hogan family who shared with me the love of their home and with whom I have shared Thanksgivings; my fellow graduate students for their friendship and encouragement during my education. of course, none of this effort would have been possible without the love and support of my family; my parents, aunts, uncles, and sisters. My special thanks to my aunt Dorah who died just after I had started my studies at UNC and well after she had left her marks upon my life.

I am deeply thankful to the entire faculty of the Department of Biostatistics at UNC for their constant support and motivation during my studies. Finally I would like to acknowledge the university of North Carolina for the opportunity to attend and learn a valuable education.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I.
Introduction and Literature Review.

1.1. Introduction.

Statistical inference methods for analysing data are aimed at estimating the parameters of a postulated underlying probability model generating the data. It is usually assumed that the investigator has access to the data following the distribution from which inference is to be drawn. This assumption is not always justified. In statistical analysis, one often has observations from a weighted distribution rather than from the original (unweighted) distribution, and from these observations one needs to make statistical inferences about the parameters of the original distribution.

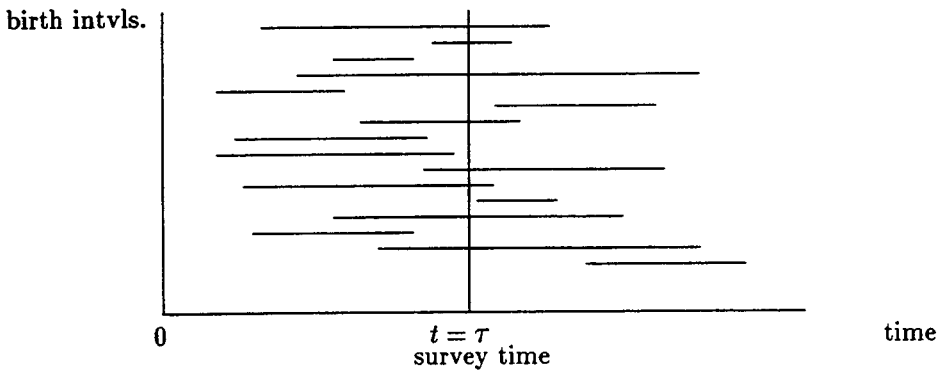
This research is motivated by the problem of statistical estimation of the population parameters from the data generated by a sampling design which gives rise to data from a length-biased distribution, which is a form of weighted distribution. Weighted distributions arise when realizations y_i of a random variable Y having a probability density function f(y) are observed and recorded according to a probability density g(y) which is a weighted version of f(y). Rao (1965) identified various sampling schemes that could be modeled by weighted distributions.

Length-biased sampling has implications for survey data collected at an

arbitrary point in time. Suppose that a component operating in a system is replaced upon failure by another component having the same life distribution, so that the sequence of component life lengths $\{Y_i\}$ form a renewal process. In order to determine the mean life of the component, components in operation in a particular system are sampled at a fixed time on a particular day and the total lives of these components are measured. Because the components selected are in operation on that particular date, they constitute a nonrandom sample of all components, having longer mean lifetimes than the class of components from which they were drawn. The probability that the sampled component has life length y is a length-biased version of the probability distribution of the lengths of all the population of life lengths of the components.

The length-biased design is illustrated in Figure 1 for a hypothetical survey data on birth intervals of a given order k, for a cohort of women married at the same time. Each line represents the duration of time between any two live births for each woman, such that the left endpoint is the time the woman experienced the event of birth, and the next birth occurred at the right endpoint of the line. Suppose at an arbitrary point in time (say $t = \tau$), a survey is undertaken to estimate the mean birth interval for all birth intervals of the given order in the population from which these birth intervals are drawn. The investigator ascertains and records the time of the last birth before the survey and the time the next birth occurs for the births covering the k^{th} birth interval. Since longer birth intervals are more likely than shorter birth intervals to overlap the survey time and hence be sampled, the sample of birth intervals thus obtained comprises of a length-biased sample of all k^{th} order of birth intervals.

Figure 1. Hypothetical data layout of the k^{th} order birth intervals.



Length-biased sampling is often encountered in event history data collection. In renewal event, such as open birth intervals in demographic studies, long intervals are usually over-represented, resulting in biased estimates of the mean birth interval of the population. Survey data taken at an arbitrary point in time to measure the time between the occurrence of an event and the next occurrence of the event tend to draw a length-biased data from the population of all such duration times. Examples of such survey questions include questions such as "when did you last see a physician?" and "when did you last give live birth?", and the duration of time since the last occurrence and the next occurrence is recorded form a length-biased data (Morrison, 1973).

The length-biased distribution is frequently appropriate for certain natural sampling plans in survival analysis, biometry and reliability (see Zelen, 1974; Gupta and Keating, 1986; Gupta and Kirmani, 1980). Section 1.2 reviews the relevant literature in the area of length-biased distributions.

Important factors in data analysis include the choice of an appropriate model for the data, which has implications for the estimation techniques, that is whether

parametric methods, for example, maximum likelihoods, the method of moments or other nonparametric methods, would be appropriate. Failure to incorporate the selection mechanism in analysing such data leads to biased results.

It is also important to consider the effects of possible covariates on the parameters to be estimated. Information on covariates affecting the response would improve the precision of parameter estimators and provide greater validity to statistical inference. Such analysis is usually carried out by regression analysis methods.

The problems of interest to us are: (*i*) To obtain unbiased and efficient estimators of the parameters of the original distribution, given that the data are from a length-biased distribution, (*ii*) To study the asymptotic distribution of such estimators and (*iii*) To examine the information about the the parameters lost or gained by not having access to the sample following the original distribution.

The objectives of the study may focus on recovering the characteristics of the original unweighted distribution based on a lenth-biased sample, and assessing the nature of distortion in determining these characteristics in case the inherent biased sampling is ignored in modelling the outcome variable. One may also be interested in the estimates of the length-biased distribution in itself.

1.2 Design Issues

The length biased or weighted sampling procedure occurs when a sample of units is drawn from a population such that the probability of a unit being selected in the sample is proportional to its length. To motivate the estimation problem, we state the problem as follows:

Let $\{Y_i.\ i = 1,\ldots,n\}$ be a sequence of independent random variables with

common pdf f(y;θ) for $\theta \in \Omega$. Instead of obtaining sample observations from the original density f, the realizations made on the random variable Y are obtained from the length-biased density

$$g(y;\theta) = \frac{w(y)f(y;\theta)}{\mu(\theta)}, \quad w(y) > 0. \tag{1.2.1}$$

where $\mu(\theta) = E[w(Y)] = \int w(y)\, f(y;\theta)\, dy$, for some nonnegative weight function w(y).

Among the weight functions commonly used are y, y^{γ} ($\gamma > 0$) and $1-(1-\beta)^{y}$ for $0 < \beta < 1$. The pdf g(y;θ) is sometimes referred to as the weighted distribution. As a special case, this research will restrict itself to length-biased distributions where the weight function w(y)=y. The random sample from length-biased distribution is called a length-biased sample (Cox, 1969 ; Patil and Rao, 1977).

The problem then is to estimate the parameter θ or functions thereof on the basis of length-biased sample from density g.

1.3. Literature Review and Related Research.

1.3.1. Origins and Applications.

The concept of weighted distributions was first formulated in a unified manner by Rao (1965) as an extension of the basic idea originally developed by Fisher (1934). Rao distinguised three major reasons, or situations, in which a sampling scheme would yield a sample that may be a biased representation of the population, and introduced the concept of weighted distributions as an adjustment

to the original unweighted distribution. Applications of the weighted distributions have since received attention in the literature. Briefly, the three situations identified by Rao are as follows.

i) Non observability / Visibility bias

This occurs frequently in ecological, aerial surveys when a group gets recorded only when at least one member in the group is sighted and each individual has an independent chance β of being sighted. If X is the number of individuals in a group with pdf f(x), the probability that an observed group has x individuals is a weighted distribution with weight $w(x) = 1-(1-\beta)^x$. In this case $E[w(X)]$ is the probability of observing a group. Rao shows that as $\beta \longrightarrow 0$, the corresponding distribution has a length biased pdf. Rao (1965) discussed the application in the study of the distribution of albino children (individuals with rare case). A convenient sampling method is to first discover an albino child and through it obtain the albino count X^w of the family to which it belongs. The pdf of X^w is a length-biased version of the original r.v. X, the number of albino children in families with proneness to produce such children, similarly, in some instances we may be able to observe patients who have a certain disease that is manifested only when some symptoms are visible and we are interested in the distribution of the entire population.

ii) Partial destruction of observations/ partial ascertainment.

This situation arises when a sample is obtained from a pdf f(x), but an observation $X = x$ is reduced to y, by a destruction process, with conditional pdf $\zeta(y/x)$. Suppose that an observation is recorded only when the original value is unchanged, and the probability that the observation $X = x$ is undamaged is $\zeta(x/x)$. Then the pdf of an obervation X^w is a weighted distribution with a

weight function $w(x) = \zeta(x/x)$. In observing human disease, the probability of detection can depend on the length of time the individual has had the disease. Zelen (1974) discussed how health-screening programs tend to sample a length biased distribution of all individuals with the disease, in which the weight is the length of time the subject is in the pre-clinical state and the disease is undetectable. Specifically Zelen argued that women with breast cancer detected in a cancer screening program are a length biased sample of all women with breast cancer. The weight is the length of time that cancer is in the detectable state. Since patients with a short pre-clinical stage may never be enrolled in the screening program, resulting in failure to report some cases, the sample tends to be a biased sample. Patients with slow-growing, less aggressive forms of the cancer will be over represented as their pre-clinical stage is longer.

iii) Sampling with unequal chances of selection.

This design is common in sampling surveys where the investigator intentionally samples units according to a probability proportional to the size (pps) of some measure. For example, let (X,Y) be a pair of nonnegative random variables with joint pdf f(x,y) and X be the main r.v. under study. An observation is made on X using the conditional distribution f(x/y) of X given Y. Patil and Rao (1974) show that the pdf of the resulting weighted version of the r.v. X, (say X^w), is given by $f^w(x) = \mathbb{E}[w(Y)/x]f(x)/\mathbb{E}[w(Y)]$. In this case the weight function $w(x,y) = w(y)$ is the regression of w(Y) on X. As is the case in sample surveys involving sampling with probability proportional to size, Y may define a design variable, and the sample is selected with probability proportional to the Y values. Thus, the original pdf is adjusted according to a given design of

selection of samples so as to improve the efficiency of estimators.

Cox (1969) discussed an idealized model in sampling textile fibers called length-biased or weighted sampling in which the choice of selection is proportional to the length of the fiber. When sampling a cross-section of a yarn, the fiber that crosses this cross-section has a length biased version of the distribution of the lengths of all the fiber in a yarn.

1.3.2. Relationship between Unweighted and Weighted Distributions

Cox (1969), Patil and Rao (1978), Mahfoud and Patil (1982) compared moments of unweighted distributions and weighted distributions. Gupta and Keating (1986); Gupta and Kirmani (1980) studied the relationship between unweighted and weighted distributions in the context of reliability and life testing, in particular, ralationships of survival functions, failure rates and residual life functions. The relationships of the measures depend on the nature of the weight functions. We review some of the relationships assuming Y is a length-biased random variable of the original unweighted random variable X. Thus

$$\mathbb{E}(Y) - \mathbb{E}(X) = V(X)/\mathbb{E}(X) \tag{1.3.1}$$

or

$$\mathbb{E}(Y) = \mathbb{E}(X)[\,1 + V(X)/\mathbb{E}^2(X)\,], \tag{1.3.2}$$

where $\mathbb{E}(X)$ denotes the expected value of X and V(X) denotes the variance of X. Hence the length-biased Y tends to record larger values of X more often.

Gupta (1981) discusses and relates the reliability measures of the original distribution to the reliability measures of the corresponding length-biased distribution. He shows that

$$Q(y) \geq S(y), \qquad \lambda_g(y) \leq \lambda_f(x) \text{ and } r_g(y) \geq r_f(x)$$

where Q(y) and S(y) are the survival functions of the length-biased distribution and unweighted distribution respectively, $\lambda_g(y)$ and $\lambda_f(y)$ are the hazard functions and $r_g(y)$ and $r_f(y)$ are the mean residual life functions of the length-biased distribution and its original distribution, respectively.

The reliability functions are defined in terms of the random variable X with expectation $\mathbb{E}X = \mu$.

The survival function is defined as:

$$S(t) = P\{X > t\} = 1 - F(t) \quad \text{and the derivative } S'(x) = -f(x)$$

and the hazard function given by

$$\lambda(t) = f(t)/S(t). \tag{1.3.3}$$

The mean residual life function is defined as:

$$r(t) = \mathbb{E}\{X - t \;/\; X > t\} = \int_t^\infty S(u)du/S(t). \tag{1.3.4}$$

The equivalence of S(t), $\lambda(t)$ and r(t) are can be established as follows:

$$S(t) = \exp\{-\int_0^t \lambda(u)du\}$$

$$= \frac{r(o)}{r(t)} \exp\{-\int_0^t \frac{du}{r(u)}\}, \tag{1.3.5}$$

showing that S(t) can be derived from $\lambda(t)$ or r(t).

The hazard function can be expressed in terms of r(t) as

$$\lambda(t) = \{1 + r'(t)\}/\, r(t), \tag{1.3.6}$$

and r can be obtained from S by the formula,

$$r(t) = \int_t^\infty S(u)du/S(t) = \{\mu - \int_0^t S(u)du\}/S(t). \tag{1.3.7}$$

Hence given any one of the three measures, the other two can be determined.

To aid in establishing the relation between the reliability measures of unweighted distribution and its length-biased form of distribution due to Gupta (1981), we use the notations of the reliability measures as defined above.

The relation between survival functions is given by observing that

$$Q'(y) = yS'(y)/\mu, \tag{1.3.8}$$

and integrating both sides to obtain

$$Q(y) = -\int_y^\infty \frac{uS'(u)}{\mu}\, du \quad = S(y)\{y + r_f(y)\}/\mu. \tag{1.3.9}$$

Using the definition of $\lambda(u)$, the relation between the hazard functions is established below as:

$$\lambda_g(y) = \frac{g(y)}{Q(y)} = y\lambda_f(y)/\{y + r_f(y)\}$$

$$= y\lambda_f(y)/\{y + \exp\{\Lambda_f(y)\}[\,\mu - \int_0^y \exp\{-\Lambda(t)\}dt]\}, \tag{1.3.10}$$

where $\Lambda(t) = \int_0^t \lambda(u)du$ is the integrated hazard function.

From algebraic manupulation and using the definition of $r(t)$, we obtain $\lambda_f(y)$ as a function of $\lambda_g(y)$ given by

$$\lambda_f(y) = \frac{f(y)}{S(y)} = \frac{g(y)/y}{\int_y^\infty \frac{1}{u}\, g(u)du} = \frac{\lambda_g(y)/y}{\int_y^\infty [\lambda_g(u)Q(u)\ \frac{1}{uQ(y)}]\ du}$$

$$= \frac{\lambda_g(y)/y}{\int_y^\infty [\lambda_g(u)\ \frac{1}{u} \exp\{-\int_y^u \lambda_g(s)ds\}]\ du}. \tag{1.3.11}$$

The relation between the mean residual functions is established below:

From the definition of r(t) and (1.3.9), the mean risidual life function, $r_g(y)$, of the length-biased distribution is expressed in terms of the mean residual life function, $r_f(y)$, of the unweighted distribution as;

$$r_g(y) = \int_y^\infty \frac{Q(u)}{Q(y)}\,du = \int_y^\infty \frac{S(u)\{u + r_f(u)\}}{S(y)\{y + r_f(y)\}}\,du$$

$$= \frac{r_f(y)}{y + r_f(y)} \int_y^\infty \frac{u + r_f(u)}{r_f(u)} \exp\{-\int_y^u \frac{ds}{r_f(s)}\}\,du \qquad (1.3.12)$$

and $r_f(y)$ is expressed in terms of $r_g(y)$ as:

$$y + r_f(y) = \frac{\exp[-\int_0^y \{du/r_g(u)\}]/r_g(y)}{\int_0^\infty \{1 + r_g'(t)\}/\{t(r(t))^2\}\exp[-\int_0^t \{du/r_g(u)\}]\,dt}. \qquad (1.3.13)$$

These results show that if the data are unknowingly obtained from a weighted version of the original distribution, the analyst will overestimate the residual life function.

1.3.3. Comparison of Unweighted and Weighted Experiments

The theory of comparison of statistical experiments was developed originally by Blackwell (1951), based on the concept of sufficiency. For two experiments $E_x = \{X, S_x ; f_\theta, \theta \in \Omega\}$ and $E_y = \{Y, S_y ; g_\theta, \theta \in \Omega\}$ where E_x denotes an experiment in which a random variable defined on some sample space S_x is to be observed, and the distribution f_θ of X depends on the unkown parameter $\theta \in \Omega$,

and likewise definition for E_y, Blackwell defines the experiment E_x to be sufficient for experiment E_y if there exists a stochastic transformation $Z(X)$ such that for each $\theta \in \Omega$, the random variable $Z(X)$ and Y have identical distributions. This means that if x is observed from performing experiment E_x , then x' obtained according to distribution $h(z/x)$ of $Z(X)$ will be as informative as x^* resulting from experiment E_y.

Lindley (1955) gives the comparison of two experiments, E_x and E_y with the same parameter $\theta \in \Omega$, based on the measure of information provided by the experiments. He defines experiment E_x to be more informative than experiment E_y if $I_x(\theta) \geq I_y(\theta)$, where $I_x(\theta)$ is the Fisher information matrix from experiment E_x . If the equality holds, the two experiments are said to be equally informative. Lindley shows that if E_x is sufficient for E_y then E_X is not less informative than E_y.

Bayarri and De Groot (1989) compared information about θ obtained from weighted distributions with that obtained from unweighted distribution, based on the concepts of sufficiency as developed by Blackwell. They defined experiment E_x to be sufficient for another experiment E_y, with the same parameter space, if $I_x(\theta) - I_Y(\theta)$ is nonnegative definite. Their results applied to the Binomial, Poisson and Negative binomial, and their respective versions of length-biased distributions show different effects for each family of distributions. For the binomial, E_x is sufficient for E_y; for the Poisson E_x and E_y are equivalent; and for the negative binomial E_y is sufficient for E_x.

Another form of weighted sample design, selection sample, arise when sample selection is restricted to a certain subset of the sample space. The weighted form of the original pdf f is expressed as

$$g(x/\boldsymbol{\theta}) = f(x/\boldsymbol{\theta})/s(\boldsymbol{\theta}), \tag{1.3.14}$$

where $s(\boldsymbol{\theta}) = P(X \in S/\boldsymbol{\theta})$. Clearly this is another form of (1.1.1) with $w(x) = 1$, if $X \in S$; $w(x) = 0$, if $X \notin S$.

Bayarri and De Groot (1986) compared the experiment in which a selection sample is obtained with that of the unweighted distribution by means of Fisher information. They considered the case where X has a normal distribution with unknown mean θ and known precision, and studied various selection models for the restricted sample. If Y is restricted to the set $Y \geq \tau$, it is shown that E_x is sufficient for E_y; if Y is restricted to the set $S=\{y: y \leq \tau_1 \text{ or } y \geq \tau_2\}$, where $\tau_1 \leq \tau_2$, then E_y is sufficient for E_x ; and for $\tau_1 \leq y \leq \tau_2$, they show E_x to be sufficient for E_y. Further it is shown by Bayarri and De Groot (1986) that for selection models from an exponential family with density

$$f(x/\theta) = a(x)c(y)\exp\{v(\theta)u(x)\} ,$$

the information matrix for the selection model is given by

$$\boldsymbol{I}_y(\theta) = \boldsymbol{I}_x(\theta) + \frac{d^2}{d\theta^2} \log s(\theta) ,$$

where $\boldsymbol{I}_x(\theta)$ denotes the information matrix from the unrestricted sample. Hence

$\boldsymbol{I}_y(\theta) \geq \boldsymbol{I}_x(\theta)$ iff log $s(\theta)$ is convex, and

$\boldsymbol{I}_y(\theta) \leq \boldsymbol{I}_x(\theta)$ iff log $s(\theta)$ is concave.

1.3.4. Estimation of parameters

Cox (1969) considered estimating the mean of the unweighted distribution f(y) based on the weighted pdf g(y), and also considered estimation of the unweighted distribution function F at a fixed $y > 0$. Denoting by E_g the mean with respect to the weighted distribution, the relations between the moments of g(y) and the moments of f(y) can be expressed as:

$$\mathbb{E}_g(Y^r) = \frac{\mu_{r+1}}{\mu} \tag{1.3.15}$$

where μ_r is the r^{th} moment of the r.v. Y with respect to pdf f.

From (1.3.15) we obtain

$$\mathbb{E}_g(Y) = \mu \left(1 + \frac{\sigma^2}{\mu^2} \right), \tag{1.3.16}$$

where μ and σ^2 are the mean and variance of the unweighted distributions.

Also from (1.3.15) we get

$$\mathbb{E}_g(Y^{-1}) = \frac{1}{\mu} \quad \text{and} \quad \mathrm{Var}(Y^{-1}) = \frac{1}{\mu^2} [\, \mu\mu_{-1} - 1],$$

which suggests using

$$\frac{1}{\tilde{\mu}} = \frac{1}{n} \{ \sum Y_i^{-1} \}, \tag{1.3.17}$$

as an unbiased estimate of $\frac{1}{\mu}$. Cox (1969) proposed $\tilde{\mu}$, the harmonic mean of the values of y in the biased sample, as the estimator of the mean of the original distribution.

As discussed by Sen (1987), $\tilde{\mu}$ is a biased though consistent and asymptotically normal estimator of $\mu = \mathbb{E}_f(Y)$ with the bias of order n^{-1}, since

$$\mathbb{E}\,[\tilde{\mu}\,] = \mu\{1 + n^{-1}(\mu\mu_{-1} - 1)\} + o(\tfrac{1}{n}), \tag{1.3.18}$$

showing the leading bias in the estimator $\tilde{\mu}$ as $n^{-1}\mu(\mu\mu_{-1} - 1)$. The bias reduction in the estimator of μ can be achieved by jackknifing, as suggested by Sen (1987). It is of interest to investigate the properties of the estimate of the mean μ for finite samples when the mean is considered a linear function of some covariate terms.

Krieger and Pfefferman (1992), considered approaches for maximum likelihood estimation from survey data. The two approaches, the classical pseudo likelihood (CPL) and MLE derived from weighted distributions (WDML) were compared for their perfomance using a simulation study.

The CPL utilizes the sample selection probabilities to estimate the census likelihood equations. Suppose the population values $\underset{\sim}{Y}_i$ are independent with a common distribution $f(\underset{\sim}{Y};\boldsymbol{\theta})$ and $L_N(\boldsymbol{\theta};\underset{\sim}{Y}) = \sum_{i=1}^{N} \log f(\underset{\sim}{Y}_i;\boldsymbol{\theta})$ define the census log-likelihood function. The MLE , $\hat{\boldsymbol{\theta}}$ is a solution to the equation,

$$\underset{\sim}{U}(\boldsymbol{\theta}) = \partial L_N(\boldsymbol{\theta};\underset{\sim}{Y}) / \partial\boldsymbol{\theta} = \sum_{i=1}^{N} u(\boldsymbol{\theta};\underset{\sim}{y}_i) = \underset{\sim}{0} \tag{1.3.19}$$

where $u(\boldsymbol{\theta};\underset{\sim}{y}_i) = \frac{\partial}{\partial\boldsymbol{\theta}} \log f(\underset{\sim}{Y}_i;\boldsymbol{\theta})$.

Assume only the sample values $\{y_i , i \in s\}$ and the sample selection probabilities $\{P_i \; i \in s\}$ are known. The pseudo MLE uses the estimator $\hat{\underset{\sim}{U}}(\boldsymbol{\theta})$ of $\underset{\sim}{U}(\boldsymbol{\theta})$ and is thus the solution to

$$\hat{\underset{\sim}{U}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} w^*_i \, u(\boldsymbol{\theta};\underset{\sim}{y}_i), \tag{1.3.20}$$

where for selection without replacement $w^*_i = [1/P_i \; (i \in s)]$, and for selection with replacement $w^*_i = (1/nP)$.

The second approach WDML is based on the concept of weighted distributions and utilizes the sample selection probabilities to adjust the original unweighted distribution. The MLE of $\boldsymbol{\theta}$ is then the solution to the maximum likelihood estimating equations of the weighted distribution. Krieger and Pfefferman illustrate the comparison with an example in which $\underset{\sim}{X}'_i = (\underset{\sim}{Y}'_i,\underset{\sim}{Z}'_i)$ are

independent realizations from the multivariate normal distribution with mean $\underset{\sim}{\mu}_X = (\underset{\sim}{\mu}'_Y , \underset{\sim}{\mu}'_Z)$ and variance-covariance, V-C, matrix

$$\Sigma_{XX} = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{bmatrix}.$$

The results obtained when estimating $\mu_Y = \mathbb{E}(Y)$ and $\sigma^2 = \text{Var } (Y)$ were compared for the two distributions for their Residual Mean Square Error (RMSE). It is shown from a simulation study that WDML dominates CPL. Krieger and Pfefferman state as a possible explanation for this result the fact that WDML is more "model dependent" whereas the CPL is viewed as "estimator maximizing the design unbiased estimator of the likelihood equations holding for the population".

Godambe and Rajarshi (1989) extended the idea of optimal estimating equations for semi-parametric models to weighted distributions for a finite population of size N. Assuming observational samples y_i are made with probability $w(y_i)$ they considered the maximum likelihood equations for $\boldsymbol{\theta}$ based on the sample data d, from a finite population $\mathcal{P}$

$$d=\{(i , y_i) : i \in s \}, \qquad s \subset \mathcal{P}.$$

The probability density of d is given by

$$\text{Prob}(d; \boldsymbol{\theta}) = \prod_{i \in s} \frac{w(y_i)\, f(y_i;\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \prod_{i \in s} \mu(\boldsymbol{\theta}) \prod_{\notin s} (1-\mu(\boldsymbol{\theta})), \tag{1.3.21}$$

where $\mu(\boldsymbol{\theta}) = \int w(y_i) f(y_i; \boldsymbol{\theta})\, dy_i$, for $i = 1,2,\ldots,N$.

The maximum likelihood estimator for $\boldsymbol{\theta}$ is then found in the usual manner as

a solution to the maximum likelihood equations given by

$$\sum_{i \in s} \frac{\partial}{\partial \theta_j} \log \frac{w(y_i) f(y_i;\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} + \frac{n - \mu(\boldsymbol{\theta}) N}{\mu(\boldsymbol{\theta})(1-\mu(\boldsymbol{\theta}))} \frac{\partial \mu(\boldsymbol{\theta})}{\partial \theta_j} = 0, \qquad j = 1,2,\ldots,k, \qquad (1.3.22)$$

where n is the sample size.

The maximum likelihood equations in (1.3.22) include an additional term

$$\frac{n - \mu(\boldsymbol{\theta}) N}{\mu(\boldsymbol{\theta})(1-\mu(\boldsymbol{\theta}))} \frac{\partial \mu(\boldsymbol{\theta})}{\partial \theta_j} \qquad (1.3.23)$$

to the usual ML equations, given by the first term on the r.h.s. of (1.3.22). The population size N can be treated as a parameter to be jointly estimated with $\boldsymbol{\theta}$ from the distribution (1.3.21).

The theory of estimating functions as discussed by Godambe and Rajarshi provides equations that are jointly optimal for $\boldsymbol{\theta}$ and N. They considered an experiment in which each individual $i \in \mathcal{P}$ has a chance of selection $m(\boldsymbol{\theta})$, and y_i is drawn from the distribution $g(y;\boldsymbol{\theta})$ in (1.2.1), such that the sample data consists of (i, y_i) . To estimate $\boldsymbol{\theta}$, the 2N estimating functions are constructed for each $i \in \mathcal{P}$ and j, as

$$h_i = \phi_i \frac{\partial}{\partial \theta_j} \log \frac{w(y_i)\, f(y_i;\, \boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \quad \text{and} \quad h_{i+N} = \phi_i - \mu(\boldsymbol{\theta}), \qquad i = 1,2,\ldots,N\ , \qquad (1.3.24)$$

where $\phi = 1$ if the i^{th} individual is selected in the sample and 0 otherwise. Since $E_g(h_i) = 0$ and $E_g(h_i h_{i'}) = 0$, for $i,\ i' = 1,2,\ldots,N$, the $2N$ estimating functions are mutually orthogonal. Then according to Godambe and Rajarshi, the optimal estimating combination of the h_i's is given by

$$h = \sum_{i=1}^{N} \{ h_i \frac{E_g \partial h_i / \partial \theta_j}{E(h^2_i)} + h_{i+N} \frac{E_g \partial h_{i+N} / \partial \theta_j}{E(h^2_{i+N})} \}$$

$$= \sum_{i=1}^{N} \phi_i \frac{\partial}{\partial \theta_j} \log \frac{w(y_i) f(y_i;\theta)}{\mu(\theta)} + \sum_{i=1}^{N} (\phi_i - \mu(\theta)) \frac{\partial \mu(\theta)/\partial \theta_j}{\mu(\theta)(1-\mu(\theta))}, \quad (1.3.25)$$

which reduces to equation (1.3.22). Godambe and Rajarshi suggest that if N is unknown, the functions h_{i+N}, $i = 1, 2, \ldots, N$ be combined into a single function $\sum_{i=1}^{N} h_{i+N} = n - \mu(\theta)N$, which is orthogonal to the h_i's. Hence equations $h = 0$ and $n - \mu(\theta)N = 0$ provide joint optimal estimation for θ and N.

Replacing the distributional assumption for the y_i's by assuming only the mean μ, variance σ^2, and coefficient of skewness γ, Godambe and Rajarshi extends the idea of estimating equations to the semi-parametric models as follows: Let

$$E_g[Y_i] = \frac{1}{\mu}(\sigma^2 + \mu^2) = c(\theta), \quad (1.3.26)$$

$$E_g[Y_i - c(\theta)]^2 = \sigma^2 \{ (\frac{\gamma\sigma}{\mu}) + 1 - (\frac{\sigma^2}{\mu^2}) \}. \quad (1.3.27)$$

For $j = 1, \ldots, k$, define the mutually orthogonal estimating equation as

$$h_i = \phi_i(y_i - c(\theta)) \quad \text{and} \quad h_{i+N} = \phi_i - \mu(\theta), \text{ for } i=1,2,\ldots,N \quad (1.3.28)$$

Then the optimum combination of the estimating functions is given by

$$h = \sum_{i=1}^{N} \phi_i(y_i - c(\theta)) \frac{\partial c(\theta)/\partial \theta_j}{E_g[y_i - c(\theta)]^2} + \sum_{i=1}^{N} (\phi_i - \mu(\theta)) \frac{\partial \mu(\theta)/\partial \theta_j}{\mu(\theta)(1-\mu(\theta))},$$

for $j = 1, \ldots, k$ (1.3.29)

where $\phi_i = 1$ if y_i is in the sample and $\phi_i = 0$ otherwise, $c(\theta)$ and $E_g[y_i - c(\theta)]^2$ are

given by (1.3.26) and (1.3.27), respectively,

and

$$\frac{\partial \mu(\boldsymbol{\theta})}{\partial \theta_j} = -\Big\{ \frac{1}{\mu(\boldsymbol{\theta})} \frac{\partial \sigma^2(\boldsymbol{\theta})}{\partial \theta_j} - [\frac{\sigma^2(\boldsymbol{\theta})}{\mu^2(\boldsymbol{\theta})} - 1] \frac{\partial \mu(\boldsymbol{\theta})}{\partial \theta_j} \Big\} \tag{1.3.30}$$

which gives

$$h = \sum_{i \in s} (y_i - c(\boldsymbol{\theta})) \frac{\Big\{(\sigma/\mu)^2 - 1\Big\}(\partial \mu/\partial \theta_j) - (\partial \sigma^2/\partial \theta_i)/\mu}{\sigma^2\Big\{(\gamma\sigma/\mu) + 1 - (\sigma/\mu)^2\Big\}}$$

$$+ \sum_{i \in s} (\partial \log \mu(\boldsymbol{\theta})/\partial \theta_j) + \sum_{i \notin s} \{\partial \log (1-\mu(\boldsymbol{\theta}))/\partial \theta_j\}. \tag{1.3.31}$$

The semi-parametric estimator of $\boldsymbol{\theta}$ is obtained from (1.3.31) by solving the optimal estimating equation $h = 0$ for $\boldsymbol{\theta}$.

1.3.5 Density Estimation

Most research has been devoted to estimating the probability distribution function F of f and the mean of the distribution based on the length-biased samples. Winter and Foldes (1988) proposed a nonparametric estimator of the cumulative hazard function and used it to construct an estimator of F. They considered a product-limit estimator for the life-testing situation where the life of n objects observed has the distribution G, the length-biased form of F, and established strong uniform consistency results for their estimator. Bhattacharyya, Franklin and Richardson (1988) compared the mean squre errors of the density estimators of the weighted and unweighted distributions. Assuming $X_1,\ldots,X_n$ are i.i.d. random variables with density $g(x)=xf(x)/\mu$, for $x \geq 0$ they considered an estimator $\hat{g}(x)$, the kernel estimator as discussed by Parzen (1962). The estimator

is of the form

$$f_n(x) = \frac{1}{nh_n}\sum K\{(x - X_j)/h_n\}, \tag{1.3.32}$$

where $\{h_n\}$ is a sequence of constants converging to zero, and K(y) is a bounded integrable function satisfying the conditions $\lim |yK(y)| = 0$.

Parzen gives the asymptotic variance of this estimator, under precise conditions assumed concerning the function K, and show asymptotic normality. The motivation of this estimator is that a common estimator of the distribution function F(x) is the empirical distribution function defined as

$F_n(x) = \frac{1}{n}\sum u(x - X_j)$, where $u(y) = 1$ if $y \geq 0$ and $u(y) = 0$ otherwise.

Bhattacharrya (1988) proposed estimating $\hat{f}(x)=\hat{\mu} \times \hat{g}(x)/x$, where $\hat{\mu}$ is an estimator of μ, $\hat{\mu} = n[\sum x_i^{-1}]^{-1}$. It was shown that the mean square error of the length biased estimator approaches zero as n increases to infinity.

Jones (1991) derived a direct kernel density estimator for the length-biased data of the form

$$\hat{f}_n(x) = n^{-1}\hat{\mu}\sum X_j^{-1}K_h(x - X_j), \tag{1.3.33}$$

where h is the smoothing parameter and $K_h=h^{-1}K(x/h)$. The measure of error used for this estimator is the mean integrated squared error (IMSE) defined as

$$E\int [\hat{f}_n(x) - f(x)]^2 dx . \tag{1.3.34}$$

Jones compared the integrated mean squared errors of the estimators based

on the weighted and unweighted models. Using simulated sample data from the density f and generated length-biased data from g the results, (Jones 1991) show the estimator based on length-biased data is inferior to the estimator based on the unweighted sample in terms of integrated mean square error, IMSE. Also the IMSE $(\hat{f})$ is less than the IMSE of the estimator provided by Bhattacharyya.

1.3.6 Discussion and Summary of Literature Review

Work on length-biased data analysis has focused mainly on estimation of the mean μ and the distribution function F of the original distribution based on length-biased data. The length-biased density function considered in the literature is of the form $g(x) = x \times f(x)/\mu$, $x \geq 0$ and $g(x) = 0$, $x \leq 0$, where μ is the mean under the unweighted distribution.

A natural consistent estimator of μ is given by $\tilde{\mu} = n[\sum x^{-1}]^{-1}$, Cox (1969) and Sen (1987). The estimators for μ and F from the length-biased data from g(x) are compared with those obtained from the original unweighted data from f(x) by finding their asymptotic variances. The estimators are derived in a nonparametric setup, and hence do not take advantage of the distributional information of the data. Furthermore, the mean is not considered dependent on a set of possible explanatory variables which might be available in research situations.

Godambe and Rajarshi considered a parametric model and semi-parametric models for the length-biased density for a finite population of size N. Their work is limited to estimating the parameters of the length-biased distribution and does not consider estimation of the functions of the parameters and their properties. Moreover, it is not clear how the optimality of the estimating functions they

consider translates to the optimality of the parameters. It is also not clear how the number of the arbitrary estimating functions, h_i's, is to be determined, especially in an infinite population. No attempt is made to compare their results with estimates that would be obtained from under the assumed original unweighted model.

It is the purpose of this research to extend the estimation to the situation where μ depends on a set of covariates in a regression setup, and to investigate the properties of such estimators. The two sampling designs, sampling under f(x) and sampling under g(x) will be compared for their lack of information about the parameters by way of optimality criteria using Fisher information matrix. The next section discusses the proposed research on length-biased distributions.

1.4 Research Proposal

The purpose of this research is to obtain estimates of the parameters of the original (unweighted) distribution from the data generated by a length-biased design. We study the large sample properties of the maximum likelihood and the jackknifed estimators for the two designs. Both the MLE and the jackknifed estimators (originally proposed by Quenoulli) are shown to have equivalent asymptotic normal distribution. The reduction in the bias of the MLE achieved by the jackknife estimator is provided. The MLE and the jackknife estimators of the regression model are also studied for their large sample properties, both the bias and the asymptotic normal distributions.

Our work differs from that in the existing literature in that we assume a parametric estimation of the parameters. Whereas attention in the existing literature emphasizes estimating a single parameter μ, the mean of the random

variate, from a length-biased distribution, we consider estimating a k-parameter vector $\boldsymbol{\theta}$ and for the regression model. We assume μ depends on a set of nonstochastic covariates whose vector of coefficients $\boldsymbol{\beta}$ is the primary objective of the estimation. Furthermore, this research aims at comparing the information about the parameter estimates from the two designs by way of the Fisher information.

We concern ourselves with a random variate Y, following an exponential probability distribution with pdf given by;

$$f(y;\boldsymbol{\theta}) = \exp\{ \sum_{j=1}^{k} \theta_j T_j(y) + \Psi(\boldsymbol{\theta}) + C(y) \}, \tag{1.4.1}$$

where $\boldsymbol{\theta}$ is a k-dimensional vector of unkown parameters. Since $f(y;\boldsymbol{\theta})$ is of exponential family type of distributions, its length-biased version $g(y;\boldsymbol{\theta})$ also belongs to the exponential family of distributions and is given by,

$$g(y;\boldsymbol{\theta}) = \exp\{ \sum_{j=1}^{k} \theta_j T_j(y) + \Psi^*(\boldsymbol{\theta}) + C^*(y)\}, \tag{1.4.2}$$

where $\Psi^*(\boldsymbol{\theta}) = \Psi(\boldsymbol{\theta}) - \log\mu(\boldsymbol{\theta})$, $\mu(\boldsymbol{\theta}) = E[Y]$ and $C^*(y) = C(y) + \log(y)$.

Our purpose is to consider estimation of the parameters or functions of the parameters $\boldsymbol{\theta}$ of the model (1.4.1) from the data generated by the probability distribution model (1.4.2). The asymptotic bias and distribution of such estimators is also established. The performance of the estimators is measured by the root mean squared error (RMSE), which is a combination of bias and variance of the estimator.

This dissertation is organized as follows:

Chapter 2 discusses the large sample properties of the MLE and the jackknife

estimator of the MLE from the data following the probability distribution (1.4.1). The jackknife variance estimator is shown to be consistent estimator of the variance of the MLE. In particular, the proposed jackknife procedure is applied to the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. The jackknife method of estimation generates replicates of the statistic $\widehat{\boldsymbol{\theta}}$ by deleting one observation from the sample and then computes the statistic from the remaining data. The bias of the jackknife estimator is compared with the bias of the MLE. It is shown that the jackknife estimators have the same asymptotic normal distribution as the MLE's.

Chapter 3 extends the results of Chapter 2 to the case of estimating the parameters of the regression model. Regression analysis concerns the study of the relationship between the response variable Y and a set of predictor variables $\mathbf{Z} = (\mathrm{Z}_1, \ldots, \mathrm{Z}_p)^T$. Given that the data follows an exponential family of distributions models, we hope to obtain the small sample results for the estimators, failing which we shall concern ourselves only with asymptotic results.

The model considered is given by

$$\mathbb{E}[\mathrm{Y}/\boldsymbol{Z}] = \mu = \mu(\boldsymbol{\beta}, \phi), \tag{1.4.3}$$

where $\boldsymbol{\beta}$ is a p-vector parameter of regression coefficients and ϕ is the scale parameter. The $\boldsymbol{Z}$ matrix is nonstochastic and is assumed to be known. The parameter of interest $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is estimated by maximum likelihood estimation methods. By deleting each row of Y and $\boldsymbol{Z}$, the jackknife estimator is derived. We study the large sample distribution of the MLE and the jackknife estimators.

Chapter 4 discusses the MLE and the jackknife estimators under the length-biased distributions. The bias and large sample properties of the estimators are derived. We compare the Fisher information matrix of the parameters of the

original distribution with the Fisher information matrix from the length-biased distribution, by way of D-optimality and E-optimality (Keifer, 1975).

Chapter 5 compares the jackknife estimates and the MLE computed from (1) the 1990 Sudan DHS birth interval data, and (2) the simulated data. This is intended to demonstrate the applications of our methodology. Two distributions considered in the simulation study for the exponential family of distributions are the lognormal and the gamma distributions. Simulation studies are undertaken on the data generated from these distributions to illustrate the computations of jackknife estimators. These studies will indicate how well our estimators perform in small to moderate sample sizes. Attention is centered on estimating and making statistical inferences about the parameters of the original distributions based on the length-biased data. Hence the bias and variance estimation is of significant importance in this study. Chapter 6 summarizes the results of the dissertation and provides suggestions for future research.

Chapter 2.

Parametric Estimation From The Exponential Family.

2.1 Introduction.

Our objective in this chapter is to derive expressions for the asymptotic bias of the MLE and the Jackknife Estimators of the parameters $\boldsymbol{\theta}$ of the Exponential family of distributions. We also study the asymptotic distribution of the estimators and show the MLE and Jackknife estimators of the parameters of the Exponential family of distributions are asymptotically equivalent, in the sense that $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ and $\sqrt{n}(\widetilde{\boldsymbol{\theta}}^J - \boldsymbol{\theta})$ have identical asymptotic normal distribution, where $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}^J$ are the MLE and jackknifed estimators respectively. The convergence of the sample variance of the pseudo-values (Tukey) to the variance of the MLE is shown.

2.2 Maximum Likelihoods Estimators

We consider the Maximum Likelihood Estimators (MLE) of parameters of the Exponential family of distributions. In this section we show the conditions

under which the MLE of the parameters of the Exponential family are consistent. The asymptotic bias and distribution of the MLE from the exponential family of distributions are derived.

Suppose that $X_1,\ldots,X_n$ are independent and identically distributed (i.i.d.) random variables with density function $f(X;\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset R^k$. For the Exponential family considered in this section, the density is $f(X;\boldsymbol{\theta})$ the form

$$f(X;\boldsymbol{\theta}) = c(X)\exp\Big\{ \sum_{j=1}^{k} \theta_j T_j(X) + \Psi(\boldsymbol{\theta}) \Big\} \tag{2.2.1}$$

where $c(.)$ and $\Psi(.)$ are some known functions. The likelihood function for the sample $X_1,\ldots,X_n$ of size n is defined by

$$L_n(\boldsymbol{\theta};\, X) = \prod_{i=1}^{n} f(X_i\,;\boldsymbol{\theta}). \tag{2.2.2}$$

Whatever its value, it is well known that the likelihood function is maximized at the true value $\boldsymbol{\theta} \in \Theta$. Thus the principle of maximum likelihood estimation consists of choosing an estimate of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$ that maximizes (2.2.2). It is often convenient to work with the log-likelihood function and since the log function is monotone, the value of $\hat{\boldsymbol{\theta}}$ that maximizes $\log L_n(\boldsymbol{\theta};\, X)$ will also maximize $L_n(\boldsymbol{\theta};\, X)$. Thus the MLE of $\boldsymbol{\theta}$ may be obtained by maximizing the sample log-likelihood function and in which case, is the solution to the likelihood equation,

$$U_n(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L_n(\boldsymbol{\theta}\,;X) \;=\; \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_i;\, \boldsymbol{\theta}) \;= 0. \tag{2.2.3}$$

The usual assumptions made for the MLE to exist and have asymptotic

normality properties are:

A1. $\frac{\partial}{\partial\theta_j}\log f(X_i;\,\boldsymbol{\theta})$ and $\frac{\partial}{\partial\theta_j\partial\theta_l}\log f(X_i;\,\boldsymbol{\theta})$ for $j,\ l\ =\ 1,\ldots,k$ exist almost everywhere, and

$$\left|\frac{\partial}{\partial\theta_j}\log f(X_i;\,\boldsymbol{\theta})\right| \le K_1(X), \quad \left|\frac{\partial^2}{\partial\theta_j\partial\theta_l}\log f(X_i;\,\boldsymbol{\theta})\right| \le K_2(X)$$

where $K_1(X)$ and $K_2(X)$ are integrable functions, and $\int K_i(X) < \infty \quad (i=1,2)$.

A2. $0 < I(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(X;\,\boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(X;\,\boldsymbol{\theta})\right\}^T < \infty$

$I(\boldsymbol{\theta})$ is called the Fisher information matrix.

A3. $E_{\boldsymbol{\theta}}\left\{\sup_{|h|<\delta}\left\|\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log f(X_i;\,\boldsymbol{\theta}) - \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}}\log f(X_i;\,\boldsymbol{\theta}+h)\right\|\right\} \longrightarrow 0, \quad \text{as } \delta\longrightarrow 0.$

The implications of the above set of assumptions for the particular exponential family of distributions (2.2.1) studied here lead to the following assumptions :

B1. $\Psi(\boldsymbol{\theta}) = \Psi(\theta_1,\ldots,\theta_n)$ has continuous partial derivatives at all $\theta_j \quad j=1,\ldots,k$

B2. $\left(-\frac{\partial^2}{\partial\theta_j\partial\theta_l}\Psi(\boldsymbol{\theta})\right), \quad j,\ l=1,\ldots k,$ is a $(k\ \times\ k)$ positive definite matrix for all $\boldsymbol{\theta}\in\Theta$.

B3. $E_{\boldsymbol{\theta}}\left|T_j(X)\right| < \infty, \quad j=1,\ldots,k \quad \boldsymbol{\theta}\in\Theta.$

To construct the estimating equations for $\boldsymbol{\theta}$ we consider the individual contribution of each individual data point. The partial derivatives of the log likelihood of (2.2.1) with respect to $\boldsymbol{\theta}$ as,

for $j, l = 1,\ldots,k$

$$\frac{\partial}{\partial\theta_j}\log f(X;\boldsymbol{\theta}) = T_j(X_i) + \frac{\partial}{\partial\theta_j}\Psi(\boldsymbol{\theta}) \tag{2.2.4}$$

and

$$\frac{\partial^2}{\partial\theta_j\partial\theta_l}\log f(X;\boldsymbol{\theta}) \;\; = \frac{\partial^2}{\partial\theta_j\partial\theta_l}\Psi(\boldsymbol{\theta}). \tag{2.2.5}$$

The MLE $\hat{\boldsymbol{\theta}}_n$, of $\boldsymbol{\theta}$ is a solution to the MLE equations (2.2.3), so that for the probability density $f(X;\boldsymbol{\theta})$ given by (2.2.1),

$$U_n(\hat{\boldsymbol{\theta}}_n) = \sum_{i=1}^{n} \underset{\sim}{T}(X_i) + n\,\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_n\,) \;= 0, \tag{2.2.6}$$

which implies that

$$\frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{T}(X_i) + \boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_n) \;= 0, \tag{2.2.7}$$

where $\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_n) = \frac{\partial}{\partial\boldsymbol{\theta}}\Psi(\boldsymbol{\theta})\,|_{\hat{\boldsymbol{\theta}}_n}$ and $\underset{\sim}{T}(X_i) = (\,T_1(X_i),\ldots,T_k(X_i)\,)^T$.

Since

$$\mathbf{0} = E_{\boldsymbol{\theta}}\frac{\partial}{\partial\boldsymbol{\theta}}\log f(X_i;\,\boldsymbol{\theta}) = \underset{\sim}{T}(X_i) + \boldsymbol{\psi}(\boldsymbol{\theta}), \tag{2.2.8}$$

we have

$$E_{\boldsymbol{\theta}} \underset{\sim}{T}(X_i) = -\boldsymbol{\psi}(\boldsymbol{\theta}). \tag{2.2.9}$$

The first term on the LHS of (2.2.7) is the mean of a sample of size n from a population having mean $-\boldsymbol{\psi}(\boldsymbol{\theta})$, so by Khintchine's SLLN, as $n \longrightarrow \infty$,

$$\frac{1}{n} \sum_{i=1}^{n} \underset{\sim}{T}(X_i) \xrightarrow{\text{a.s.}} = -\boldsymbol{\psi}(\boldsymbol{\theta}). \tag{2.2.10}$$

By (2.2.3)

$$\frac{\partial U_n'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log f(X_i; \boldsymbol{\theta}) = n\, \boldsymbol{\varphi}(\boldsymbol{\theta}) = n \left(\frac{\partial^2}{\partial \theta_j \partial \theta_l} \Psi(\boldsymbol{\theta}) \right) \tag{2.2.11}$$

and differentiating (2.2.1) twice with respect to $\boldsymbol{\theta}$, under the integral sign, and using B2, we have

$$-\boldsymbol{\varphi}(\boldsymbol{\theta}) = \left(- \frac{\partial^2}{\partial \theta_j \partial \theta_l} \Psi(\boldsymbol{\theta}) \right) = \mathrm{Var}_{\boldsymbol{\theta}}(\underset{\sim}{T}(X)) > 0, \tag{2.2.12}$$

we see that the log-likelihood is concave. Therefore if a solution to the likelihood equation (2.2.3) exists, it must be an MLE.

By assumption B1, $\boldsymbol{\psi}(\boldsymbol{\theta}^*)$ is continuous in a neighborhood of the true parameter value $\boldsymbol{\theta} \in \Theta$, i.e. for every $\epsilon > 0$ there exist $\eta > 0$ such that

$$\| \boldsymbol{\psi}(\boldsymbol{\theta}^*) - \boldsymbol{\psi}(\boldsymbol{\theta}) \| < \epsilon, \qquad \text{for all } \| \boldsymbol{\theta}^* - \boldsymbol{\theta} \| < \eta. \tag{2.2.13}$$

If $-\psi(\theta) = E_{\theta}[\underset{\sim}{T}(X)]$ is a one-to-one function of θ, then (2.2.13) necessarily hold.

Example 1: Let $X_1,\ldots,X_n$ be i.i.d. r.v.'s with a lognormal $(\mu\ ,\ \sigma^2)$ distribution. The likelihood function is

$$L_n(\mu,\sigma^2) = (2\pi\sigma)^{-n/2}(\prod_{i=1}^{n} X_i)^{-1} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\log X_i - \mu)^2\},$$

Then $\theta_1 = \frac{\mu}{\sigma^2}$, $\theta_2 = -\frac{1}{2\sigma^2}$, $T_1(X_1) = \log X_1$, $T_2(X_1) = (\log X_1)^2$ and

$\Psi(\theta) = \frac{1}{2}(\frac{\theta_1^2}{2\theta_2} - \log\frac{-1}{2\theta_2})$, so that we have

$$E_{\theta}\underset{\sim}{T}(X_i) = -\psi(\theta) = \frac{1}{2}(-\frac{\theta_1}{\theta_2},\ \frac{\theta_1^2}{2\theta_2^2} - \frac{1}{\theta_2})^T = (\mu,\ \mu^2+\sigma^2)^T,$$

which is a one-to-one function of θ and hence (2.2.13) is satisfied.

Example 2 : Let $X_1,\ldots,X_n$ be i.i.d. r.v.'s following a Poisson(λ) distribution. The likelihood of θ is $L_n(\lambda) = \exp\{-\lambda + X\log\lambda - \log X!\}$. So that $T(X) = X$, $\theta = \log\lambda$.

$E_{\theta}T(X) = -\psi(\theta) = \lambda = \exp(\theta)$ which is a one-to-one monotone function of θ satisfying (2.2.13).

There are situations when the likelihood equation (2.2.3) has at least one consistent solution, so that the likelihood function has a relative maxima at such solutions. When the likelihood has multiple roots we will require that

$$|| \psi(\theta^{**}) - \psi(\theta) || > 0\ , \qquad \text{if} \quad || \theta^{**} - \theta || > c \tag{2.2.14}$$

where $\boldsymbol{\theta}^{**}$ is outside the η neighborhood of the true parameter $\boldsymbol{\theta}$ and $c > 0$. That is, we can choose c sufficiently small in the neighborhood of $\boldsymbol{\theta}^{**}$ such that (2.2.14) is satisfied. This requirement ensures the likelihood evaluated at $\boldsymbol{\theta}$ is larger than at any other $\boldsymbol{\theta}^{**} \in \Theta$.

Now since the first term on the LHS of (2.2.7) converges a.s. to $-\boldsymbol{\psi}'(\boldsymbol{\theta})$ and $U_n(\widehat{\boldsymbol{\theta}}) = 0$,

$$\boldsymbol{\psi}(\widehat{\boldsymbol{\theta}}) \xrightarrow{\text{a.s.}} \boldsymbol{\psi}(\boldsymbol{\theta}) \tag{2.2.15}$$

By B2 and (2.2.13) we have $\widehat{\boldsymbol{\theta}} \xrightarrow{\text{a.s.}} \boldsymbol{\theta}$.

2.2.1 Asymptotic Distribution and Bias of the MLE

The asymptotic distribution of MLE have been studied extensively in the literature, (Kendal and Stuart, 1958; Cramer, 1946; LeCam, 1957; Wilks, 1963 and others). In this study we apply the theory of asymptotic properties of the MLE specifically to the exponential family (2.2.1), furthermore the form of bias of the MLE $\widehat{\boldsymbol{\theta}}_n$ is presented to justify the jackknife estimation to reduce bias of the MLE in large samples. Our results relies heavily on Taylor expansion of the gradient of the log-likelihood function in some neighborhood of the true parameter $\boldsymbol{\theta}$.

To establish the asymptotic normality of $\widehat{\boldsymbol{\theta}}_n$ we let

$$U_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i; \theta) \tag{2.2.16}$$

using (2.2.8), assumption B2 and the Central Limit Theorem on $\partial/\partial\theta \log f(X_i; \theta)$, $(i = 1,2,\ldots,n)$ we note that $U_n(\theta) \longrightarrow^{\mathfrak{D}} N\left(0, I(\theta)\right)$, where $I(\theta)$ is the Fisher information matrix.

Expanding $U_n(\hat{\theta}_n)$ in the neighborhood of θ we write

$$0 = U_n(\hat{\theta}_n) - U_n(\theta) = U_n'(\theta)|_{\theta^*}(\hat{\theta}_n - \theta), \tag{2.2.17}$$

where θ^* is a point on the line segment connecting $\hat{\theta}_n$ and θ. Since from (2.2.13) and (2.2.15), $\hat{\theta}_n$ is strongly consistent for θ and for the exponential family of distributions (2.2.1), and with the assumption that $\Psi(\theta)$ are continuously differentiable we have,

$$\varphi(\theta^*) = \varphi(\theta) + o_p(\tfrac{1}{\sqrt{n}}), \qquad \text{as } n \longrightarrow \infty. \tag{2.2.18}$$

Hence we write (2.2.17) as

$$\psi(\hat{\theta}_n) - \psi(\theta) = \varphi(\theta)|_{\theta^*}(\hat{\theta}_n - \theta)$$

$$= \varphi(\theta)(\hat{\theta}_n - \theta) + o_p(\tfrac{1}{\sqrt{n}}), \tag{2.2.19}$$

where the last equality in (2.2.19) is by virtue of (2.2.18) and by using the fact that $(\hat{\theta}_n - \theta)$ is O(1). Now by assumption B2, $-\varphi(\theta)$ is positive definite. From (2.2.19) we get

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \sqrt{n}\,[\boldsymbol{\varphi}(\boldsymbol{\theta})]^{-1}\,[\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_n) - \boldsymbol{\psi}(\boldsymbol{\theta})] + o_p(1), \qquad \text{as } n \longrightarrow \infty. \tag{2.2.20}$$

So that by Slutsky's Theorem, the LHS of (2.2.20) converges in distribution to the same asymptotic distribution of $\sqrt{n}\,[\boldsymbol{\varphi}(\boldsymbol{\theta})]^{-1}\,[\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}_n) - \boldsymbol{\psi}(\boldsymbol{\theta})]$. Now the first term on the RHS of (2.2.20) is

$$\mathrm{I}^{-1}(\boldsymbol{\theta})\ \mathrm{U}_n(\boldsymbol{\theta}) \longrightarrow^{\mathfrak{D}} \mathrm{N}\left(0,\ \mathrm{I}^{-1}(\boldsymbol{\theta})\right), \quad \text{as } n \longrightarrow \infty \text{ and hence}$$

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \longrightarrow^{\mathfrak{D}} \mathrm{N}\left(0,\ \mathrm{I}^{-1}(\boldsymbol{\theta})\right), \tag{2.2.21}$$

thus establishing the asymptotic normality of the MLE $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ for the exponential family (2.2.1). The result (2.2.21) for general distributions is classical and is contained in Cramer, 1946 and Kendal and Stuart, 1958 among other books .

Although the MLE generally possess nice asymptotic properties of consistency and asymptotic normality, they are often not unbiased estimators. We wish to show the form of the bias in the MLE and later show how the jackknifing method can reduce the bias. To study the bias of $\hat{\boldsymbol{\theta}}_n$, note that $-\boldsymbol{\psi}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i)]$ is a one-to-one function of $\boldsymbol{\theta}$ (eg. if X follows a Poisson distribution with mean λ, then $\mathrm{E}[\mathrm{T}(\mathrm{X})] = -\boldsymbol{\psi}(\boldsymbol{\theta}) = \mathrm{e}^{\theta} = \lambda$, so that there exists a transformation $\mathbf{H} = \mathbf{H}(-\boldsymbol{\psi}(\boldsymbol{\theta}))$ satisfying

$$\mathbf{H}(-\boldsymbol{\psi}(\boldsymbol{\theta})) = \boldsymbol{\theta}. \tag{2.2.22}$$

Proposition 1. Assume the set of assumptions "B". Define $\boldsymbol{\eta} = -\boldsymbol{\psi}(\boldsymbol{\theta})$ and a

function $\mathbf{H}$: $\mathbf{R}^k \longrightarrow \mathbf{R}^k$ given by $\mathbf{H}(\boldsymbol{\eta}) = \boldsymbol{\theta}$. Assume $\mathbf{H}(\boldsymbol{\eta})$ is continuously differentiable in some neighborhood of $\boldsymbol{\eta}$. If $\mathrm{E}||\underset{\sim}{\mathrm{T}}(\mathrm{X})||^4 < \infty$ then

$$\mathrm{E}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}] = \frac{1}{n}\boldsymbol{a}(\boldsymbol{\theta}) + \frac{1}{n^2}\boldsymbol{b}(\boldsymbol{\theta}) + \mathrm{O}(\frac{1}{n^3}), \tag{2.2.23}$$

where $\boldsymbol{a}(\boldsymbol{\theta})$ and $\boldsymbol{b}(\boldsymbol{\theta})$ are unknown functions of $\boldsymbol{\theta}$.

Proof.

Let $\boldsymbol{\eta}$ and $\mathbf{H}(\boldsymbol{\eta})$ be as defined in proposition 1, and define the sample mean of $\underset{\sim}{\mathrm{T}}(\mathrm{X})$ as

$$\underset{\sim}{\bar{\mathrm{T}}}_n(\mathrm{X}) = \frac{1}{n}\sum_{i=1}^{n}\underset{\sim}{\mathrm{T}}(\mathrm{X}_i).$$

To obtain the central moment of $\underset{\sim}{\bar{\mathrm{T}}}_n = \underset{\sim}{\bar{\mathrm{T}}}_n(\mathrm{X})$, let us first define the corresponding central moments of the elements $\bar{\mathrm{T}}_{nj}(\mathrm{X})$ of $\underset{\sim}{\bar{\mathrm{T}}}_n$, as the product moment of the variates about their respective means η_j's and note that the expected value of the a random vector matrix is the expected value of it's random elements.

Writing $\underset{\sim}{\bar{\mathrm{T}}}_n$ for $\underset{\sim}{\bar{\mathrm{T}}}_n(\mathrm{X}_i)$ the first central moment of $\bar{\mathrm{T}}_{nj}$ is,

$$\mathrm{E}[\bar{\mathrm{T}}_{nj} - \eta_j] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[\mathrm{T}_j(\mathrm{X}_i) - \eta_j] = 0 \text{ so that}$$

$$\mathrm{E}[\underset{\sim}{\bar{\mathrm{T}}}_n - \boldsymbol{\eta}] = \mathbf{0}. \tag{2.2.24}$$

The second central moment of $\bar{\mathrm{T}}_{nj}$ is

$$E[\bar{T}_{nj}-\eta_j][\bar{T}_{nj'}-\eta_{j'}] = \frac{1}{n^2}\sum_{i=1}^{n} E[T_j(X_i)-\eta_j][T_{j'}(X_i)-\eta_{j'}] = \frac{1}{n}\sigma_{jj'},$$

so that

$$E[\underset{\sim}{\bar{T}}n-\boldsymbol{\eta}][\underset{\sim}{\bar{T}}n-\boldsymbol{\eta}]^T = \frac{1}{n}\underset{\sim}{\Sigma}_T, \tag{2.2.25}$$

where $\underset{\sim}{\Sigma}_T = (\sigma_{jj'})$ $j, j' = 1,2,\ldots,k$ is the covariance matrix of $\underset{\sim}{T}(X)$.

For the third central moment the $\bar{T}_n$ we have

$$E[\bar{T}_{nj}-\eta_j][\bar{T}_{nj'}-\eta_{j'}][\bar{T}_{nj''}-\eta_{j''}]$$

$$= \frac{1}{n^3}E\Big\{\sum_{i=1}^{n}[T_j(X_i)-\eta_j][T_{j'}(X_i)-\eta_{j'}][T_{j''}(X_i)-\eta_{j''}]\Big\}$$

$$= \frac{1}{n^2}E[T_j(X_i)-\eta_j][T_{j'}(X_i)-\eta_{j'}][T_{j''}(X_i)-\eta_{j''}]$$

$$= \frac{1}{n^2}\gamma_{jj'j''}$$

Note that the X_i's are independent and hence $E[T_j(X_i)-\eta_j][T_j(X_{i'})-\eta_j]=0$, $i \neq i'$ and $E[T_j(X_i)-\eta_j][T_j(X_{i'})-\eta_j][T_j(X_{i''})-\eta_j]=0$, for $i \neq i' \neq i''$.

Hence,

$$E[\underset{\sim}{\bar{T}}n-\boldsymbol{\eta}][\underset{\sim}{\bar{T}}n-\boldsymbol{\eta}]^T[\underset{\sim}{\bar{T}}n-\boldsymbol{\eta}] = \frac{1}{n^2}\underset{\sim}{J}_T, \tag{2.2.26}$$

where $\underset{\sim}{J}_T = \left(\gamma_{jj'j''}\right)$ $j, j', j'' = 1, \ldots, k$ is the third central moment of $\underset{\sim}{\mathrm{T}}(\mathrm{X})$.

Similarly the fourth central moment of $\underset{\sim}{\overline{\mathrm{T}}}_n(\mathrm{X})$ is

$$\mathrm{E}[\underset{\sim}{\overline{\mathrm{T}}}_n - \boldsymbol{\eta}][\underset{\sim}{\overline{\mathrm{T}}}_n - \boldsymbol{\eta}]^T[\underset{\sim}{\overline{\mathrm{T}}}_n - \boldsymbol{\eta}][\underset{\sim}{\overline{\mathrm{T}}}_n - \boldsymbol{\eta}]^T = \frac{1}{n^2}\underset{\sim}{K}_T, \tag{2.2.27}$$

where $\underset{\sim}{K}_T = \left(\lambda_{jj'j''j'''}\right)$ $j, j', j'', j''' = 1, \ldots, k$ is the fourth central moment of $\underset{\sim}{\mathrm{T}}(\mathrm{X})$

Similar results of the central moments of order k of the sample mean in a one dimensional variate (Cramer, 1946) are

$$\mathrm{E}[\overline{\mathrm{X}}_n - \mathrm{E}(\overline{\mathrm{X}}_n)]^{2k} = \frac{1}{n^k}\boldsymbol{a}_k(\boldsymbol{\theta}) + \mathrm{O}\left(\frac{1}{n^{(k+1)}}\right),$$

$$\mathrm{E}[\overline{\mathrm{X}}_n - \mathrm{E}(\overline{\mathrm{X}}_n)]^{2k-1} = \frac{1}{n^k}\boldsymbol{b}_k(\boldsymbol{\theta}) + \mathrm{O}\left(\frac{1}{n^{(k+1)}}\right) \tag{2.2.28}$$

Now consider the following Taylor's expansion of $\mathbf{H}(\underset{\sim}{\overline{\mathrm{T}}}_n)$ about $\boldsymbol{\eta} = \mathrm{E}[\underset{\sim}{\overline{\mathrm{T}}}_n]$:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} &= \mathbf{H}(\underset{\sim}{\overline{\mathrm{T}}}_n) - \mathbf{H}(\boldsymbol{\eta}) \\
&= \frac{\partial\, \mathbf{H}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}(\underset{\sim}{\overline{\mathrm{T}}}_n - \boldsymbol{\eta}) + \frac{1}{2}(\underset{\sim}{\overline{\mathrm{T}}}_n - \boldsymbol{\eta})^T \frac{\partial^2 \mathbf{H}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T \partial \boldsymbol{\eta}}(\underset{\sim}{\overline{\mathrm{T}}}_n - \boldsymbol{\eta}) \\
&+ \frac{1}{3!}(\underset{\sim}{\overline{\mathrm{T}}}_n - \boldsymbol{\eta})(\underset{\sim}{\overline{\mathrm{T}}}_n - \boldsymbol{\eta})^T \frac{\partial^3 \mathbf{H}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T \partial \boldsymbol{\eta}}(\underset{\sim}{\overline{\mathrm{T}}}_n - \boldsymbol{\eta})
\end{aligned}$$

$$+\tfrac{1}{4!}((\bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta})(\bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta})^T \frac{\partial^4 \mathbf{H}(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}^T\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T\partial\boldsymbol{\eta}} (\bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta})(\bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta})^T + \boldsymbol{\gamma}_n, \qquad (2.2.29)$$

where $\boldsymbol{\gamma}_n$ is a random variable which represents the remainder term. Let $\mathbf{H}(\boldsymbol{\eta}) = \left(\mathrm{h}_1(\boldsymbol{\eta}),\ldots,\mathrm{h}_l(\boldsymbol{\eta}),\ldots,\mathrm{h}_k(\boldsymbol{\eta})\right)^T$. Consider Taylor's expansion of a particular element $\mathrm{h}_l(\bar{\underset{\sim}{\mathrm{T}}}n)$ of $\mathbf{H}(\bar{\underset{\sim}{\mathrm{T}}}n)$, about $\boldsymbol{\eta}$ and, similar to (2.2.29), we have

$$\begin{aligned}
&\mathrm{h}_l(\bar{\mathrm{T}}n)-\mathrm{h}_l(\boldsymbol{\eta}) \\
&= \frac{\partial\,\mathrm{h}_l(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}} (\bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta}) + \tfrac{1}{2}(\bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta})^T \frac{\partial^2\mathrm{h}_l(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}^T\partial\boldsymbol{\eta}} (\bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta}) \\
&\quad + \tfrac{1}{3!}(\bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta})(\bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta})^T \frac{\partial^3\mathrm{h}_l(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T\partial\boldsymbol{\eta}} (\bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta}) \\
&\quad + \tfrac{1}{4!}(\bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta})(\bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta})^T \frac{\partial^4\mathrm{h}_l(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}^T\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T\partial\boldsymbol{\eta}} (\bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta})(\bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta})^T + \boldsymbol{\gamma}_{nl}. \qquad (2.2.30)
\end{aligned}$$

Note the expected value of the first term on the RHS of (2.2.30) is zero. To study the second and subsequent terms on the RHS of (2.2.30), we introduce the following notation:

Let

$$\mathbf{y} = \bar{\underset{\sim}{\mathrm{T}}}n-\boldsymbol{\eta}\,, \qquad \underset{\sim}{\mathbf{B}} = \mathbf{y}\mathbf{y}^T = \left(b_{ji}\right)\,,$$

where $b_{ji} = (\bar{\mathrm{T}}_{nj}-\eta_j)(\bar{\mathrm{T}}_{ni}-\eta_i)$ and

$$\underset{\sim}{\mathbf{A}} = \frac{\partial^2\mathrm{h}_l(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}^T\partial\boldsymbol{\eta}} = \left(\, a_{ij}^{(l)} \,\right) i,j = 1,\ldots,k$$

where $a_{ji}^{(l)} = \dfrac{\partial^2 \mathrm{h}_l(\boldsymbol{\eta})}{\partial \eta_j \partial \eta_i}$.

We introduce the following quantities which will be used in evaluating the bias of $\widehat{\boldsymbol{\theta}}_n$ from (2.2.29).

$$a_{rij}^{(l)} = \Big(\frac{\partial a_{ij}^{(l)}}{\partial \boldsymbol{\eta}} \Big)_r = \frac{\partial^3 \mathrm{h}_l(\boldsymbol{\eta})}{\partial \eta_r \partial \eta_j \partial \eta_i} \ ,$$

$$a_{srij}^{(l)} = \Big(\frac{\partial a_{rij}^{(l)}}{\partial \boldsymbol{\eta}^T} \Big)_{\mathrm{s}} = \frac{\partial^4 \mathrm{h}_l(\boldsymbol{\eta})}{\partial \eta_s \partial \eta_r \partial \eta_j \partial \eta_i} \ ,$$

and

$$c_r = \bar{\mathrm{T}}_{nr} - \eta_r.$$

Then the second term on the RHS of (2.2.30), which is a scalar is written as

$$\mathrm{E}[(\underset{\sim}{\bar{\mathrm{T}}} n - \boldsymbol{\eta})^T \frac{\partial^2 \mathrm{h}_l(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T \partial \boldsymbol{\eta}} (\underset{\sim}{\bar{\mathrm{T}}} n - \boldsymbol{\eta})]$$

$$= \mathrm{E}[\mathrm{Tr}(\underset{\sim}{\mathrm{A}}\,\underset{\sim}{\mathrm{B}})] = \mathrm{E}[\sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij}^{(l)} b_{ji}]$$

$$= \frac{1}{n} [\sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij}^{(l)} \sigma_{ji}]. \qquad (2.2.31)$$

The last equality is a consequence of (2.2.25), noting that the $a_{ij}^{(l)}$'s are fixed.

The subsequent higher order terms on the RHS of (2.2.30) are obtained by differentiating the second term with respect to $\boldsymbol{\eta}$ and taking the expectation as follows. The expected value of the third term in (2.2.30) becomes

$$\mathrm{E}[(\underset{\sim}{\bar{\mathrm{T}}} n - \boldsymbol{\eta})(\underset{\sim}{\bar{\mathrm{T}}} n - \boldsymbol{\eta})^T \frac{\partial^3 \mathrm{h}_l(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T \partial \boldsymbol{\eta}} (\underset{\sim}{\bar{\mathrm{T}}} n - \boldsymbol{\eta})]$$

$$= \mathrm{E}[\sum_{r=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{k} c_r a_{rij}^{(l)} b_{ji}]$$

$$= \frac{1}{n^2}[\sum_{r=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{k} a_{rij}^{(l)}\gamma_{rji}], \tag{2.2.32}$$

where the first equal sign in (2.2..32) is a result of taking the derivative of $a_{ij}^{(l)}$ in (2.2.31), and γ_{rji} are given by (2.2.26).

From (2.2.26) we have $\mathrm{E}[c_r a_{rij}^{(l)} b_{ji}] = \mathrm{O}(n^{-2})$ and since k is fixed (2.2.32) is $\mathrm{O}(n^{-2})$.

The expected value of the fourth term in (2.2.30) is given as,

$$\mathrm{E}[(\bar{\mathbb{T}}n-\boldsymbol{\eta})(\bar{\mathbb{T}}n-\boldsymbol{\eta})^T \frac{\partial^4 \mathrm{h}_l(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}^T\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T\partial\boldsymbol{\eta}} (\bar{\mathbb{T}}n-\boldsymbol{\eta})(\bar{\mathbb{T}}n-\boldsymbol{\eta})^T]$$

$$= \mathrm{E}[\sum_{s=1}^{k}\sum_{r=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{k} c_s c_r a_{srij}^{(l)} b_{ji}]$$

$$= \frac{1}{n^2}[\sum_{s=1}^{k}\sum_{r=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{k} a_{srij}^{(l)}\lambda_{srji}]. \tag{2.2.33}$$

From (2.2.27), $\mathrm{E}[c_s c_r a_{srij}^{(l)} b_{ji}] = \mathrm{O}(n^{-2})$ and similarly (2.2.33) is $\mathrm{O}(n^{-2})$.

Note that the term with the fifth and sixth partial derivatives of $\mathrm{h}_l(\boldsymbol{\eta})$ will contribute $\mathrm{O}(n^{-3})$ to (2.2.30). So that from (2.2.30), as $n \longrightarrow \infty$, we obtain

$$\mathrm{E}[\mathrm{h}_l(\bar{\mathbb{T}}n) - \mathrm{h}_l(\boldsymbol{\eta})] = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}^{(l)}\sigma_{ji}]}{n} + \frac{\sum_{r=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{k} a_{rij}^{(l)}\gamma_{rji} + \sum_{s=1}^{k}\sum_{r=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{k} a_{srij}^{(l)}\lambda_{srji}}{n^2}$$

$$+ \mathrm{O}(n^{-3}), \quad l = 1,\ldots,k \tag{2.2.34}$$

and hence taking expectation of (2.2.29) we have

$$\mathrm{E}[\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}] = \frac{1}{n}\,\boldsymbol{a}(\boldsymbol{\theta}) + \frac{1}{n^2}\,\boldsymbol{b}(\boldsymbol{\theta}) + \mathrm{o}(\frac{1}{n^2}), \qquad \text{as } n \longrightarrow \infty, \tag{2.2.35}$$

where

$$\boldsymbol{a}(\boldsymbol{\theta}) = \Big(\frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij}^{(l)} \sigma_{ji} \Big) \qquad l = 1, \ldots, k$$

and

$$\boldsymbol{b}(\boldsymbol{\theta}) = \Big(\frac{1}{3!} \sum_{r=1}^{k} \sum_{i=1}^{k} \sum_{j=1}^{k} a_{rij}^{(l)} \gamma_{rji} + \frac{1}{4!} \sum_{s=1}^{k} \sum_{r=1}^{k} \sum_{i=1}^{k} \sum_{j=1}^{k} a_{srij}^{(l)} \lambda_{srji} \Big), \quad l = 1, \ldots, k$$

are functions of $\boldsymbol{\theta}$ independent of n.

Note that $\widehat{\boldsymbol{\theta}}_n$ will be unbiased for $\boldsymbol{\theta}$ if the k^{th} order differential of $\mathbf{H}(\boldsymbol{\eta}) \equiv \mathbf{0}$ for $k > 1$.

We show, in section 2.3.1, that the jackknifing method reduces the bias in the MLE by eliminating the leading term of order n^{-1} from the bias of the MLE (2.2.30). In case the leading term in the bias of $\widehat{\boldsymbol{\theta}}_n$ is zero, the classical delete-one jackknife would not achieve the desired advantage of bias reduction in the estimator $\widehat{\boldsymbol{\theta}}_n$.

2.3 Jackknife Estimator of the MLE

Next we investigate the properties of the delete one observation in the jackknife estimation. The jackknife estimation was originally introduced by

Quenoulli as a method of bias reduction and was later extended by Tukey for robust interval estimation. Considerable research work has since been made to study the properties of the jackknife estimators in a variety of statistical estimation problems. Detailed studies of the technique appear in the literature of, among others, Schucany, Gray and Owen, 1971; Gray, Watkins and Adams, 1972; Miller, 1974; Sen, 1977.

The purpose of this chapter is to investigate the form of bias reduction achieved by the jackknife method applied to the maximum likelihood estimators of the parameters of an exponential family of distributions and furthermore, study the distribution of the jackknife estimators of the parameters in large samples. The results of this chapter are then applied to the parameter estimators of the length-biased version of the original exponential family of distributions, in the next chapter.

It should be noted that the jackknife method studied here is different from the one considered by Brillinger, 1964. Brillinger defines jackknife by dividing the sample observations into a fixed number of groups, and successively deleting the groups to obtain the pseudo-estimates. The asymptotic results are then obtained by increasing the sample size so that the number of observations in each group increases while the number of groups remains fixed. In our classical delete-one jackknife the large sample properties of the estimates the number of observations per group tends to infinity as the sample size increases without bound.

Given the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$, of $\boldsymbol{\theta}$ we define the pseudo-values

$$\widetilde{\boldsymbol{\theta}}_i = n\widehat{\boldsymbol{\theta}} - (n-1)\widehat{\boldsymbol{\theta}}^{(i)} \qquad i = 1, \ldots, n \tag{2.3.1}$$

where $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^{(i)}$ are the MLE computed from all the n and $n-1$ data points respectively, the superscript (i) denotes the i^{th} observation deleted from the data vector.

The classical Jackknife estimator of $\boldsymbol{\theta}$ defined as

$$\widetilde{\boldsymbol{\theta}}^{J} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\boldsymbol{\theta}}_i$$

$$= \widehat{\boldsymbol{\theta}} - \frac{(n-1)}{n} \sum_{i=1}^{n} (\widehat{\boldsymbol{\theta}}^{(i)} - \widehat{\boldsymbol{\theta}}\,). \tag{2.3.2}$$

The Jackknife uses the estimates $\widehat{\boldsymbol{\theta}}^{(i)}$ $(i = 1,\ldots,n)$ which are obtained by successively deleting one observation from the data and estimating the MLE from the remaining observations. So we need to study the behavior of the terms $(\widehat{\boldsymbol{\theta}}^{(i)} - \widehat{\boldsymbol{\theta}}_n)$, in order to establish the asymptotic results of the Jackknife estimators. The large sample properties of the Jackknife estimators are established through the use of estimating equations for the MLE.

Let $U^{(i)}_{n-1}(\boldsymbol{\theta})$ be the solution of the estimating equation (2.2.3) with the i^{th} observation deleted $(i = 1,\ldots,n)$. Since $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ we have $U_n(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$ and similarly $U^{(i)}_{n-1}(\widehat{\boldsymbol{\theta}}^{(i)}) = \mathbf{0}$.

By definition

$$U^{(i)}_{n-1}(\boldsymbol{\theta}) = \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_j;\, \boldsymbol{\theta})$$

$$= U_n(\boldsymbol{\theta}) - \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_i;\, \boldsymbol{\theta}) \tag{2.3.3}$$

so that

$$U^{(i)}_{n-1}(\hat{\theta}) = -\frac{\partial}{\partial\theta}\log f(X_i;\,\theta)\,|_{\hat{\theta}}$$

$$= -\underset{\sim}{T}(X_i) - \psi(\hat{\theta}). \tag{2.3.4}$$

From (2.3.4) we write

$$U^{(i)}_{n-1}(\hat{\theta}) - U^{(i)}_{n-1}(\hat{\theta}^{(i)}) = -\frac{\partial}{\partial\theta}\log f(x_i;\,\theta)\,|_{\hat{\theta}} + \frac{\partial}{\partial\theta}\log f(X_i;\,\theta)\,|_{\hat{\theta}^{(i)}}$$

$$= \psi(\hat{\theta}^{(i)}) - \psi(\hat{\theta}) \tag{2.3.5}$$

and similarly using (2.2.6)

$$U_n(\hat{\theta}) - U_n(\hat{\theta}^{(i)}) = n[\,\psi(\hat{\theta}) - \psi(\hat{\theta}^{(i)})\,]. \tag{2.3.6}$$

Since $U_n(\hat{\theta}) = 0$, we have

$$\underset{\sim}{T}(X_i) + \psi(\hat{\theta}^{(i)}) = n[\psi(\hat{\theta}^{(i)}) - \psi(\hat{\theta})\,] \tag{2.3.7}$$

which can be written as

$$\underset{\sim}{T}(X_i) + \psi(\hat{\theta}) = (n-1)[\,\psi(\hat{\theta}^{(i)}) - \psi(\hat{\theta})\,]. \tag{2.3.8}$$

Note that $\underset{\sim}{T}(X_i)$ $i = 1, \ldots, n$ are iid with finite second moment, as a result

$$\max_{1 \le k \le n} \|\,\underset{\sim}{T}(X_i)\| = o_p(\sqrt{n}), \tag{2.3.9}$$

hence

$$\max_{1 \le i \le n} \| \psi(\hat{\theta}^{(i)}) - \psi(\hat{\theta}) \| = o_p(\tfrac{1}{\sqrt{n}}). \tag{2.3.10}$$

Given the strong consistency of $\hat{\theta}$ and the continuity assumptions on $\psi(\theta)$ in (2.2.13), we conclude from (2.3.10) that

$$\max_{1 \le i \le n} \| \hat{\theta}^{(i)} - \hat{\theta} \| = o_p(\tfrac{1}{\sqrt{n}}). \tag{2.3.11}$$

The following lemma will be used in the study of the asymptotic properties of the Jackknife estimate.

Lemma 2.1. Let $\underset{\sim}{h}_{n_i} = h(\hat{\theta}^{(i)} - \hat{\theta})$, where $0 < h < 1$.

Then $\max_{1 \le i \le n} \| \varphi(\hat{\theta} + \underset{\sim}{h}_{n_i}) - \varphi(\hat{\theta}) \| = o_p(\tfrac{1}{\sqrt{n}})$, uniformly in i.

Proof.

By (2.2.11) $\max_{1 \le i \le n} \| \hat{\theta}^{(i)} - \hat{\theta} \| = o_p(\tfrac{1}{\sqrt{n}})$, so that uniformly in i $(1 \le i \le n)$

$$\varphi(\hat{\theta} + \underset{\sim}{h}_{n_i}) = \varphi(\hat{\theta} + o_p(\tfrac{1}{\sqrt{n}})) = \varphi(\hat{\theta}) + o_p(\tfrac{1}{\sqrt{n}}). \tag{2.3.12}$$

The last equality is due to the assumed continuity of $\varphi(\theta)$ around θ and almost sure convergence of $\hat{\theta}$ to θ.

□

Lemma 2.2. Let $\hat{\theta}$ and $\hat{\theta}^{(i)}$ be the MLE of θ obtained from a complete sample and a delete one sample respectively. Then

$$\sum_{i=1}^{n}(\widehat{\boldsymbol{\theta}}^{(i)}-\widehat{\boldsymbol{\theta}}\ )\ =\ o_p(\tfrac{1}{n}). \tag{2.3.13}$$

Proof.

Consider Taylor's expansion of $\boldsymbol{\psi}(\widehat{\boldsymbol{\theta}}^{(i)})$ around $\widehat{\boldsymbol{\theta}}$ and using (2.3.8) we have

$$\tfrac{1}{n}\sum_{i=1}^{n}[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i)+\boldsymbol{\psi}(\widehat{\boldsymbol{\theta}})][\underset{\sim}{\mathrm{T}}(\mathrm{X}_i)+\boldsymbol{\psi}(\widehat{\boldsymbol{\theta}})]^T$$

$$=\ \frac{(n-1)^2}{n}\ \sum_{i=1}^{n}[\ \boldsymbol{\psi}(\widehat{\boldsymbol{\theta}}^{(i)})-\boldsymbol{\psi}(\widehat{\boldsymbol{\theta}})][\ \boldsymbol{\psi}(\widehat{\boldsymbol{\theta}}^{(i)})-\boldsymbol{\psi}(\widehat{\boldsymbol{\theta}})]^T$$

$$=\frac{n-1)^2}{n}\ \boldsymbol{\varphi}(\widehat{\boldsymbol{\theta}}^{*})\sum_{i=1}^{n}(\widehat{\boldsymbol{\theta}}^{(i)}-\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}^{(i)}-\widehat{\boldsymbol{\theta}})^T\boldsymbol{\varphi}(\widehat{\boldsymbol{\theta}}). \tag{2.3.14}$$

Since the LHS of (2.3.14) is $\mathrm{O}_p(1)$ and the continuity assumption on $\boldsymbol{\varphi}(\boldsymbol{\theta})$, (2.3.14) implies that,

$$\sum_{i=1}^{n}(\widehat{\boldsymbol{\theta}}^{(i)}-\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}^{(i)}-\widehat{\boldsymbol{\theta}})^T\ =\mathrm{O}_p(\tfrac{1}{n})\ \text{ and hence}$$

$$\sum_{i=1}^{n}||\ \widehat{\boldsymbol{\theta}}^{(i)}-\widehat{\boldsymbol{\theta}}\ ||=\mathrm{O}_p(\tfrac{1}{\sqrt{n}}). \tag{2.3.15}$$

Also by Taylor's expansion of $\boldsymbol{\psi}(\widehat{\boldsymbol{\theta}}^{(i)})$ around $\widehat{\boldsymbol{\theta}}$ we write

$$\mathbf{0}=\ \sum_{i=1}^{n}[\ \boldsymbol{\psi}(\widehat{\boldsymbol{\theta}}^{(i)})-\boldsymbol{\psi}(\widehat{\boldsymbol{\theta}})]\ =\sum_{i=1}^{n}(\widehat{\boldsymbol{\theta}}^{(i)}-\ \widehat{\boldsymbol{\theta}})^T\ \boldsymbol{\varphi}(\widehat{\boldsymbol{\theta}}+\underset{\sim}{\mathrm{h}}_{n_i})$$

$$=\ \sum_{i=1}^{n}(\widehat{\boldsymbol{\theta}}^{(i)}-\widehat{\boldsymbol{\theta}})^T\ \boldsymbol{\varphi}(\widehat{\boldsymbol{\theta}})+\sum_{i=1}^{n}(\widehat{\boldsymbol{\theta}}^{(i)}-\widehat{\boldsymbol{\theta}})^T[\ \boldsymbol{\varphi}(\widehat{\boldsymbol{\theta}}+\underset{\sim}{\mathrm{h}}_{n_i})-\boldsymbol{\varphi}(\widehat{\boldsymbol{\theta}})]$$

so that,

$$\| \sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}_{-i} - \hat{\boldsymbol{\theta}})^T \varphi(\hat{\boldsymbol{\theta}}) \| \leq \sum_{i=1}^{n} \| \hat{\boldsymbol{\theta}}_{-i} - \hat{\boldsymbol{\theta}})^T \| \max_i \| \varphi(\hat{\boldsymbol{\theta}} + \underset{\sim}{h}\, n_i) - \varphi(\hat{\boldsymbol{\theta}}) \|$$

$$= O_p(\tfrac{1}{\sqrt{n}})\, o_p(\tfrac{1}{\sqrt{n}}) \quad = o_p(\tfrac{1}{n}), \tag{2.3.16}$$

which implies

$$\sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}^{(i)} - \hat{\boldsymbol{\theta}}) = o_p(\tfrac{1}{n}).$$

∎

2.3.1 Asymptotic Distribution and Bias of the Jackknife Esimator

Theorem 2.1 establishes that the Jackknife estimate $\hat{\boldsymbol{\theta}}^J$ and the MLE $\hat{\boldsymbol{\theta}}_n$ are equivalent in the sense that they both have the same asymptotic normal distribution.

Theorem 2.1. For $\tilde{\boldsymbol{\theta}}^J$ defined in (2.2.1)

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}^J - \boldsymbol{\theta}) \longrightarrow_d N(0, I^{-1}(\boldsymbol{\theta})) \quad \text{as } n \longrightarrow \infty \tag{2.3.17}$$

Proof.

From (2.3.1)

$$\tilde{\boldsymbol{\theta}}^J - \hat{\boldsymbol{\theta}} = -\frac{(n-1)}{n} \sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}^{(i)} - \hat{\boldsymbol{\theta}}) \tag{2.3.18}$$

Since the MLE $\hat{\boldsymbol{\theta}}$ is asymptotically normal, $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \longrightarrow_d \mathrm{N}(0, \mathrm{I}^{-1}(\boldsymbol{\theta}))$, and by

Lemma 2.2, the RHS of (2.3.18) is $o_p(\frac{1}{n})$, which implies $\sqrt{n}\left(\tilde{\boldsymbol{\theta}}^J - \hat{\boldsymbol{\theta}}\right) = o_p(\frac{1}{\sqrt{n}})$, and

using Slutskys Theorem we conclude $\sqrt{n}(\tilde{\boldsymbol{\theta}}^J - \boldsymbol{\theta}) \longrightarrow_d \mathrm{N}(0, \mathrm{I}^{-1}(\boldsymbol{\theta}))$.

□

Since the Jackknife estimator $\tilde{\boldsymbol{\theta}}^J$ is a linear combination of the MLE $\hat{\boldsymbol{\theta}}_n$ from (2.3.1) the bias of the Jackknife estimator is expressed as:

$$\mathrm{E}[\,\tilde{\boldsymbol{\theta}}^J - \boldsymbol{\theta}\,] = \mathrm{E}[\,\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\,] - \frac{(n-1)}{n}\sum_{i=1}^{n}\mathrm{E}[(\hat{\boldsymbol{\theta}}^{(i)} - \hat{\boldsymbol{\theta}}_n)]$$

$$= n\mathrm{E}[\,\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\,] - (n-1)\mathrm{E}[(\hat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta})] \qquad (2.3.19)$$

From (2.2.26) $n\mathrm{E}[\,\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\,] = a(\boldsymbol{\theta}) + \frac{1}{n}\mathbf{b}(\boldsymbol{\theta}) + \mathrm{O}(\frac{1}{n^2})$ so that (2.3.17) becomes

$$\mathrm{E}[\,\tilde{\boldsymbol{\theta}}^J - \boldsymbol{\theta}\,] = -\frac{1}{n(n-1)}\boldsymbol{b}(\boldsymbol{\theta}) + o(n^{-2}) \qquad (2.3.20)$$

Comparing the bias of the MLE $\hat{\boldsymbol{\theta}}_n$ in (2.2.24) with the bias of the Jackknife estimator $\tilde{\boldsymbol{\theta}}^J$ in (2.3.20) we note that the Jackknife estimate has reduced the bias term of order $\frac{1}{n}$. This is exactly the purpose for which the Jackknife was proposed by Quenouille. We also noted that if $a(\boldsymbol{\theta})$ vanishes to zero then the Jackknife estimator will not reduce the bias.

Gray, Watkins and Adams, 1972; Schucany, Gray and Owen, 1971 proposed

higher order Jackknife where, instead of delete-one, k observations are deleted successively and the statistic is recomputed from the remaining observations. For example if instead of deleting one observation, two observations are deleted, then the additional bias term of order n^{-2} will be eliminated.

Theorem 2.1 establishes that the variance of the Jackknife estimator $\tilde{\boldsymbol{\theta}}^J$ is a consistently estimates the variance of $\hat{\boldsymbol{\theta}}_n$. Tukey proposed the variance of the sample variance of the pseudo values (2.2.1) divided by n would consistently estimate the variance of $\tilde{\boldsymbol{\theta}}^J$. This is shown in Theorem 2.2.

Theorem 2.2. Consider the Jackknifed dispersion matrix

$$S_{nJ} = \frac{1}{n-1} \sum_{i=1}^{n} (\tilde{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}}^J)(\tilde{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}}^J)^T \tag{2.3.21}$$

Then $S_{nJ} \longrightarrow \mathrm{I}^{-1}(\boldsymbol{\theta})$, as $n \longrightarrow \infty$

where $\mathrm{I}(\boldsymbol{\theta})$ is the Fisher information matrix defined by A2.

Proof.

Consider the variates $(\tilde{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}}^J)$, and using (2.3.10) and (2.3.11) substitute the expression (2.3.10) for $\tilde{\boldsymbol{\theta}}_i$, which gives

$$(\tilde{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}}^J) = (n-1)\left\{ (\hat{\boldsymbol{\theta}}^{(i)} - \hat{\boldsymbol{\theta}}) - \frac{1}{n} \sum_{j=1}^{n} (\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}) \right\}, \tag{2.3.22}$$

so that

$$S_{Jn} = \frac{1}{n-1} \sum_{i=1}^{n} (\tilde{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}}^J)\,(\tilde{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}}^J)^T$$

$$= (n-1) \sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}^{(i)} - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{(i)} - \hat{\boldsymbol{\theta}})^T - n(n-1) \left\{ \frac{1}{n} \sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}^{(i)} - \hat{\boldsymbol{\theta}}) \right\} \left\{ \frac{1}{n} \sum_{j=1}^{n} (\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}) \right\}^T$$

(2.3.23)

From Lemma 2.2, $\sum_{i=1}^{n} (\widehat{\theta}^{(i)} - \widehat{\theta}) = o_p(\frac{1}{\sqrt{n}})$, so that n times the second term on the

RHS of (2.3.23) is $o_p(1)$.

To study the asymptotic behavior of the first term on RHS of (2.3.22) we use Taylor's expansion of $\psi(\widehat{\theta}^{(i)})$ around $\widehat{\theta}$, together with Lemma 2.1 and write

$$\psi(\widehat{\theta}^{(i)}) - \psi(\widehat{\theta}) = \varphi(\widehat{\theta})\,(\widehat{\theta}^{(i)} - \widehat{\theta}) + o_p(n^{-1/2}) \tag{2.3.24}$$

so that

$$(n-1)\sum_{i=1}^{n} [\psi(\widehat{\theta}^{(i)}) - \psi(\widehat{\theta})][\psi(\widehat{\theta}^{(i)}) - \psi(\widehat{\theta})]^T$$

$$= (n-1)\Big\{ \varphi(\widehat{\theta}) \Big(\sum_{i=1}^{n} (\widehat{\theta}^{(i)} - \widehat{\theta})(\widehat{\theta}^{(i)} - \widehat{\theta})^T \Big) \varphi(\widehat{\theta}) + o_p(n^{-1}) \Big\}$$

$$= \varphi(\widehat{\theta})\, S_{nJ}\, \varphi(\widehat{\theta}) + o_p(1). \tag{2.3.25}$$

Now $\varphi(\widehat{\theta}) \longrightarrow \varphi(\theta)$ a.s. and from (2.3.6) the LHS of (2.3.25) is given by

$$\frac{1}{n-1} \sum_{i=1}^{n} [\underset{\sim}{T}(X_i) + \psi(\theta)]\,[\underset{\sim}{T}(X_i) + \psi(\theta)]^T. \tag{2.3.26}$$

Since by Khintchine (WLLN) the term in (2.3.26) converges to $I(\theta) = -\varphi(\theta)$, then,

$S_{nJ} \longrightarrow -[\varphi(\theta)]^{-1}\varphi(\theta)[\varphi(\theta)]^{-1} = I^{-1}(\theta)$, as $n \longrightarrow \infty$.

So that the sample variance of the pseudo values is a consistent estimator of the Jackknife variance estimator.

Chapter 3.

Parametric Regression Model: Exponential Family

3.1 Introduction.

This chapter provides a study of the Jackknife estimator of the regression parameters in a multiple regression setup. The bias of the Jackknife estimator of the regression parameters is derived and compared with the bias of the corresponding MLE. It is shown the Jackknife estimator and the MLE of the regression coefficients have the same asymptotic normal distribution. The convergence of the Jackknife variance estimate to the variance of the MLE is also studied.

We consider the statistical estimation in a linear regression model, where the mean of the response variate is a linear function of some fixed and known set of covariates. The notations and assumptions of the model are given below:

Let $\mathbf{X} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n)^T$ denote an $n \times 1$ vector of response r.v.'s

$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^T$ be the corresponding vector of expected value parameters.

$\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{ik})^T$ a vector of nonstochastic covariates,

$\underset{\sim}{\mathbf{Z}} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)^T$ denote a $n \times k$ design matrix with full rank $k < n$

$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)^T$ is $k \times 1$ vector of unknown parameters to be estimated.

Assume $(\mathrm{X}_i, \mathbf{z}_i)$, $i = 1, \ldots, n$, are the observations on the i^{th} subject. In some

applications it may be of interest to express the mean of X_i or some function of X_i, say $T(X_i)$, as

$$E[T(X_i)] = \mu_i = \mu(z_i^T\beta,\phi)\,, \qquad i = 1,\ldots,n \tag{3.1}$$

The usual method of estimation in the univariate linear models is the least squares method. Least squares estimators are generally easier to handle mathematically and computationally, however they become less efficient when the underlying distribution of the error terms is significantly different from normal. Other types of robust estimators for linear models appearing in the literature are: 1) the maximum likelihood type of estimators (M-estimators) and 2) estimators derived from ranks (R-estimators). M-estimators introduced by Huber, 1965 is defined as an estimator T_n for β to be the solution to the equations

$$\sum_{i=1}^{n}\rho(X_i - z_i^T T_n) = \min_{\theta\in\Theta}\sum_{i=1}^{n}\rho(x_i - z_i^T T_n)$$

where Θ is in $\mathbb{R}^k$. Huber, 1964 advocated replacing $\rho(y)$ by $\psi(y) = \partial\rho/\partial y$. Particular forms of $\rho(y)$ are

(*i*) $\rho(y) = y^2$

(*ii*) $\rho(y) = |y|$

(*iii*) $\rho(y) = -\log f(y)$, where $f(y)$ is the density of Y.

Our discussion will be focused on the general linear model with the score function of $\rho(y) = -\log(f(y))$. Specifically the X_i $i = 1,2,\ldots,n$ are assumed to follow an exponential family of distributions given by

$$f(X_i;\beta,\phi) = c(X_i,\phi)\exp\Big\{\, b(z_i^T\beta,\phi)^T \underset{\sim}{T}(X_i) + \Psi(b(z_i^T\beta,\phi))\,\Big\}, \tag{3.2}$$

where

$b(z_i^T\beta,\phi) = \Big(\, b_l(z_i^T\beta,\phi) \quad l = 1,2,\ldots,q\,\Big)^T$ is a q-vector of independent quantities,

and ϕ is the scale parameter.

$\Psi(.)$ and $\mathbf{b}(.)$ are known function and

$\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) = (\ \mathrm{T}_1(\mathrm{X}_i), \ldots, \mathrm{T}_q(\mathrm{X}_i)\ )^T$ is a q-vector of sufficient statistics.

The specification of (3.2) includes a variety of regression models such as the log-linear models, survival data models and linear models, (see McCullagh and Nelder, (1989). Statistical analysis of linear regression models often involves estimation of the parameters $\boldsymbol{\beta}$ and making statistical inference about the parameter. We focus on estimating the parameters $\boldsymbol{\beta}$ via Jackknife estimation of the MLE $\widehat{\boldsymbol{\beta}}$. Both the bias and asymptotic distribution of $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\beta}}^J$ (the Jackknife estimator) are studied.

To derive the maximum likelihood estimators of $\boldsymbol{\beta}$ and ϕ, we impose the same set of assumptions, "A" in Chapter 2, on the density $\mathrm{f}(\mathrm{X};\boldsymbol{\beta},\phi)$ in (3.2) with $(\boldsymbol{\beta},\phi)$ replacing $\boldsymbol{\theta}$, in order to derive asymptotic properties of the MLE $\widehat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$. For simplicity of notation we write $\boldsymbol{\beta}^0$ for $(\boldsymbol{\beta},\phi)$ and $\mathbf{b}(\mathbf{z}_i;\boldsymbol{\beta}^0)$ for $\mathbf{b}(\mathbf{z}_i^T\boldsymbol{\beta},\phi)$.

Assumptions

(i). for $l, j = 1, \ldots, (k+1)$

$\frac{\partial}{\partial\beta_l^0} \log \mathrm{f}(\mathrm{X}_i;\boldsymbol{\beta}^0)$ and $\frac{\partial^2}{\partial\beta_l^0\partial\beta_j^0} \log \mathrm{f}(\mathrm{X}_i;\boldsymbol{\beta}^0)$ exist and are bounded by integrable

functions,

i.e. $|\frac{\partial}{\partial\beta_l^0} \log \mathrm{f}(\mathrm{X}_i;\boldsymbol{\beta}^0)| \leq \mathrm{K}_1(\mathrm{x})\mathrm{dx}$ and $|\frac{\partial^2}{\partial\beta_l^0\partial\beta_j^0} \log \mathrm{f}(\mathrm{X}_i;\boldsymbol{\beta}^0)| \leq \mathrm{K}_2(\mathrm{x})$

where $\int \mathrm{K}_i(\mathrm{x})\mathrm{dx} < \infty \quad i = 1,2$

(ii). The Fisher Information

$\mathbf{I}_i(\boldsymbol{\beta}^0) = \mathrm{E}_{\boldsymbol{\beta}^0}[\frac{\partial}{\partial \boldsymbol{\beta}^0} \log f(X_i;\boldsymbol{\beta}^0)][\frac{\partial}{\partial \boldsymbol{\beta}^0} \log f(X_i;\boldsymbol{\beta}^0)]^T$ exists and is positive definite.

(*iii*). $-\mathrm{E}[\frac{\partial^2}{\partial \boldsymbol{\beta}^{0T} \partial \boldsymbol{\beta}^0} \log f(X_i;\boldsymbol{\beta}^0)] = \mathbf{I}_i(\boldsymbol{\beta}^0)$ is a positive definite matrix.

In addition we assume

4. $\mathrm{E}\,[||\underset{\sim}{\mathrm{T}}(X_i)||^2 < \infty$, for all $i = 1,\ldots,n$ and all $\boldsymbol{\beta}^0 \in \Theta$

5. $\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0)$ is a q-vector of linearly independent components that are continuously differentiable, with continuous partial derivatives of $\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0)$ in some neighborhood of $\boldsymbol{\beta}^0$. Also assume

$||\frac{\partial}{\partial \boldsymbol{\beta},\phi} \mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0)||$ is nonsingular

6. $\Psi(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))$ is continuously differentiable with respect to $\boldsymbol{\beta}^0$ and all order derivatives of $\Psi(.)$ are continuous in the neighborhood of $\boldsymbol{\beta}^0$.

Note that the moments of $\underset{\sim}{\mathrm{T}}(X_i)$ can be obtained from (3.2), we derive the first moment and the second moment of $\underset{\sim}{\mathrm{T}}(X_i)$ by differentiating $\int f(X_i;\boldsymbol{\beta}^0)dx$ with respect to $\boldsymbol{\beta}^0$.

$$\mathbf{0} = \int \frac{\partial}{\partial \boldsymbol{\beta}} f(X_i;\boldsymbol{\beta}^0)dx = \int (\mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0))^T\,[\underset{\sim}{\mathrm{T}}(X_i) + \boldsymbol{\psi}(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))]\, f(X_i;\boldsymbol{\beta}^0)dx$$

where

$$\mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0))^T = \frac{\partial\, \mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))^T}{\partial \boldsymbol{\beta}^0} \qquad \text{is a } (k+1)\times q \text{ vector,}$$

$$\boldsymbol{\psi}(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0)) = \frac{\partial\, \Psi(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))}{\partial \mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0)} \qquad \text{is a } (q\times 1) \text{ vector}$$

so that,

$$E[\underset{\sim}{T}(X_i)] = -\psi(b(z_i,\beta^0)) \qquad i=1,\ldots,n. \tag{3.3}$$

By differentiating $\int \frac{\partial}{\partial \beta^0} f(X_i;\beta^0)dx$ again with respect to β^{0^T} we obtain,

$$\begin{aligned}
0 &= \int \frac{\partial^2}{\partial\beta^0\partial\beta^{0^T}} f(X_i;\beta^0)dx \\
&= \int \Big\{(b'(z_i,\beta^0))^T\,[\underset{\sim}{T}(X_i)+\psi(b(z_i,\beta^0))][\underset{\sim}{T}(X_i)+\psi(b(z_i,\beta^0))]^T b'(z_i,\beta^0)\, f(X_i;\beta^0)dx \\
&\quad + \int [(\, b''(z_i,\beta^0))^T\,[\underset{\sim}{T}(X_i)+\psi(b(z_i,\beta^0))]\, f(X_i;\beta^0)dx \\
&\quad + (b'(z_i,\beta^0))^T\,\varphi(b(z_i,\beta^0))\, b'(z_i,\beta^0),
\end{aligned} \tag{3.4}$$

where

$$\varphi(b(z_i,\beta^0)) = \frac{\partial^2\,\Psi(b(z_i,\beta^0))}{\partial(b(z_i,\beta^0))^T\partial b(z_i,\beta^0)}.$$

It follows from (3.4) and (3.3) that

$$\begin{aligned}
&E[(b'(z_i,\beta^0))^T\,[\underset{\sim}{T}(X_i)+\psi(b(z_i,\beta^0))][\underset{\sim}{T}(X_i)+\psi(b(z_i,\beta^0))]^T\, b'(z_i,\beta^0) \\
&\qquad = -(b'(z_i,\beta^0))^T\,\varphi(b(z_i,\beta^0))\, b'(z_i,\beta^0),
\end{aligned} \tag{3.5}$$

hence the covariance matrix of $\underset{\sim}{T}(X_i)$ is given by the $(q\times q)$ matrix

$$E[\underset{\sim}{T}(X_i)-E\underset{\sim}{T}(X_i)]\,[\underset{\sim}{T}(X_i)-E\underset{\sim}{T}(X_i)]^T$$

$$= - \varphi(\mathbf{b}(\mathbf{z}_i,\beta^0)). \tag{3.6}$$

Higher order central moments of $\underset{\sim}{\mathrm{T}}(\mathrm{X}_i)$ can also be obtained from the (3.2). The third and fourth central moments of $\underset{\sim}{\mathrm{T}}(\mathrm{X}_i)$ are given by

$$\mathrm{E}[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) - \mathrm{E}\underset{\sim}{\mathrm{T}}(\mathrm{X}_i)]^T[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) - \mathrm{E}\underset{\sim}{\mathrm{T}}(\mathrm{X}_i)]\ [\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) - \mathrm{E}\underset{\sim}{\mathrm{T}}(\mathrm{X}_i)]^T$$

$$= -\Gamma(\mathbf{b}(\mathbf{z}_i,\bar{\beta})) \tag{3.7}$$

and

$$\mathrm{E}[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) - \mathrm{E}\underset{\sim}{\mathrm{T}}(\mathrm{X}_i)]^T[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) - \mathrm{E}\underset{\sim}{\mathrm{T}}(\mathrm{X}_i)]\ [\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) - \mathrm{E}\underset{\sim}{\mathrm{T}}(\mathrm{X}_i)]^T[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) - \mathrm{E}\underset{\sim}{\mathrm{T}}(\mathrm{X}_i)]$$

$$= 3\ \varphi(\mathbf{b}(\mathbf{z}_i,\beta^0))\ \varphi(\mathbf{b}(\mathbf{z}_i,\beta^0))^T - \Upsilon(\mathbf{b}(\mathbf{z}_i,\beta^0)), \tag{3.8}$$

where

$$\Gamma(\mathbf{b}(\mathbf{z}_i,\beta^0)) = \frac{\partial^3\ \Psi(\mathbf{b}(\mathbf{z}_i,\beta^0))}{\partial(\mathbf{b}(\mathbf{z}_i,\beta^0))^T\partial\mathbf{b}(\mathbf{z}_i,\beta^0)\partial(\mathbf{b}(\mathbf{z}_i,\beta^0))^T} \quad \text{and}$$

$$\Upsilon(\mathbf{b}(\mathbf{z}_i,\beta^0)) = \frac{\partial^4\ \Psi(\mathbf{b}(\mathbf{z}_i,\beta^0))}{\partial(\mathbf{b}(\mathbf{z}_i,\beta^0))^T\partial\mathbf{b}(\mathbf{z}_i,\beta^0)\partial(\mathbf{b}(\mathbf{z}_i,\beta^0))^T\partial\mathbf{b}(\mathbf{z}_i,\beta^0)}.$$

3.2 Estimation of Parameters of the Regression Model

The maximum likelihood estimator of β^0, denoted by β_n^0, is a function of the

sample data, of size n, which maximizes the log-likelihood function. To study the maximum likelihood estimators we present the partial derivatives of the log-likelihood with respect to the parameters β^0.

The first and second partial derivatives of the log $f(X_i;\beta^0)$ in (3.2) with respect to β^0 are obtain as:

$$\frac{\partial \log f(X_i;\beta^0)}{\partial \beta^0} = (b'(z_i,\beta^0))^T \left\{ \underset{\sim}{T}(X_i) + \psi(b(z_i,\beta^0)) \right\} \tag{3.9}$$

and

$$\frac{\partial^2 \log f(X_i;\beta^0)}{\partial \beta^0 \partial \beta^{0^T}} = (b'(z_i,\beta^0))^T \varphi(b(z_i,\beta^0))\, b'(z_i,\beta^0)$$

$$+ [(b''(z_i,\beta^0))[\underset{\sim}{T}(X_i) + \psi(b(z_i,\beta^0))]. \tag{3.10}$$

Then from (3.3) and (3.10) the Fisher information from the i^{th} data point is

$$I_i(\beta^0) = E\Big\{(b'(z_i,\beta^0))^T[\underset{\sim}{T}(X_i) + \psi(b(z_i,\beta^0))][\underset{\sim}{T}(X_i) + \psi(b(z_i,\beta^0))]^T b'(z_i,\beta^0)$$

$$= - (b'(z_i,\beta^0))^T \varphi(b(z_i,\beta^0)\, b'(z_i,\beta^0). \tag{3.11}$$

The estimating equations for β^0 are developed in the same fashion as in Chapter 2, and are given by

$$U_n(\beta^0) = \frac{\partial}{\partial \beta^0} \log L_n(\beta^0; X, \underset{\sim}{Z})$$

$$= \sum_{i=1}^{n} (\mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0))^T \, [\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \boldsymbol{\psi}(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))]. \tag{3.12}$$

Note that the likelihood estimating equation (3.12) is now a $(k+1)\times 1$ vector which can be written as

$$\mathrm{U}_n(\boldsymbol{\beta}^0) = \frac{\partial}{\partial \boldsymbol{\beta}^0} \sum_{i=1}^{n} \log \mathrm{f}(\mathrm{X}_i;\boldsymbol{\beta}^0)$$

$$= (\mathrm{U}_{n1}(\boldsymbol{\beta}^0), \, \mathrm{U}_{n2}(\boldsymbol{\beta}^0) \,)^T,$$

where the first $(k\times 1)$ rows are defined by

$$\mathrm{U}_{n1}(\boldsymbol{\beta}^0) = \sum_{i=1}^{n} \partial \log \mathrm{f}(\mathrm{X}_i;\boldsymbol{\beta}^0)/\partial \boldsymbol{\beta}^0$$

and the $(k+1)^{th}$ element is

$$\mathrm{U}_{n2}(\boldsymbol{\beta}^0) = \sum_{i=1}^{n} \partial \log \mathrm{f}(\mathrm{X}_i;\boldsymbol{\beta}^0)/\partial \phi \, .$$

So that the MLE $\widehat{\boldsymbol{\beta}}^0$ is the solution to the equation (3.12) set equal to zero, and satisfies the equation

$$\mathrm{U}_n(\widehat{\boldsymbol{\beta}}^0_n) = \mathbf{0}. \tag{3.13}$$

In general, explicit closed-form expressions for the MLE $\widehat{\boldsymbol{\beta}}^0_n$ may not exist. In such cases iterative methods like the **Newton-Raphsom** techniques for finding the MLE can be used as described below. Consider the first order Taylor's series expansion of the estimating equation around some initial guess $\boldsymbol{\beta}^{0(0)}_n$:

$$\mathbf{0} = \mathrm{U}_n(\boldsymbol{\beta}^0)\Big|_{\widehat{\boldsymbol{\beta}}^0_n}$$

$$= \mathrm{U}_n(\boldsymbol{\beta}^0)\Big|_{\boldsymbol{\beta}^{0(0)}_n} + \left(\widehat{\boldsymbol{\beta}}^0 - \boldsymbol{\beta}^{0(0)}_n\right) \mathrm{U}'_n(\boldsymbol{\beta}^0)\Big|_{\boldsymbol{\beta}^*_n},$$

where $\boldsymbol{\beta}^*_n$ is on the line segment joining $\widehat{\boldsymbol{\beta}}^0_n$, $\boldsymbol{\beta}^{0(0)}_n$ and $\mathrm{U}'_n(\boldsymbol{\beta}^0) = \dfrac{\partial \mathrm{U}_n(\boldsymbol{\beta}^0)}{\partial \boldsymbol{\beta}^{0^T}}$.
Then we have

$$\widehat{\boldsymbol{\beta}}^0_n = \boldsymbol{\beta}^{0(0)}_n - \left\{\mathrm{U}_n(\boldsymbol{\beta}^0)\Big|_{\boldsymbol{\beta}^{0(0)}_n}\right\}\left\{\mathrm{U}'_n(\boldsymbol{\beta}^0)\Big|_{\boldsymbol{\beta}^*_n}\right\}^{-1}$$

and by choosing $\boldsymbol{\beta}^{0(0)}_n$ in some neighborhood of $\widehat{\boldsymbol{\beta}}^0_n$ we may use the iteration procedure to obtain a first step estimator:

$$\widehat{\boldsymbol{\beta}}^{0(1)}_n = \widehat{\boldsymbol{\beta}}^{0(0)}_n - \left\{\mathrm{U}_n(\boldsymbol{\beta}^0)\Big|_{\widehat{\boldsymbol{\beta}}^{0(0)}_n}\right\}\left\{\mathrm{U}'_n(\boldsymbol{\beta}^0)\Big|_{\widehat{\boldsymbol{\beta}}^{0(0)}_n}\right\}^{-1}.$$

The process is repeated by updating $\widehat{\boldsymbol{\beta}}^{0(1)}_n$. So that at the i^{th} iteration estimates obtained from the previous $(i-1)^{th}$ are used to refine the current estimates until the process is stopped when $\widehat{\boldsymbol{\beta}}^{0(i)}_n - \widehat{\boldsymbol{\beta}}^{0(i-1)}_n$ is sufficiently small.

To derive the variance covariance matrix of $\widehat{\boldsymbol{\beta}}^0_n$, which is the inverse of the Fisher information matrix, we proceed as follows. Observe that

$$\mathrm{U}_n(\boldsymbol{\beta}^0)\mathrm{U}_n(\boldsymbol{\beta}^0)^T$$

$$= \sum_{i=1}^{n} (\mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0))^T \left[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \boldsymbol{\psi}(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))\right]\left[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \boldsymbol{\psi}(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))\right]^T \mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0)$$

$$+ \sum_{i \neq i'} (\mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0))^T \, [\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \boldsymbol{\psi}(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))] \, [\underset{\sim}{\mathrm{T}}(\mathrm{X}_{i'}) + \boldsymbol{\psi}(\mathbf{b}(\mathbf{z}_{i'},\boldsymbol{\beta}^0))]^T \mathbf{b}'(\mathbf{z}_{i'},\boldsymbol{\beta}^0) \tag{3.14}$$

Since the X_i's are independent, taking expectation of the vector (3.14), and noting from (3.3) that $\mathrm{E}[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) = -\boldsymbol{\psi}(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))$, the second term $\equiv \mathbf{0}$, hence we get

$$\tfrac{1}{n}\mathrm{E}[\mathrm{U}_n(\boldsymbol{\beta}^0)\mathrm{U}_n(\boldsymbol{\beta}^0)^T]$$

$$= \tfrac{1}{n}\mathrm{E}\sum_{i=1}^{n} (\mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0))^T \, [\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \boldsymbol{\psi}(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))] \, [\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \boldsymbol{\psi}(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))]^T \mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0)$$

$$= -\tfrac{1}{n} \sum_{i=1}^{n} (\mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0))^T \, \boldsymbol{\varphi}(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0)) \, \mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0)$$

$$= \bar{\mathrm{I}}_n(\boldsymbol{\beta}^0). \tag{3.15}$$

where $\bar{\mathrm{I}}_n(\boldsymbol{\beta}^0)$ is the average Fisher information matrix for the n data points which from (3.11) is also given by

$$\bar{\mathrm{I}}_n(\boldsymbol{\beta}^0) = \tfrac{1}{n} \sum_{i=1}^{n} \mathrm{I}_i(\boldsymbol{\beta}^0)$$

The second partial derivatives with respect to $\boldsymbol{\beta}^0$ of the log-likelihood function is given by

$$\frac{\partial}{\partial \boldsymbol{\beta}^{0^T}} \mathrm{U}_n(\boldsymbol{\beta}^0) = \sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\beta}^0 \partial \boldsymbol{\beta}^{0^T}} \log \mathrm{f}(\mathrm{X}_i;\boldsymbol{\beta}^0)$$

$$= \sum_{i=1}^{n} (\mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0))^T \boldsymbol{\varphi}(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0)) \, \mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0) + (\mathbf{b}''(\mathbf{z}_i,\boldsymbol{\beta}^0))^T \, [\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \boldsymbol{\psi}(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))] \tag{3.16}$$

Taking the expected value of the expression in (3.16) the second term is zero and we have

$$-\frac{1}{n}\mathrm{E}[\frac{\partial}{\partial\overline{\beta}^{T}}\mathrm{U}_n(\overline{\beta})] = -\frac{1}{n}\sum_{i=1}^{n}(\mathbf{b}'(\mathbf{z}_i,\beta^0))^T\varphi(\mathbf{b}(\mathbf{z}_i,\beta^0))\mathbf{b}'(\mathbf{z}_i,\beta^0)$$

$$= \overline{\mathrm{I}}_n(\beta^0). \tag{3.17}$$

Since $\mathrm{E}[\mathrm{U}_n(\beta^0)] = \mathbf{0}$, $\overline{\mathrm{I}}_n(\beta^0)$ is also the covariance matrix of $\mathrm{U}_n(\beta^0)$.

Hence from (3.16) we write

$$-\frac{\partial}{\partial\beta^{0^T}}\mathrm{U}_n(\beta^0) = \sum_{i=1}^{n}\mathrm{I}_i(\beta^0) - \sum_{i=1}^{n}\Big\{ [\, \mathbf{b}''(\mathbf{z}_i,\beta^0)[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \psi(\mathbf{b}(\mathbf{z}_i,\beta^0))] \tag{3.18}$$

To study the behavior of the second term in (3.16) in large samples, we assume

$$-\,\mathbf{b}''(\mathbf{z}_i,\beta^0)\,\varphi(\mathbf{b}(\mathbf{z}_i,\beta^0))\,\mathbf{b}''(\mathbf{z}_i,\beta^0)^T = \gamma_i(\beta^0) > 0, \quad (i = 1,\ldots,n)$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\gamma_i(\beta^0) \longrightarrow \underset{\sim}{\Gamma}(\beta^0) \qquad \text{(p.d.)} \qquad \text{as } n \longrightarrow \infty \tag{3.19}$$

Let

$$\boldsymbol{Q}_n(\beta^0) = -\sum_{i=1}^{n}\Big\{ (\mathbf{b}''(\mathbf{z}_i,\beta^0))^T\,[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \psi(\mathbf{b}(\mathbf{z}_i,\beta^0))]$$

and note that $\boldsymbol{Q}_n(\beta^0)$ is a sum of independent random variates centered around their respective means and

$$E[\tfrac{1}{n}\boldsymbol{Q}_n(\boldsymbol{\beta}^0)] = \mathbf{0}$$

and by (3.6) and (3.19) as $n \longrightarrow \infty$

$$\mathrm{Cov}[\tfrac{1}{n}\boldsymbol{Q}_n(\boldsymbol{\beta}^0)] = -\frac{1}{n^2}\sum_{i=1}^{n} (\mathbf{b}''(\mathbf{z}_i,\boldsymbol{\beta}^0))^T\, \varphi(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))\, \mathbf{b}''(\mathbf{z}_i,\boldsymbol{\beta}^0))^T \longrightarrow \mathbf{0} \tag{3.20}$$

So by Chebyschev's inequality, the second term in (3.16) divided by n is

$$\frac{1}{n}\sum_{i=1}^{n} (\mathbf{b}''(\mathbf{z}_i,\boldsymbol{\beta}^0))^T\left\{\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \psi(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))\right\} \longrightarrow^{p} 0 \quad \text{as } n \longrightarrow \infty \tag{3.21}$$

Hence the LHS of (3.18) divided by n is the sample average of the $\mathbf{I}_i(\boldsymbol{\beta}^0)$ plus the remainder term which is asymptotically negligible.

To study the asymptotic behavior of $\hat{\boldsymbol{\beta}}_n^0$ we make following assumptions parallel to the assumptions in Chapter I about the behavior of the second partial derivative of the log-likelihood in some neighborhood of $\boldsymbol{\beta}^0$. Note that the loglikelihood is a function of the $\mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0)$, hence we need some assumptions about the continuity behavior of the $\mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0)$ and $\psi(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))$ in some small neighborhood of $\boldsymbol{\beta}^0$. In particular we assume for

$$\| \boldsymbol{\beta}^{0*} - \boldsymbol{\beta}^0)\| < \delta, \text{ and } \delta \longrightarrow 0$$

1a. $$\sup \frac{1}{n}\sum_{i=1}^{n} \| \left\{ \mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0)\Big|_{\boldsymbol{\beta}^{0*}} - \mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0) \right\} \underset{\sim}{\mathrm{T}}(\mathrm{X}_i) \| \longrightarrow \mathbf{0}$$

1b. $$\sup \frac{1}{n}\sum_{i=1}^{n} \| (\mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0))^T\psi(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0))\Big|_{\boldsymbol{\beta}^{0*}} - (\mathbf{b}'(\mathbf{z}_i,\boldsymbol{\beta}^0))^T\psi(\mathbf{b}(\mathbf{z}_i,\boldsymbol{\beta}^0)) \| \longrightarrow \mathbf{0}$$

2a. $\sup \frac{1}{n} \sum_{i=1}^{n} \| \left\{ \mathbf{b}''(\mathbf{z}_i,\beta^0)\big)_{|\beta^{0*}} - \mathbf{b}''(\mathbf{z}_i,\beta^0) \right\} \underset{\sim}{\mathrm{T}}(\mathrm{X}_i) \| \longrightarrow 0$

2b. $\sup \frac{1}{n} \sum_{i=1}^{n} \| \mathbf{b}''(\mathbf{z}_i,\beta^0)\psi(\mathbf{b}(\mathbf{z}_i,\beta^0))_{|\beta^{0*}} - \mathbf{b}''(\mathbf{z}_i,\beta^0)\psi(\mathbf{b}(\mathbf{z}_i,\beta^0)) \| \longrightarrow 0$

2c.

$$\sup \frac{1}{n} \sum_{i=1}^{n} \| \left\{ \left(\mathbf{b}'(\mathbf{z}_i,\beta^0)\right)^T \varphi(\mathbf{b}(\mathbf{z}_i,\beta^0))\mathbf{b}'(\mathbf{z}_i,\beta^0)_{|\beta^{0*}} - \left(\mathbf{b}'(\mathbf{z}_i,\beta^0)\right)^T \varphi(\mathbf{b}(\mathbf{z}_i,\beta^0))\mathbf{b}'(\mathbf{z}_i,\beta^0) \| \right\}$$

$$\longrightarrow 0$$

and

$$\mathrm{E}_{\beta^0} \Big\{ \sup \frac{1}{n} \| \sum_{i=1}^{n} \left[(\mathbf{b}''(\mathbf{z}_i,\beta^0))^T \left[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \psi(\mathbf{b}(\mathbf{z}_i,\beta^0))\right]_{|\beta^{0*}} \right.$$

$$\left. - \frac{1}{n} \sum_{i=1}^{n} \left[(\mathbf{b}''(\mathbf{z}_i,\beta^0))^T \left[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \psi(\mathbf{b}(\mathbf{z}_i,\beta^0))\right] \| \right\} \longrightarrow 0 \quad \text{as} \quad \delta \longrightarrow 0 \qquad (3.22)$$

We also make the assumption that as $n \longrightarrow \infty$,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_i(\beta^0) = \overline{\mathrm{I}}_n(\beta^0) \longrightarrow \mathrm{I}(\beta^0). \qquad (3.23)$$

So that by (3.22) and (3.23)

$$-\frac{1}{n} \frac{\partial}{\partial \beta^{0T}} \mathrm{U}_n(\beta^0)_{|\beta^{0*}} \longrightarrow^{p} \mathrm{I}(\beta^0). \qquad (3.24)$$

Note that $\frac{1}{n}\mathrm{U}_n(\beta^0)$ is the sum of independent variates having finite second moment $\overline{\mathrm{I}}_n(\beta^0)$ moreso the variates are centered around their means. Since the second moment of the $\underset{\sim}{\mathrm{T}}(\mathrm{X}_i)$ and hence of the $\mathrm{U}_n(\beta^0)$ exist and is finite, we can applying the Liaponov Theorem to claim

$$\frac{1}{\sqrt{n}} U_n(\beta^0) \longrightarrow^{\mathcal{D}} N\left(0, I(\beta^0)\right). \tag{3.25}$$

Next we consider the first order Taylor's expansion of $U_n(\hat{\beta}^0_n)$ around β^0.

$$0 = \frac{1}{n} U_n(\hat{\beta}^0{}_n) = \frac{1}{n} U_n(\beta^0) + \frac{1}{n}\frac{\partial}{\partial\beta^0} U_n(\beta^0)\Big|_{\beta^{0*}} (\hat{\beta}^0_n - \beta^0). \tag{3.26}$$

Furthermore since $\frac{1}{\sqrt{n}} U_n(\beta^0)$ is $O_p(1)$, $\frac{1}{n}\frac{\partial}{\partial\beta^0} U_n(\beta^0)\Big|_{\beta^{0*}} = O_p(1)$ and the LHS of (3.26) is $\mathbf{0}$ and also using (3.24) we conclude that

$$(\hat{\beta}^0_n - \beta^0) = o_p(1). \tag{3.27}$$

From (3.26) we have

$$\sqrt{n}(\hat{\beta}^0_n - \beta^0) = \frac{1}{\sqrt{n}} U_n(\beta^0)\, I^{-1}(\beta^0) + o(1). \tag{3.28}$$

By (3.25) the first term in the RHS of (3.28) is asymptotically normal and hence by the Slutsky's theorem LHS of (3.28) is

$$\sqrt{n}(\hat{\beta}^0_n - \beta^0) \longrightarrow^{\mathcal{D}} N\left(0,\ I^{-1}(\beta^0)\right). \tag{3.29}$$

Special Case

In the classical linear model we have $b(z_i^T\beta,\phi) = z_i^T\beta$, $-\psi(b(z_i^T\beta)) = z_i^T\beta$ and $\underset{\sim}{T}(X_i) = T(X_i)$, in this case ϕ is a nuisance parameter which does not influence the

estimating equations for β, hence without loss of generality ϕ can be taken to equal unity, so that the maximum likelihood estimating equations (3.12), for β becomes

$$\sum_{i=1}^{n} \mathbf{z}_i \, [\mathrm{T}(\mathrm{X}_i) - \mathbf{z}_i^T \boldsymbol{\beta} \,] \tag{3.30}$$

which leads to the MLE $\widehat{\boldsymbol{\beta}}_n$ given by

$$\widehat{\boldsymbol{\beta}}_n = \Big\{ \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^T \Big\}^{-1} \sum_{i=1}^{n} \mathbf{z}_i \mathrm{T}(\mathrm{X}_i). \tag{3.31}$$

Example 3.1 Let $\log \mathrm{X}_1, \ldots, \log \mathrm{X}_n$ be i.i.d. r.v.'s following a normal distribution with mean $\mathbf{z}_i^T \boldsymbol{\beta}$ and common variance $\phi \equiv \sigma^2$ $(i = 1, \ldots, n)$. Then the X_i have a lognormal density given by

$$\mathrm{f}(\mathrm{X}_i; \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi}\, \sigma \mathrm{X}_i} \, \mathrm{e}^{-\frac{1}{\sigma^2}(\log \mathrm{X}_i - \mathbf{z}_i^T \boldsymbol{\beta})^2}$$

$$= \mathrm{c}(\mathrm{X}) \, \mathrm{e}^{\underset{\sim}{\mathrm{T}}(\mathrm{X}_i)^T \mathbf{b}(\mathbf{z}_i^T \boldsymbol{\beta}, \sigma^2) + \Psi(\mathbf{b}(\mathbf{z}_i^T \boldsymbol{\beta}, \sigma^2))}, \qquad \mathrm{X}_i > 0 \quad (i = 1, \ldots, n)$$

where

$$\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) = (\, \mathrm{T}_1(\mathrm{X}_i), \, \mathrm{T}_2(\mathrm{X}_i))^T = (\, \log \mathrm{X}_i, \; (\log \mathrm{X}_i)^2)^T,$$

$$\mathbf{b}(\mathbf{z}_i^T \boldsymbol{\beta}, \sigma^2) = \big(\mathrm{b}_1(\mathbf{z}_i^T \boldsymbol{\beta}, \sigma^2), \, \mathrm{b}_2(\mathbf{z}_i^T \boldsymbol{\beta}, \sigma^2) \big) = \Big(\frac{\mathbf{z}_i^T \boldsymbol{\beta}}{\sigma^2}, \; -\frac{1}{\sigma^2} \Big)^T$$

and

$$\Psi(\mathbf{b}(\mathbf{z}_i^T \boldsymbol{\beta}, \sigma^2) = -\tfrac{1}{2} \Big(\frac{\mathrm{b}_1(\mathbf{z}_i^T \boldsymbol{\beta}, \sigma^2)^2}{\mathrm{b}_2(\mathbf{z}_i^T \boldsymbol{\beta}, \sigma^2)} + \log -2\mathrm{b}_2(\mathbf{z}_i^T \boldsymbol{\beta}, \sigma^2) \Big).$$

The partial derivatives of $\mathbf{b}(\mathbf{z}_i^T\boldsymbol{\beta},\sigma^2)$ are

$$\mathbf{b}'(\mathbf{z}_i^T\boldsymbol{\beta},\sigma^2) = \left(\mathrm{b}_1'(\mathbf{z}_i^T\boldsymbol{\beta},\sigma^2)\ ,\ \mathrm{b}_2'(\mathbf{z}_i^T\boldsymbol{\beta},\sigma^2) \right) = \left(\frac{\mathbf{z}_i^T\boldsymbol{\beta}}{\sigma^2}\ ,\ -\frac{1}{\sigma^2} \right)$$

$$= \begin{bmatrix} \frac{1}{\sigma^2}\mathbf{z}_i & 0 \\ -\frac{1}{\sigma^4}\mathbf{z}_i^T\boldsymbol{\beta} & \frac{1}{2\sigma^4} \end{bmatrix}$$

and

$$\boldsymbol{\psi}(\mathbf{b}(\mathbf{z}_i^T\boldsymbol{\beta},\sigma^2)) = \frac{1}{2}\begin{bmatrix} \dfrac{\mathrm{b}_1(\mathbf{z}_i^T\boldsymbol{\beta},\sigma^2)}{\mathrm{b}_2(\mathbf{z}_i^T\boldsymbol{\beta},\sigma^2)} \\ -\dfrac{\mathrm{b}_1(\mathbf{z}_i^T\boldsymbol{\beta},\sigma^2)^2}{2\mathrm{b}_2(\mathbf{z}_i^T\boldsymbol{\beta},\sigma^2)^2} + \dfrac{1}{\mathrm{b}_2(\mathbf{z}_i^T\boldsymbol{\beta},\sigma^2)} \end{bmatrix},$$

so that the MLE of $(\boldsymbol{\beta},\sigma^2)$ is a solution to the $(k+1)\times 1$ estimating equations

$$\mathrm{U}_n(\boldsymbol{\beta},\sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}\sum(\log \mathrm{X}_i - \mathbf{z}_i^T\boldsymbol{\beta})\mathbf{z}_i \\ \frac{1}{2\sigma^4}\sum(\log \mathrm{X}_i - \mathbf{z}_i^T\boldsymbol{\beta})^2 - \frac{n}{2\sigma^2} \end{bmatrix} = \mathbf{0}.$$

3.2.1 Bias of the Regression Parameter Estimates

We have already shown, in chapter II, that the MLE are generally not

unbiased and that the leading term in the bias of the MLE is of order n^{-1}, which was reduced by jackknifing the MLE. In this section we establish the bias of the regression parameters and demonstrate the extent to which the classical Jackknife estimator can reduce the bias of the MLE of the regression coefficients.

Since the $\mathbf{z}_i$ are most often not the same, the $E[\underset{\sim}{T}(X_i)]$ which are functions of the $\mathbf{z}_i$, would generally be different and may not be identically distributed. Then the estimating equations are a sum of independent but not identically distributed random variables. The parameter vector β^0 enters in the estimating functions as a product of the $\underset{\sim}{T}(X_i)$, and further the estimating functions are an implicit function of the vector β^0, hence some iteration procedures are needed to estimate the parameters. Given that we can not write an expression of $\widehat{\beta}_n^0$ explicitly from the estimating equations, we rely on Taylor expansion of the estimating equations, to study the large-sample bias in $\widehat{\beta}_n^0$.

Proposition 3.1 The MLE $\widehat{\beta}_n^0$ has an asymptotic bias of order n^{-1}. For simplicity of presentation we write U_n for $U_n(\beta^0)$. The j^{th} component of $U_n(\beta^0)$ will be denoted by U_{nj} $(j = 1, 2, \ldots, k+1)$.

$$E[\,\widehat{\beta}_n^0 - \beta^0] = \frac{a(\beta^0)}{n} + \frac{b(\beta^0)}{n^2} + O_p(\frac{1}{n^2})$$

where $a(.)$ and $b(.)$ are known functions of β^0 and are independent of n.

Proof.

Note that the $\mathbf{z}_i$ and $\mathbf{b}$ are fixed and known vectors.

Assume there is a function $H : \mathbb{R}^{k+1} \to \mathbb{R}^{k+1}$ given by $H(U_n(\beta^0)) = \beta^0$, so that

$\mathbf{H}$ is a mapping from the estimating equations space into the parameter space. Let $\mathbf{H} = (h_1, \ldots, h_{k+1})$ where the elements h_l ($l = 1, \ldots, k+1$) of $\mathbf{H}$ have continuous fourth order partial derivatives with respect to the vector $U_n(\beta^0)$. We also assume the fourth moment of $U_n(\beta^0)$ exist and is finite, that is,

$E||U_n(\beta^0)||^4 < \infty$

Note that $E[U_n] = \mathbf{0}$ and by definition $U_n(\beta_n^0) = \mathbf{0}$. Then by Taylor's series expansion we write

$$\hat{\beta}_n^0 - \beta^0 = \mathbf{H}(U_n(\hat{\beta}_n^0)) - \mathbf{H}(U_n(\beta^0))$$

$$= -\frac{\partial \mathbf{H}(U_n)}{\partial U_n} U_n + \frac{1}{2} U_n{}^T \frac{\partial^2 \mathbf{H}(U_n)}{\partial U_n \partial U_n^T} U_n - \frac{1}{3!} U_n U_n^T \frac{\partial^3 \mathbf{H}(U_n)}{\partial U_n \partial U_n^T \partial U_n} U_n$$

$$+ \frac{1}{4!} U_n U_n{}^T \frac{\partial^4 \mathbf{H}(U_n)}{\partial U_n \partial U_n^T \partial U_n \partial U_n^T} U_n U_n{}^T$$

$$+ \cdots \qquad (3.32)$$

As in (2.1.30) consider the l^{th} component of $\mathbf{H}(U_n)$ say $h_l(U_n)$ and present the corresponding Taylor's expansion of $h_l(U_n)$ about it's expected null vector as

$h_l(\, U_n(\hat{\beta}_n^0)\,) - h_l(U_n(\beta^0)\,)$

$$= -\frac{\partial h_l(U_n)}{\partial U_n} U_n + \frac{1}{2} U_n{}^T \frac{\partial^2 h_l(U_n)}{\partial U_n \partial U_n^T} U_n - \frac{1}{3!} U_n U_n^T \frac{\partial^3 h_l(U_n)}{\partial U_n \partial U_n^T \partial U_n} U_n$$

$$+ \frac{1}{4!} U_n U_n^T \frac{\partial^4 h_l(U_n)}{\partial U_n \partial U_n^T \partial U_n \partial U_n^T} U_n U_n^T$$

$$+ \cdots \tag{3.33}$$

Taking the expectation of both sides of (3.33) the first term becomes zero. The second term can be written as

$$\mathrm{E}[\mathrm{U}_n{}^T \frac{\partial^2 \mathrm{h}_l(\mathrm{U}_n)}{\partial \mathrm{U}_n^T \partial \mathrm{U}_n} \mathrm{U}_n]$$

$$= \mathrm{E}[\mathrm{Tr}(\underset{\sim}{\mathrm{A}}\,\underset{\sim}{\mathrm{B}})] = \mathrm{E}[\sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij}^{(l)} b_{ji}]$$

$$= \tfrac{1}{n} [\sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij}^{(l)} \sigma_{ji}], \tag{3.34}$$

where

$$\underset{\sim}{\mathrm{A}} = \frac{\partial^2 \mathrm{h}_l(\mathrm{U}_n)}{\partial \mathrm{U}_n^T \partial \mathrm{U}_n} = \left(a_{ij}^{(l)} \right) i, j = 1, \ldots, k$$

$$\left(a_{ji}^{(l)}\right) = \frac{\partial^2 \mathrm{h}_l(\mathrm{U}_n)}{\partial \mathrm{U}_{nj} \partial \mathrm{U}_{ni}} .$$

$$\underset{\sim}{\mathrm{B}} = \mathrm{U}_n \mathrm{U}_n^T = \left(b_{ji}\right) ,$$

$$b_{ji} = (\mathrm{U}_n)_j (\mathrm{U}_n)_i \qquad \text{and}$$

$$\sigma_{ji} = \mathrm{E}[b_{ji}]$$

$$= \Big(- \tfrac{1}{n} \sum_{i=1}^{n} (\mathbf{b}'(\mathbf{z}_i, \beta^0))^T \varphi(\mathbf{z}_i, \beta^0)\, \mathbf{b}'(\mathbf{z}_i, \beta^0) \Big)_{ji} \quad \text{is the } (j,i)^{th} \text{ element of the}$$

covariance matrix of $U_n(\hat{\beta}_n^0)$.

The third term in (3.33) is obtained from differentiating (3.34) with respect to U_n and taking the expectation and is expressed as

$$E[\sum_{r=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{k} c_r a_{rij}^{(l)} b_{ji}] = \frac{1}{n^2}\sum_{r=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{k} a_{rij}^{(l)}\gamma_{rji}$$

where

$$a_{rij}^{(l)} = \Big(\frac{\partial a_{ij}^{(l)}}{\partial U_{nji}}\Big)_r = \frac{\partial^3 h_l(U_n)}{\partial U_{nr}\partial U_{nj}\partial U_{ni}}$$

$$\gamma_{rji} = E[c_r b_{jr}] = E[\, U_{nr}\, U_{nj}\, U_{ni}\,]$$

$$= \Big(-\frac{1}{n}\sum_{i=1}^{n} \mathbf{b}'(\mathbf{z}_i,\beta^0)^T\, \mathbf{b}'(\mathbf{z}_i,\beta^0)\, \Gamma(\mathbf{z}_i,\beta^0)\, \mathbf{b}'(\mathbf{z},\beta^0)\Big)_{rji}$$

Differentiating the third term once more with respect to U_n, the fourth term in (3.33) is expressed as

$$E\,[\sum_{s=1}^{k}\sum_{r=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{k} c_s c_r a_{rij}^{(l)} b_{ji}]$$

$$= \frac{1}{n^2}\,[\sum_{s=1}^{k}\sum_{r=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{k} a_{srij}^{(l)}\lambda_{srji}]$$

where λ_{srji} is the $(s\ r\ j\ i)^{th}$ element of the fourth central moment of $U_n(\hat{\beta}_n^0)$
As in (2.1.34) we can then express (3.33) as

$$E[h_l(\,U_n(\hat{\beta}_n^0)\,) - h_l(\,U_n(\beta^0)\,)\,] = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}^{(l)}\sigma_{ji}}{n}$$

$$+ \frac{\sum_{r=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{k} a_{rij}^{(l)}\gamma_{rji} + \sum_{s=1}^{k}\sum_{r=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{k} a_{srij}^{(l)}\lambda_{srji}}{n^2}$$

$$+ O(n^{-3}),\ \ l = 1,\ldots,k \tag{3.35}$$

and hence the bias of $\hat{\beta}_n^0$ from (3.32) becomes

$$E[\hat{\beta}_n^0 - \beta^0] = \tfrac{1}{n}\, a(\beta^0) + \tfrac{1}{n^2}\, b(\beta^0) + o(\tfrac{1}{n^2}) \qquad \text{as } n \longrightarrow \infty, \tag{3.36}$$

where

$$a(\beta^0) = \Big(\tfrac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}^{(l)}\sigma_{ji}\Big) \text{ and}$$

$$b(\beta^0) = \Big(\tfrac{1}{3!}\sum_{r=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{k} a_{rij}^{(l)}\gamma_{rji} + \tfrac{1}{4!}\sum_{s=1}^{k}\sum_{r=1}^{k}\sum_{i=1}^{k}\sum_{j=1}^{k} a_{srij}^{(l)}\lambda_{srji}\Big)$$

are functions of β, ϕ and z_i's independent of n.

In section 3.2 we shall be concerned with the Jackknife estimation aimed at eliminating the leading term in the bias of $\hat{\beta}_n^0$ given by $\frac{1}{n}\, a(\beta^0)$.

3.3 Jackknife Estimator of the Regression Model Parameters

We wish to establish the asymptotic bias and distribution of the Jackknife

estimate $\widetilde{\beta}^J$ of β^0. The Jackknife estimator is obtained by deleting in succession one row of the observation vector X and computing the MLE $\widehat{\beta}^{0(i)}_{n-1}$ and is defined below.

Let $\widehat{\beta}^0{}_n$ and $\widehat{\beta}^{0(i)}_{n-1}$ be the MLE's of β^0 based on all the n and $n-1$ observations, respectively, and define the pseudo values

$$\widehat{\beta}^0_{n,i} = n\,\widehat{\beta}^0_n - (n-1)\,\widehat{\beta}^{0(i)}_{n-1}, \qquad i = 1,\ldots,n \tag{3.37}$$

The Jackknife estimate of β^0 is obtained by averaging the pseudo-values (3.37) and is defined as

$$\begin{aligned}\widehat{\beta}^J_n &= \tfrac{1}{n}\sum_{i=1}^{n}\widehat{\beta}^0_{n,i} \\ &= \widehat{\beta}^0_n - (n-1)\,\tfrac{1}{n}\sum_{i=1}^{n}(\widehat{\beta}^{0(i)}_{n-1} - \widehat{\beta}^0_n).\end{aligned} \tag{3.38}$$

In order to study the Jackknife estimators of β^0 we need establish the properties of the delete-one estimators $\widehat{\beta}^{0(i)}_{n-1}$, obtained from the estimating equations with one observation deleted from the data. Since the explicit form of $\widehat{\beta}^0_n$ may not be available we rely on the estimating equations $\mathrm{U}_n(\widehat{\beta}^0_n)$ and it's Taylor's expansion around β^0 to derive an approximate expression for $\widehat{\beta}^0_n$.

Let $\mathrm{U}^{(i)}_{n-1}(\beta^0)$ denote the delete-one estimating equation for the estimate $\widehat{\beta}^{0(i)}_{n-1}$ of β^0 defined as

$$\mathrm{U}^{(i)}_{n-1}(\beta^0) = \mathrm{U}_n(\beta^0) - (\mathbf{b}'(\mathbf{z}_i,\beta^0))^T\,[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \psi(\mathbf{b}(\mathbf{z}_i,\beta^0))], \tag{3.39}$$

whereby $\widehat{\beta}^{0(i)}_{n-1}$ is the solution to the equation $U^{(i)}_{n-1}(\beta^0) = 0$. Note that when evaluated at $\widehat{\beta}^0_n$, (3.39) satisfies

$$U^{(i)}_{n-1}(\widehat{\beta}^0_n) = -(b'(z_i,\beta^0))^T \left\{ \underset{\sim}{T}(X_i) + \psi(b(z_i,\beta^0)) \right\}\Big|_{\widehat{\beta}^0_n}. \tag{3.40}$$

From (3.40) one obtains the following

$$U^{(i)}_{n-1}(\widehat{\beta}^0_n) - U^{(i)}_{n-1}(\widehat{\beta}^{0(i)}_{n-1}) = -(b'(z_i,\beta^0))^T \left\{ \underset{\sim}{T}(X_i) + \psi(b(z_i,\beta^0)) \right\}\Big|_{\widehat{\beta}^0_n}. \tag{3.41}$$

Unlike in the non regression estimation problem considered in Chapter 2, (3.41) the terms $\underset{\sim}{T}(X_i)$ appear as a product with functions of the estimates of β^0. So that we need a different approach, from the one used in Chapter 2, to study the behavior of the difference in the estimates $(\widehat{\beta}^{0(i)}_{n-1} - \widehat{\beta}^0_n)$, the delete-one estimates and the estimates from the entire sample.

To examine the behavior of $\sum_{i=1}^{n} \left\{ \widehat{\beta}^{0(i)}_{n-1} - \widehat{\beta}^0_n \right\}$ in (3.38) we need expressions of $\widehat{\beta}^{0(i)}_{n-1}$ and $\widehat{\beta}^0_n$. Since the explicit expressions are often not available in various estimation problems, we derive an approximate representation of $\widehat{\beta}^0_n$ by considering the Taylor's expansion of $U_n(\widehat{\beta}^0_n)$. It will be helpful at this time to introduce introduce the following notations and results from matrix algebra which will be used in studying the approximate expression for $\widehat{\beta}^0_n$.

Let

$$v_i(\beta^0) = -(b'(z_i,\beta^0))^T \varphi(b(z_i,\beta^0)) b'(z_i,\beta^0), \qquad i = 1, \ldots, n,$$

$$V_n(\beta^0) = \sum_{i=1}^{n} v_i(\beta^0),$$

$$\mathbf{u}_i(\beta^0) = (\mathbf{b}'(\mathbf{z}_i,\beta^0))^T\ [\mathbf{T}(\mathbf{X}_i) + \ \psi(\mathbf{b}(\mathbf{z}_i,\beta^0))],$$

$$\mathrm{U}_n(\beta^0) = \sum_{i=1}^{n} \mathbf{u}_i(\beta^0).$$

So that for the delete-one expressions we have

$$\mathbf{V}^{(i)}_{n-1}(\beta^0) = \mathbf{V}_n(\beta^0) - \mathbf{v}_i(\beta^0) \text{ and}$$
$$\mathrm{U}^{(i)}_{n-1}(\beta^0) = \mathrm{U}_n(\beta^0) - \mathbf{u}_i(\beta^0), \qquad i = 1,\ldots,n.$$

The matrix $\mathbf{V}_n(\beta^0)$ is symmetric and positive definite and note that $\frac{1}{n}\mathbf{V}_n(\beta^0)$ converges to $\bar{\mathrm{I}}_n(\beta^0)$ which is the average Fisher information. Furthermore assuming each of the $\mathbf{v}_i(\beta^0)$ is small relative to $\mathbf{V}_n(\beta^0)$ we can obtain the $\left(\mathbf{V}^{(i)}_{n-1}(\beta^0)\right)^{-1}$, the inverse of $\mathbf{V}^{(i)}_{n-1}(\beta^0)$ as

$$\left(\mathbf{V}^{(i)}_{n-1}(\beta^0)\right)^{-1} = \left(\mathbf{V}_n(\beta^0) - \mathbf{v}_i(\beta^0)\right)^{-1}$$

$$= \left\{\mathrm{I} + \left(\mathbf{V}_n(\beta^0)\right)^{-1}\mathbf{v}_i(\beta^0) + \left(\mathbf{V}_n(\beta^0)\right)^{-1}\mathbf{v}_i(\beta^0)\mathbf{v}_i(\beta^0)^T\left(\mathbf{V}_n(\beta^0)\right)^{-1} + \ldots\right\}\left(\mathbf{V}_n(\beta^0)\right)^{-1}$$

$$= \left(\mathbf{V}_n(\beta^0)\right)^{-1} + \left(\mathbf{V}_n(\beta^0)\right)^{-1}\mathbf{v}_i(\beta^0)\left(\mathbf{V}_n(\beta^0)\right)^{-1} + \mathrm{O}(n^{-3}).$$

Also the following result can be proved by matrix algebra,

$$\left(\mathbf{V}_n(\beta^0)\right)^{-1} - \left(\mathbf{V}^{(i)}_{n-1}(\beta^0)\right)^{-1} = \left(\mathbf{V}_n(\beta^0)\right)^{-1} - \left\{\mathrm{I} - \left(\mathbf{V}_n(\beta^0)\right)^{-1}\mathbf{v}_i(\beta^0)\right\}^{-1}\left(\mathbf{V}_n(\beta^0)\right)^{-1}$$

$$= \left(\mathbf{V}_n(\beta^0)\right)^{-1} \mathbf{v}_i(\beta^0) \left(\mathbf{V}_n(\beta^0)^{-1} + \mathrm{O}(n^{-3}).\right.$$

Now the Taylor's expansion of $\mathrm{U}_n(\hat{\beta}^0_n)$ around β^0 given by

$$\mathrm{U}_n(\hat{\beta}^0_n) = \mathrm{U}_n(\beta^0) + \frac{\partial \mathrm{U}_n(\beta^0)}{\partial \beta^0} (\hat{\beta}^0_n - \beta^0) + \tfrac{1}{2}(\hat{\beta}^0 n - \beta^0)^T \frac{\partial \mathrm{U}_n(\beta^0)}{\partial \beta^0} (\hat{\beta}^0_n - \beta^0)$$

$$+ \cdots \tag{3.43}$$

and multiplying both sides of (3.43) by $\left(\mathbf{V}_n(\beta^0)\right)^{-1}$ and noting that $\mathrm{U}_n(\hat{\beta}^0 n) = \mathbf{0}$, for sufficiently large n, $\partial \mathrm{U}_n(\beta^0)/\partial \beta^0 \longrightarrow \mathbf{V}_n(\beta^0)$, we obtain

$$\hat{\beta}^0_n - \beta^0 = \left(\mathbf{V}_n(\beta^0)\right)^{-1} \left\{ \mathrm{U}_n(\beta^0) + \mathrm{W}_n(\beta^0) \right\},$$

where

$\mathbf{W}_n(\beta^0) = \frac{1}{2}(\hat{\beta}^0_n - \beta^0)^T \frac{\partial \mathrm{U}_n(\beta^0)}{\partial \beta^0} (\hat{\beta}^0_n - \beta^0) = \mathrm{O}(1)$ and similarly for the delete-one estimating equations,

$$\hat{\beta}^{0(i)}_{n-1} - \beta^0 = \left(\mathbf{V}^{(i)}_{n-1}(\beta^0)\right)^{-1} \left\{ \mathrm{U}^{(i)}_{n-1}(\beta^0) + \mathbf{W}^{(i)}_{n-1}(\beta^0) \right\}. \tag{3.44}$$

By substituting the expressions for $\left(\mathbf{V}^{(i)}_{n-1}(\beta^0)\right)^{-1}$ and $\mathrm{U}^{(i)}_{n-1}(\beta^0)$ given above we obtain, for large n,

$$\hat{\beta}^0_n - \hat{\beta}^{0(i)}_{n-1} = \left(\mathbf{V}_n(\beta^0)\right)^{-1} \mathbf{u}_i(\beta^0) - \left(\mathbf{V}_n(\beta^0)\right)^{-1} \mathbf{v}_i(\beta^0) \left(\mathbf{V}_n(\beta^0)\right)^{-1} \mathrm{U}_n(\beta^0)$$

$$+ \left(\mathbf{V}_n(\beta^0)\right)^{-1} \mathbf{v}_i(\beta^0) \left(\mathbf{V}_n(\beta^0)\right)^{-1} \mathbf{u}_i(\beta^0) - \left(\mathbf{V}_n(\beta^0)\right)^{-1} \mathbf{v}_i(\beta^0) \left(\mathbf{V}_n(\beta^0)\right)^{-1} \mathrm{O}(1)$$

$$+ O(n^{-3}).$$

If we write $\mathbf{W}_{n-1}^{(i)}(\beta^0) = \mathbf{W}_n(\beta^0) - (\mathbf{W}_n(\beta^0) - \mathbf{W}_{n-1}^{(i)}(\beta^0))$, then

$$\sum_{i=1}^{n} \hat{\beta}_n^0 - \hat{\beta}^{0(i)}{}_{n-1} = \sum_{i=1}^{n} \left(\mathbf{V}_n(\beta^0)\right)^{-1} \mathbf{v}_i(\beta^0) \left(\mathbf{V}_n(\beta^0)\right)^{-1} \left(\mathbf{u}_i(\beta^0) - \mathbf{W}_n(\beta^0)\right) + O(n^{-2}). \tag{3.45}$$

Since $\mathbf{u}_i(\beta^0) = O_p(1)$ and $\left(\mathbf{V}_n(\beta^0)\right)^{-1} = O(n^{-1})$, the first term on the RHS of (3.45) has a negligible contribution to the left hand side of (3.45) and
the second term is $\left(\mathbf{V}_n(\beta^0)\right)^{-1} O_p(1) = O_p(n^{-1}) = o_p(n^{-1/2})$.

Hence $\sum_{i=1}^{n} \hat{\beta}_n^0 - \hat{\beta}^{0(i)}{}_{n-1} = o_p(n^{-1/2})$, which implies that

$$\sqrt{n} \sum_{i=1}^{n} \left\{\hat{\beta}^{0(i)}{}_{n-1} - \hat{\beta}_n^0\right\} \longrightarrow^{p} 0, \qquad \text{as } n \longrightarrow \infty. \tag{3.46}$$

So that from (3.38) we have

$$\sqrt{n}(\hat{\beta}^{0J}{}_n - \hat{\beta}_n^0) = o_p(1).$$

Since $\hat{\beta}_n^0$ is asymptotically normal by (3.29) we conclude by Slutsky's theorem that for large n

$$\sqrt{n}(\hat{\beta}^{0J}{}_n - \beta^0) \longrightarrow^{\mathcal{D}} \mathrm{N}\left(0,\ \bar{\mathrm{I}}_n{}^{-1}(\beta^0)\right). \tag{3.47}$$

For inference purposes then, the Jackknifed estimate could be used in place of the MLE.

To demonstrate the reduction in bias of the MLE $\hat{\beta}_n^0$ achieved by jackknifing, we

express the bias of the Jackknife estimator as

$$\mathrm{E}[\,\tilde{\beta}^{J} - \beta^{0}] \;=\; \mathrm{E}[\,\hat{\beta}^{0}_{n} - \beta^{0}\,] - \frac{(n-1)}{n}\sum_{i=1}^{n}\mathrm{E}[(\hat{\beta}^{0(i)}_{n-1} - \hat{\beta}^{0}_{n})]$$

$$= \mathrm{E}[\,\hat{\beta}^{0}_{n} - \beta^{0}\,] - \frac{(n-1)}{n}\sum_{i=1}^{n}\mathrm{E}[(\hat{\beta}^{0}_{n-1} - \hat{\beta}^{0}_{n})]$$

$$= n\mathrm{E}[\,\hat{\beta}^{0} - \beta^{0}] - (n-1)\mathrm{E}[(\hat{\beta}^{0}_{n-1} - \hat{\beta}^{0}_{n})]$$

and using the results obtained earlier (3.36) on the bias of $\hat{\beta}^{0}_{n}$ we have

$$\mathrm{E}[\,\tilde{\beta}^{J} - \beta^{0}] \;=\; \frac{1}{n(n-1)}\,\mathrm{b}(\beta^{0}) \;+\; \mathrm{o}(\frac{1}{n^{2}}) \qquad \text{as } n \longrightarrow \infty$$

where $\mathbf{b}(\beta^{0})$ is as defined in (3.36).

Chapter 4

Length-Biased Distribution for Exponential Families.

4.1 Introduction.

Let X be a continuous r.v. following a length-biased distribution with p.d.f., $g(X;\boldsymbol{\theta})$ which is a weighted p.d.f. of an exponential family of distributions given by

$$g(X;\boldsymbol{\theta}) = \frac{X\, f(X;\boldsymbol{\theta})}{\gamma(\boldsymbol{\theta})} = c^*(X)\exp\left\{ \underset{\sim}{T}(X)^T\boldsymbol{\theta} + \Psi^*(\boldsymbol{\theta}) \right\}, \qquad X > 0 \tag{4.1}$$

where

$\boldsymbol{\theta} = (\theta_1,\ldots,\theta_k)$ is a k-parameter vector

$\Psi^*(\boldsymbol{\theta}) = \Psi(\boldsymbol{\theta}) - \log\,\gamma(\boldsymbol{\theta})$

$\gamma(\boldsymbol{\theta}) = E_f[X]$

$\underset{\sim}{T}(X) = (\,T_1(X_i),\ldots,T_k(X_i))^T$ is a k-vector of sufficient statistics for $\boldsymbol{\theta}$

$\Psi^*(.)$ and $\gamma(.)$ are known functions of their arguments.

The density $g(X;\boldsymbol{\theta})$ is said to be a length-biased version of the pdf $f(X;\boldsymbol{\theta})$. We study the bias and asymptotic distribution of the MLE ant jackknifed MLE of $\boldsymbol{\theta}$ obtained from the sample of size n drawn from the distribution $g(X;\boldsymbol{\theta})$. Note that the densities $f(X;\boldsymbol{\theta})$ and $g(X;\boldsymbol{\theta})$ are both of exponential family of distributions,

hence with appropriate assumptions on $\Psi^*(\boldsymbol{\theta})$, the results we have for the density $f(X;\boldsymbol{\theta})$ apply as well to the density $g(X;\boldsymbol{\theta})$. Also the parameter $\boldsymbol{\theta}$ is the same for the two distributions, however as will be illustrated in section 4.2 the estimates of $\boldsymbol{\theta}$ under the different likelihoods will differ.

4.2 Maximum Likelihood Estimators of the Parameters

Following similar development discussed in Chapter 2, we have, under the density $g(X;\boldsymbol{\theta})$,

$$\begin{aligned} E_g[\underset{\sim}{T}(X)] &= -\psi^*(\boldsymbol{\theta}) \\ &= -\psi(\boldsymbol{\theta}) + \zeta'(\boldsymbol{\theta}) \end{aligned}$$

where $\zeta'(\boldsymbol{\theta}) = \dfrac{\partial \log \gamma(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \dfrac{\gamma'(\boldsymbol{\theta})}{\gamma(\boldsymbol{\theta})}$

This clearly indicates the sample mean $\bar{\underset{\sim}{T}}_n(X)$ from the length-biased data will be a biased estimator of the original population mean from the density $f(X;\boldsymbol{\theta})$. Furthermore, in a univariate setup with $T(X_i) = X_i$ the bias is the ratio of the variance to the mean of X_i, which is a positive quantity.

Next we consider the estimation of the parameter vector $\boldsymbol{\theta}$ under the length-biased distribution (4.1). For the maximum likelihood estimation of $\boldsymbol{\theta}$, we assume the same set of assumptions "B" as in Chapter 2 with $\Psi(\boldsymbol{\theta})$ replaced by $\Psi^*(\boldsymbol{\theta})$. The estimating equations for $\boldsymbol{\theta}$ are then given by

$$U_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial \log g(X_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \underset{\sim}{T}(X_i) + n\,\psi^*(\boldsymbol{\theta}) = 0. \tag{4.2}$$

So that the MLE $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ is the solution for $\boldsymbol{\theta}$ in (4.2). The gradient of the estimating equation is obtained by differentiating (4.2) with respect to $\boldsymbol{\theta}^T$ and is expressed as

$$-E_{\boldsymbol{\theta}} \frac{\partial U_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = -n\varphi^*(\boldsymbol{\theta}), \quad \text{finite and p.d.} \tag{4.3}$$

Note that the MLE $\hat{\boldsymbol{\theta}}_n$ obtained from solving (4.2) for $\boldsymbol{\theta}$, will be different from the MLE obtained from the original unweighted distribution and the the two estimates can not be compared unless their explicit expressions are available, even then the expressions may not be simple enough for any general comparison to be made. This will be true for comparing the bias of the two estimates, however since the expected Fisher's information from the two probability models is a function of only the parameter $\boldsymbol{\theta}$ conditions under which one model is more informative than the other will be investigated.

4.3 Asymptotic Distribution and Bias of the MLE

First we establish the asymptotic normal distribution of $\hat{\boldsymbol{\theta}}_n$ from (4.2). As noted above the difference in the length-biased version of the exponential family of distributions (4.1) from the original distribution is in the reparametrization of the function $\Psi(.)$, under the regularity set of assumptions "B" we conclude using the same development, that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathrm{N}(0,\ \mathrm{I}_g^{-1}(\boldsymbol{\theta})), \tag{4.4}$$

where $\mathrm{I}_g(\boldsymbol{\theta}) = -\varphi^*(\boldsymbol{\theta})$, the subscript g denotes the length-biased distribution, is the Fisher information from the length-biased distribution characterized by $g(X;\boldsymbol{\theta})$. In section 4.4 we compare the Fisher information from the density $f(X;\boldsymbol{\theta})$ with the Fisher information from $g(X;\boldsymbol{\theta})$.

The structure of the bias of $\hat{\boldsymbol{\theta}}_n$, obtained from (4.2) is similar to the form of bias studied in Chapter 2 except for the quantities involved. So that as in (2.1.35) the leading term in the bias of $\hat{\boldsymbol{\theta}}_n$ under the distribution $g(X;\boldsymbol{\theta})$ is given as

$$\tfrac{1}{n}\boldsymbol{a}^*(\boldsymbol{\theta}) = \Big(\tfrac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij}^{(l)*} \sigma_{ji}^* \Big),$$

which is a similar expression for the leading term in the bias of $\hat{\boldsymbol{\theta}}_n$ under $f(X;\boldsymbol{\theta})$ with $a_{ij}^{(l)}$ and σ_{ji} replaced by $a_{ij}^{(l)*}$ and σ_{ji}^* respectively.
In particular, as in proposition 1, let $\boldsymbol{\eta}^* = -\boldsymbol{\psi}^*(\boldsymbol{\theta})$ and a function H: $\mathbb{R}^k \longrightarrow \mathbb{R}^k$ given by $\mathrm{H}^*(\boldsymbol{\eta}^*) = \boldsymbol{\theta}$, with $\mathrm{H}(\boldsymbol{\eta}^*) = \big(\mathrm{h}_1^*(\boldsymbol{\eta}^*),\ldots,\mathrm{h}_l^*(\boldsymbol{\eta}^*),\ldots,\mathrm{h}_k^*(\boldsymbol{\eta}^*)\big)^T$. Define the sample mean

$\underset{\sim}{\bar{\mathrm{T}}}\, n = (\underset{\sim}{\bar{\mathrm{T}}}\, n1,\ldots,\underset{\sim}{\bar{\mathrm{T}}}\, nk)$, where

$$\underset{\sim}{\bar{\mathrm{T}}}\, n1 = \tfrac{1}{n}\sum_{i=1}^{n} \mathrm{T}_{nj}(\mathrm{X}_i) \qquad j = 1,\ldots,k.$$

The $\sigma_{jj'}^*$ are the elements of the covariance matrix of $\underset{\sim}{\bar{\mathrm{T}}}\, n$ based on the length-biased distribution and are given by

$$\sigma_{jj'}^* = \mathrm{E}_g[\bar{\mathrm{T}}_{nj} - \eta_j^*][\bar{\mathrm{T}}_{nj'} - \eta_{j'}^*] \qquad j, j' = 1,\ldots,k$$

$$= \mathrm{E}_g[\bar{\mathrm{T}}_{nj} - \eta_j - \zeta'_j(\boldsymbol{\theta})\,][\bar{\mathrm{T}}_{nj'} - \eta_{j'} - \zeta'_{j'}(\boldsymbol{\theta})\,]$$

$$= \sigma_{jj'} - \left(\zeta'(\boldsymbol{\theta})\zeta'(\boldsymbol{\theta})^T\right)_{jj'}$$

where $\zeta'_j(\boldsymbol{\theta})$ and $\left(\zeta'(\boldsymbol{\theta})\zeta'(\boldsymbol{\theta})^T\right)_{jj'}$ are the j^{th} and $jj'^{\ th}$ elements of $\zeta'(\boldsymbol{\theta})$ and $\left(\zeta'(\boldsymbol{\theta})\zeta'(\boldsymbol{\theta})^T\right)$ respectively and $\sigma_{jj'}$ are the jj' elements of the covariance matrix of $\underset{\sim}{\bar{\mathrm{T}}}_n$ under the original probability distribution f.

$$a_{ij}^{(l)*} = \frac{\partial^2 \mathrm{h}_l^*(\boldsymbol{\eta}^*)}{\partial\eta_j^*\partial\eta_i^*}$$

Recall that by definition $\mathrm{H}^*(\boldsymbol{\eta}^*) = \boldsymbol{\theta}$, and $\mathrm{H}(\boldsymbol{\eta}) = \boldsymbol{\theta}$, where $\boldsymbol{\eta}^* = \boldsymbol{\eta} + \zeta'(\boldsymbol{\theta})$ so that if we let $\zeta_{jj'} = \left(\zeta'(\boldsymbol{\theta})\zeta'(\boldsymbol{\theta})^T\right)_{jj'}$ we observe that

$$\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}^{(l)*}\sigma_{ji}^* = \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}^{(l)*}\left(\sigma_{ji} - \zeta_{ji}\right).$$

This shows the leading term in the bias of the length-biased distribution involves both the bias from the original distribution and the bias induced by the length-biased design. Next we compare the information about $\boldsymbol{\theta}$ provided by the estimates $\hat{\boldsymbol{\theta}}_n$ from the samples drawn from f(X;$\boldsymbol{\theta}$) and g(X;$\boldsymbol{\theta}$).

The goal here is to investigate theoretically which of the two designs provide greater information about the unkown parameters. Fisher information is be used to compare the efficiency for the parameters in two design models under which samples are obtained by comparing their respective optimality. Optimal design criteria is defined by a function Φ_p defined on the matrix C, where Φ_p satisfies the

following conditions.

i) Φ_p is nondecreasing and invariant with respect to row and colmn permutations.

ii) If $C_1 \geq C_2$, i.e. $C_1 - C_2$ is non negative definite, then

$$\Phi(C_1) \leq \Phi(C_2)$$

Keifer (1975) defines some of the widely used optimality criteria as;

$$\Phi_p(C) = (\frac{1}{q\text{-}1} \sum \lambda^{-p})^{\frac{1}{p}},$$

where $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{q-1}$ are the eigenvalues of matrix C.

Silvey, 1980; Shah, 1989 discussed optimality criteria measures applicable in the comparison of any two information matrices C_1 and C_2. If $C_1 - C_2$ is nonnegative definite matrix then $\Phi_p(C_1) \geq \Phi_p(C_2)$ where $\Phi_p(C)$ is a measure of the amount of information in C. The three ψ measures proposed are for values of $p = 0$, 1, and ∞ , in Φ_p corresponding to D-optimality, A-optimality and E-optimality respectively. If C is an inverse of the Fisher information matrix for θ, D-optimality minimizes det(C), provides a measure of the generalized variance of the components of $\hat{\theta}$. A-optimality minimizes the trace(C) and represents the average variance of the elements of $\hat{\theta}$. E-optimality minimizes the largest variance of the components of $\hat{\theta}$ and is a measure of the largest variance of the components of $\hat{\theta}$.

From (4.2) we obtain the average Fisher information about θ as

$$\begin{aligned} I_g(\theta) = E[\frac{1}{n}\frac{\partial U_n(\theta)}{\partial \theta^T}] &= -\varphi^*(\theta) \\ &= -\varphi(\theta) + \zeta''(\theta) \\ &= I_f(\theta) + \zeta''(\theta), \qquad \text{finite and p.d.} \end{aligned} \tag{4.5}$$

From (4.5) we have the following implications

$$|\mathbf{I}_g(\boldsymbol{\theta})| = |\mathbf{I}_f(\boldsymbol{\theta}) + \zeta''(\boldsymbol{\theta})| \leq |\mathbf{I}_f(\boldsymbol{\theta})| + |\zeta''(\boldsymbol{\theta})| \tag{4.6}$$

where |A| denotes determinant of A.

So that

$|\mathbf{I}_g(\boldsymbol{\theta})| \leq |\mathbf{I}_f(\boldsymbol{\theta})|$ if $|\zeta''(\boldsymbol{\theta})| \leq 0$, that is, if $\zeta''(\boldsymbol{\theta})$ is negative semi-definite. The equality of the two information matrices will be realized if $\zeta''(\boldsymbol{\theta}) \equiv \mathbf{0}$.

On the other hand if $\zeta''(\boldsymbol{\theta})$ is positive definite then $|\zeta''(\boldsymbol{\theta})| > 0$, and D-optimality criteria is indeterminate.

The trace of (4.5) is

$$\operatorname{tr}(\mathbf{I}_g(\boldsymbol{\theta})) \quad \begin{cases} < \operatorname{tr}(\mathbf{I}_f(\boldsymbol{\theta})) & \text{if } \operatorname{tr}(\zeta''(\boldsymbol{\theta})) < 0, \\ > \operatorname{tr}(\mathbf{I}_f(\boldsymbol{\theta})) & \text{if } \operatorname{tr}(\zeta''(\boldsymbol{\theta})) > 0 \end{cases}$$

which implies that by E-optimality criteria, the length-biased design would be more informative than the original design if $\operatorname{tr}(\zeta''(\boldsymbol{\theta})) > 0$ and otherwise sampling from the original distribution would be more informative about $\boldsymbol{\theta}$.

The following remarks are also noteworthy in making the comparisons.

Remarks

1) If $\zeta(\boldsymbol{\theta})$ is a polynomial in $\boldsymbol{\theta}$, of degree $r < 2$, then $\zeta''(\boldsymbol{\theta}) \equiv \mathbf{0}$, in which case the information about $\boldsymbol{\theta}$ obtained from the length-biased design and the original design is equivalent.

2) For $k = 1$ and $\zeta(\theta)$ is a convex function of θ, then $\zeta''(\theta) > 0$, so that the length-biased design is more informative about θ than the original design.

Example 4.1 Let X_i, $i = 1,\ldots,n$, be i.i.d.r.v.'s following a lognormal distribution(μ,σ^2) as given in example 1 of Chapter 2. It is easy to show that $E_f(X_i) = \exp(\mu + \frac{\sigma^2}{2})$. Recall that $\theta_1 = \frac{\mu}{\sigma^2}$, $\theta_2 = -\frac{1}{2\sigma^2}$ and so that $\zeta(\boldsymbol{\theta}) = \log E_f[X] = -\frac{1}{2\theta_2}(\theta_1 + \frac{1}{2\theta_2})$ and

$$\Psi(\boldsymbol{\theta}) = \tfrac{1}{2}\left(\frac{\theta_1^2}{2\theta_2} - \log\frac{-1}{2\theta_2}\right).$$

This implies that

$$\varphi(\boldsymbol{\theta}) = \tfrac{1}{2}\begin{bmatrix} \frac{1}{\theta_2} & -\frac{\theta_1}{\theta_2^2} \\ -\frac{\theta_1}{\theta_2^2} & \frac{\theta_1^2}{\theta_2^3} + \frac{1}{\theta_2^2} \end{bmatrix} \text{ and } \underset{\sim}{\zeta}''(\boldsymbol{\theta}) = \tfrac{1}{2}\begin{bmatrix} 0 & \frac{1}{\theta_2^2} \\ \frac{1}{\theta_2^2} & -\frac{2\theta_1}{\theta_2^3} - \frac{1}{\theta_2^3} \end{bmatrix}. \tag{4.7}$$

Since $|\underset{\sim}{\zeta}''(\boldsymbol{\theta})| = -\frac{1}{4\theta_2^4} < 0$, by D-optimality we conclude that length-biased sampling design is less informative than sampling from the original original distribution for the lognormal probability model, in the sense that the generalized variance from the length-biased design is greater than the generalized variance from sampling from the original distribution.

The trace of $\underset{\sim}{\zeta}''(\boldsymbol{\theta})$ is $\mathrm{tr}(\underset{\sim}{\zeta}''(\boldsymbol{\theta})) = -(\frac{\theta_1}{\theta_2^3} + \frac{1}{2\theta_2^3}) > 0$, since $\theta_1 > 0$ and $\theta_2 < 0$, so that by A-optimality the length-biased design would be desirable since it provides the minimum average variance of the components of $\widehat{\boldsymbol{\theta}}$.

4.4. Estimating Regression Parameters from the Length-Biased Distribution

Let X_i $(i = 1,\ldots,n)$ be a r.v. following a pdf $g(X_i;\beta,\phi)$, which is a length-biased version of the distribution of $f(X_i;\beta,\phi)$ in (3.2), given by

$$g(X_i;\beta,\phi) = \frac{X_i\, f(X_i;\beta,\phi)}{\gamma(\mathbf{b}(z_i,\beta,\phi))} = c^*(X_i)\exp\Big\{ \mathbf{b}(z_i,\beta,\phi))^T \underset{\sim}{T}(X_i) + \Psi^*(\mathbf{b}(z_i,\beta,\phi)) \Big\} \ , \ X_i > 0$$

where $\Psi^*(\mathbf{b}(z_i,\beta,\phi)) = \Psi(\mathbf{b}(z_i,\beta,\phi)) - \log\, \gamma(\mathbf{b}(z_i,\beta,\phi))$ and
$\gamma(\mathbf{b}(z_i,\beta,\phi)) = E_f[X_i] < \infty$, where E_f denotes expectation with respect to the density f(.).
Recall that $\underset{\sim}{T}(X_i) = (T_1(X_i),\ldots,T_q(X_i))^T$, so that we may find a one-to-one transformation h, such that

$$h(T_j(X_i)) = X_i \text{ and } E_f|h(T_j(X_i)| < \infty, \text{ for some } 1 \le j \le k$$

This is certainly the case with the lognormal distribution with $T_1(X_i) = \log(X_i)$, and $b_1(z_i,\beta,\sigma^2) = z_i^T\beta\ /\sigma^2$.

As before we derive the moments of $\underset{\sim}{T}(X_i)$ from length-biased density as

$$E_g[\underset{\sim}{T}(X_i)] = -\psi^*(\mathbf{b}(z_i,\beta^0)).$$

The second central moment of $\underset{\sim}{T}(X_i)$ is given by

$$E_g[\underset{\sim}{T}(X_i) - \psi^*(b(z_i,\beta^0))][\underset{\sim}{T}(X_i)] - \psi^*(b(z_i,\beta^0))]^T$$

$$= -\varphi^*(b(z_i,\beta^0)).$$

Expressing the moments of $\underset{\sim}{T}(X_i)$ in terms of the moments of the original density we have

$$E_g[\underset{\sim}{T}(X_i)] = -\left\{ \psi(b(z_i,\beta^0)) - \frac{\gamma'(b(z_i,\beta^0))}{\gamma(b(z_i,\beta^0))} \right\}.$$

As an illustration if $q = 1$, $T_1(X_i) = X_i$. (The gamma distribution $G(\lambda,\nu)$ with ν known and $\gamma = z^T\beta$ would fit this case).

In this case,

$$E_f[X_i] = -\psi(b(z_i,\beta^0)) = \gamma(b(z_i,\beta^0)$$

and

$$E_g[X_i] = -\left\{ \psi(b(z_i,\beta^0)) + \frac{\varphi(b(z_i,\beta^0))}{\psi(b(z_i,\beta^0))} \right\}$$

$$= E_f[X_i] + \frac{Var_f(X_i)}{E_f[X_i]}.$$

This implies the sample mean from the length-biased distribution would be a biased estimate of the population mean of the true distribution. To obtain the MLE of β, and ϕ, of the length-biased data from $g(X_i;\beta,\phi)$ we maximize the log-likelihood function in the same fashion as before. So the MLE $\hat{\beta}_n$, and $\hat{\phi}_n$ of

β, and ϕ, are the solutions to the equations

$$\mathrm{U}_n(\beta^0) = \sum_{i=1}^{n} (\mathbf{b}'(\mathbf{z}_i,\beta^0))^T \left[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \psi^*(\mathbf{b}(\mathbf{z}_i,\beta^0)) \right] = \mathbf{0}. \tag{4.8}$$

Note that the estimating equation is not a sum of identically distributed r.v.'s, however the terms are independent, hence we can still use some large sample results to establish the asymptotic normality of the MLE. The estimating equations (4.8) are similar to the estimating equations (3.12) and by similar arguments leading to (3.25) we conclude $\mathrm{U}_n(\beta^0)$ in (4.7) has an asymptotic normal distribution

$$\tfrac{1}{\sqrt{n}}\, \mathrm{U}_n(\beta^0) \longrightarrow^{\mathfrak{D}} \mathrm{N}\left(0,\, \mathrm{I}_g(\beta^0)\right), \tag{4.9}$$

where $\mathrm{I}_g(\beta^0)$ is the limiting average Fisher information from the length-biased likelihood given below. Also in line with (3.29) we have the following result,

$$\sqrt{n}(\widehat{\beta}_n^0 - \beta^0) \longrightarrow^{\mathfrak{D}} \mathrm{N}\left(0,\; \mathrm{I}_g^{-1}(\beta^0)\right) \tag{4.10}$$

The gradient of the estimating equation $\mathrm{U}_n(\beta^0)$ is given by

$$\frac{\partial \mathrm{U}_n(\beta^0)}{\partial \beta^{0^T}} = \sum_{i=1}^{n} (\mathbf{b}''(\mathbf{z}_i,\beta^0))^T \left[\underset{\sim}{\mathrm{T}}(\mathrm{X}_i) + \psi^*(\mathbf{b}(\mathbf{z}_i,\beta^0))\right]$$

$$+ \sum_{i=1}^{n} (\mathbf{b}'(\mathbf{z}_i,\beta^0))^T\, \varphi^*(\mathbf{b}(\mathbf{z}_i,\beta^0))]^T \mathbf{b}'(\mathbf{z}_i,\beta^0)).$$

For every $i = 1, \ldots, n$ we define

$$I_i(\beta^0) = E_g[\frac{\partial \log g(X_i;\beta^0)}{\partial \beta^0}][\frac{\partial \log g(X_i;\beta^0)}{\partial \beta^0}]^T$$

$$= -(b'(z_i,\beta^0))^T \varphi^*(b(z_i,\beta^0))\, b'(z_i,\beta^0),$$

so that, as $n \longrightarrow \infty$ the average Fisher information is

$$\bar{I}_{g(\beta^0)} = -\frac{1}{n}E[\frac{\partial U_n(\beta^0)}{\partial \beta^{0T}}]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}(b'(z_i,\beta^0))^T \varphi^*(b(z_i,\beta^0))\, b'(z_i,\beta^0).$$

The average Fisher information matrix for β^0, from the length-biased distribution can be expressed in terms of the Fisher information for β^0 from the original distribution by substituting in the expression for $\varphi^*(b(z_i,\beta^0))$, to obtain

$$\bar{I}_{g(\beta^0)} = \bar{I}_{f(\beta^0)} + \zeta''(\beta^0),$$

where $\zeta''(\beta^0) = \frac{1}{n}\sum_{i=1}^{n}(b'(z_i,\beta^0))^T\{\frac{\gamma''(b(z_i,\beta^0))}{\gamma(b(z_i,\beta^0))} - \frac{\gamma'(b(z_i,\beta^0))\,\gamma'(b(z_i,\beta^0))^T}{(\gamma(b(z_i,\beta^0)))^2}\}\, b'(z_i,\beta^0))$.

so that the Fisher information from the length-biased distribution will be less than or greater or equal to the Fisher information depending on the nature of $\log\gamma(b(z_i,\beta^0))$.

Remarks:

(1). If $\log\gamma(b(z_i,\beta^0))$ is a polynomial of degree r, $(r \leq 1)$, in $b(z_i,\beta^0)$, then $\zeta''(\beta^0)$

vanishes. In this case we get the same amount of information from the two probability models.

(2). If $q = 1$, and $\log\gamma(b(z_i,\beta^0))$ is a convex function of $b(z_i,\beta^0)$ then the length-biased design is more informative about β^0 than the original design.

Chapter 5

Example and Simulations

5.1 Introduction

The previous chapters studied the asymptotic bias reduction and optimality of the designs. Since the asymptotics tells us how the estimates behave in large samples, it is important to study also how well the proposed estimators perform in small-to-moderate sample sizes. In this chapter we study applications to a real data problem which is of length-biased design, and also study the perfomance of our estimators in moderate sample sizes.

The data set for analysis in this exercise is from the 1989/1990 Sudan Demograhic and Health Surveys (DHS). Our objective is to estimate the mean birth interval between the first and second births among married women, 14-49 years old, who have been married for at least 10 years. Without loss of generality we assume the women constitutes a cohort married at the same time ($t = 0$) and with complete birth intervals, that is, both the first and second births dates are available. The data then consist of $N = 3211$ complete birth intervals. In addition to the birth intervals, the covariates age, duration of marriage use for the women were also extracted for the regression analysis.

In order to obtain a length-biased sample of all the complete birth intervals, we select the birth intervals straddling the 5-year period (survey time, $\tau = 5$) for the sample. Such a sampling scheme is known to induce a length-biased sample

since longer birth intervals are more likely to overlap any arbitrarily chosen time, and hence be sampled, than shorter birth intervals. The length-biased sample selected consists of $n = 681$ birth intervals. From the 681 birth intervals we obtained $M = 1000$ sub-samples of size $n = 50$, 100 and 200 to study and compare the performance of the jackknife estimates and the MLE. The estimates are the averages of the 1000 estimates from each of the $n = 50$, 100 and 200 sample sizes.

We assume the population of $N = 3211$ birth intervals follow a lognormal distribution with the baseline survival function given by

$$S_W(w) = 1 - \Phi((w-\mu)/\sigma), \tag{5.1}$$

where w is the natural logarithm of the birth intervals, μ and σ are the location and shape parameters, respectively. Φ is the cumulative distibution function of the normal distribution.

The lognormal regression model with the covariates age (months) and duration of marriage (months) is fit to the data to estimate biases and the estimation of the effects of the covariate terms. The model assumed is

$$W = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \sigma\epsilon, \tag{5.2}$$

where

W is the natural logarithm of the birth interval,

β_0, β_1 and β_2 are the regression coefficients,

σ is scale parameter,

Z_1 and Z_2 denote Age and Duration of marriage for the women, respectively, and

ϵ is the error term.

The baseline survival function is of the form (5.1) with $\mu = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2$

For the model (5.1) without covariates, we denote the parameter vector by $\boldsymbol{\theta} = (\mu, \sigma)$ and its MLE's and jackknife estimates are denoted by $\hat{\boldsymbol{\theta}}_l$ and $\hat{\boldsymbol{\theta}}_l^J$, respectively, $l = 1, \ldots, M$. The sample mean and variance of $\boldsymbol{\theta}$ computed from all the M replicate samples are

$$\hat{\bar{\boldsymbol{\theta}}} = \frac{1}{M} \sum_{l=1}^{M} \hat{\boldsymbol{\theta}}_l$$

and

$$\hat{\hat{\mathbf{V}}} = \frac{1}{M-1} \sum_{l=1}^{M} (\hat{\boldsymbol{\theta}}_l - \hat{\bar{\boldsymbol{\theta}}})(\hat{\boldsymbol{\theta}}_l - \hat{\bar{\boldsymbol{\theta}}})^T.$$

The standard error of the components of $\hat{\boldsymbol{\theta}}$ is estimated as the square root of the corresponding diagonal entry of $\hat{\hat{\mathbf{V}}}$.

The overall performance of our estimators is shown by the root mean squared error (RMSE). The RMSE measures the combined effect of bias and sampling variation, and for the parameter μ the expected value of the MSE is equal to:

$$E(\hat{\mu} - \mu)^2 = \sigma^2(\hat{\mu}) + (E(\hat{\mu}) - \mu)^2. \tag{5.3}$$

The RMSE for the estimate $\hat{\mu}$ of μ is estimated by

$$\text{RMSE} = \sqrt{\frac{1}{1000} \sum_{l=1}^{1000} (\hat{\mu}_l - \mu)^2}. \tag{5.4}$$

The RMSE of $\hat{\sigma}^2$ is defined similarly, with $\hat{\mu}$ in (5.4) replaced by $\hat{\sigma}^2$.

In addition to the parameter estimates, the distribution of the M estimates from the replicate samples is provided to support our methodological claim that the jackknifed estimates and MLE are asymptotically normal. The distributions are presented graphically for the parameters (μ and σ) by way of both the smoothed density function obtained from the empirical distribution of the estimates.

5.2 Results and Discussion

The parameters for the model (5.1) for the $N = 681$ birth intervals are $\mu = 3.5984$ and $\sigma = 0.5550$. The results of our estimates from fitting model (5.3) to the Sudan birth interval data are given in Table 5.1.

Table 5.1
Mean, Bias and RMSE of the MLE and Jackknife Estimates
for M = 1000 samples of size n from a Lognormal Distribution Model.
1990 Sudan DHS birth interval data.

	MLE						Jackknife Estimates					
n	$\widehat{\mu}$	Bias	RMSE	$\widehat{\sigma}$	Bias	RMSE	$\widehat{\mu}$	Bias	RMSE	σ	Bias	RMSE
50	3.6005	0.0021	0.4179	0.5458	-0.0092	0.0720	3.6005	0.0021	0.4179	0.5546	-0.0004	0.0782
100	3.5986	0.0002	0.4122	0.5518	-0.0032	0.0632	3.5986	0.0002	0.4122	0.5562	0.0012	0.0669
200	3.5989	0.0005	0.4103	0.5540	-0.0010	0.0570	3.5989	0.0005	0.4103	0.5562	0.0012	0.0590
681	3.5984	_		0.5550	_							

Discussion

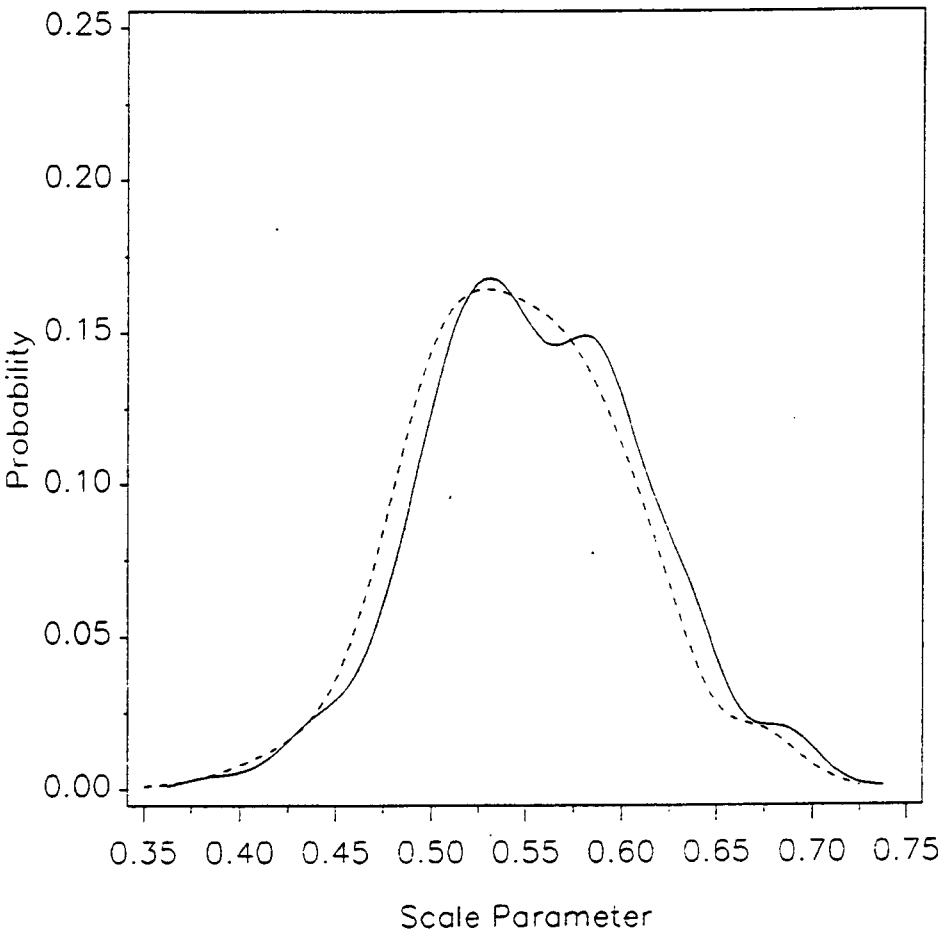
Both the MLE and jackknifed estimates of the location parameter μ were identical suggesting the leading term in the bias of MLE for μ is zero, hence the classical jackknifing did not improve the MLE. The jackknife estimate for the scale parameter, for each of the sample sizes 50, 100, and 200 is less biased for the length-biased estimate of the $n = 681$ compared to the corresponding MLE estimate. The MLE of the scale parameter performed better than the jackknife estimate in terms of a smaller RMSE, however, the difference in the RMSE of the two estimates is small. This lack of gain in the jackknifed estimator over the MLE for the location parameter, in this example, of the lognormal model (5.1), is be explained by the lack of contribution of the leading term in the bias of the MLE. Recall that the jackknifed estimator reduces the bias of order n^{-1} from the bias of the MLE and for large samples this reduction is minimal. The leading term in the bias of the MLE $\widehat{\boldsymbol{\theta}}_n$ is a function $\boldsymbol{a}(\boldsymbol{\theta})$ whose components are linear combinations of the third partial derivatives of the loglikelihood. In the case of the model (5.1) the coordinate of $\boldsymbol{a}(\boldsymbol{\theta})$ for μ vanishes asymptotically, resulting in no gain in the bias reduction of the Jackknife estimator. To see this, recall from Chapter 4 that the first element of $\Psi(\boldsymbol{\theta})$ is θ_1^2/θ_2, so that the $(1,1)^{th}$ element of $\boldsymbol{\Gamma}(\boldsymbol{\theta})$ is zero.

Figure 5.1 shows the distribution of the 1000 estimates of the scale parameter for the samples of size $n = 100$. The smoothed density plots shows a good approximation to the normal distribution for the sample size of $n = 100$. The distribution of the shape parameter μ for the same samples (not provided here) also exhibit a good approximation to the normal distribution. The same result, of no difference in the MLE and the jackknifed estimate for the location parameter is

also shown by the coincidence of the density plots for the two estimates of μ. The distribution of the estimates from the sample sizes $n(=50,200)$ also showed a good approximation to the normal distribution.

Figure 5.1

Maximum Likelihood and Jackknifed Estimates
of the Scale Parameter of the Lognormal Distribution
for the SUDAN First Order Birth Interval Data
SUDAN DHS (1990)



— Jackknife Estimate ----- Maximum Likelihood Estimate

Regression Model

The lognormal regression model with covariates, duration of marriage and age of the woman, is fit to the length-biased sample from the Sudan birth interval data. The length-biased sample consists of n = 681 birth intervals. The population regression coefficients are $\beta_0 = 3.8228$, $\beta_1 = -0.0006$, $\beta_2 = -0.0002$ and $\sigma = 0.5514$, where β_0, β_1 and β_2 are the intercept duration and age coefficients respectively. From the n=681 we obtained M = 1000 samples of size n=50, 100 and 200 to study and compare the performance of our the jackknife estimator and the MLE. The estimates are the averages from the 1000 samples for each sample size. The parameter vector for the regression model is denoted by $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \sigma)$ and it's corresponding MLE and jackknife estimates are obtained in the same fashion as discussed above for estimating the parameter $\boldsymbol{\theta}$.

The results are shown in Table 5.2. The largest bias reduction is observed for the scale parameter for all sample sizes, where the MLE have a negative bias and the jackknifed estimates are biased upwards. The RMSE's from the MLE and the jackknifed MLE for the intercept term and the regression coefficients are comparable, whereas the MLE performs better than the jackknifed MLE for the scale parameter in term of smaller RMSE.

Table 5.2

Mean, Bias and RMSE of the MLE and Jackknife Estimates
of estimates of mean birth interval from the Lognormal Regression Model,
1990 Sudan DHS birth interval data.
M = 1000 samples of size n = 50, 100, 200

		MLE			Jackknife Estimates		
n		Mean	Bias	RMSE	Mean	Bias	RMSE
50	β_0	3.8083	-0.0145	0.7141	3.8037	-0.0191	0.0798
	β_1	-0.0007	-0.0001	0.0022	-0.0007	-0.0001	0.0022
	β_2	-0.0001	0.0001	0.0020	-0.0001	0.0001	0.0020
	σ	0.5321	-0.0193	0.0658	0.5543	0.0029	0.0787
100	β_0	3.8211	-0.0017	0.6147	3.8176	-0.0052	0.6117
	β_1	-0.0006	0.0000	0.0016	-0.0006	0.0000	0.0016
	β_2	-0.0002	0.0000	0.0013	-0.0001	0.0001	0.0013
	σ	0.5443	-0.0071	0.0579	0.5543	0.0029	0.0659
200	β_0	3.8233	0.0005	0.5651	3.8254	0.0026	0.5630
	β_1	-0.0006	0.0000	0.0009	-0.0006	0.0000	0.0009
	β_2	-0.0002	0.0000	0.0087	-0.0002	0.0001	0.0013
	σ	0.5484	-0.0030	0.0529	0.5534	0.0020	0.0574

$N = 681$ with true parameters $\beta_0 = 3.8228$, $\beta_1 = -0.0006$, $\beta_2 = -0.0002$ and $\sigma = 0.5514$

The density estimates of the estimates for the M = 1000 samples of size $n = 50$ are shown in Figures 5.2, 5.3, 5.4 and 5.5 for the intercept term, Age and Duration regression coefficients and the scale parameters, respectively. The graphs illustrates similar approximately normal distributions for both the MLE and the Jackknife estimates. The graphs of the estimates from the other sample sizes (not shown here) showed similar patterns.

Figure 5.2

Maximum Likelihood and jackknifed Estimates of the
Intercept Parameter of the Lognormal Regression Model
for the SUDAAN First Order Birth Interval Data
SUDAAN DHS (1990)

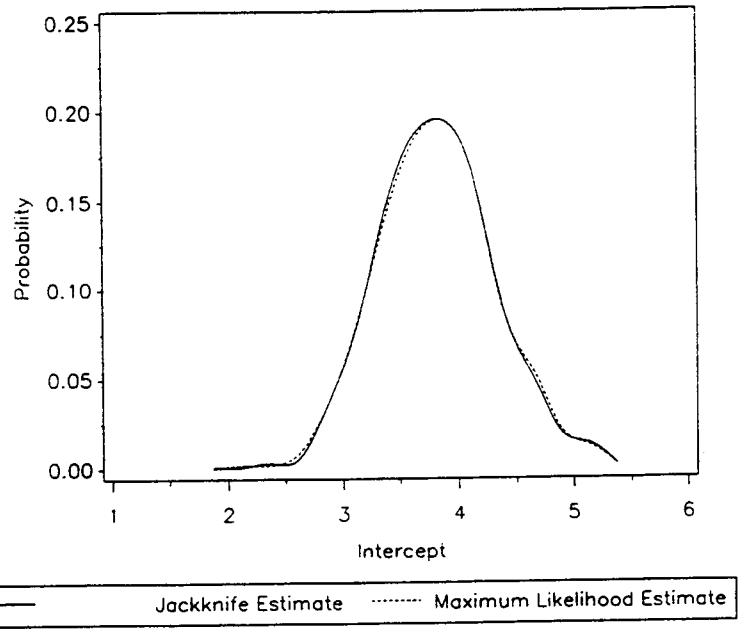


Figure 5.3

Maximum Likelihood and jackknifed Estimates of the
Marital Duration Parameter of the Lognormal Regression Model
for the SUDAAN First Order Birth Interval Data
SUDAAN DHS (1990)

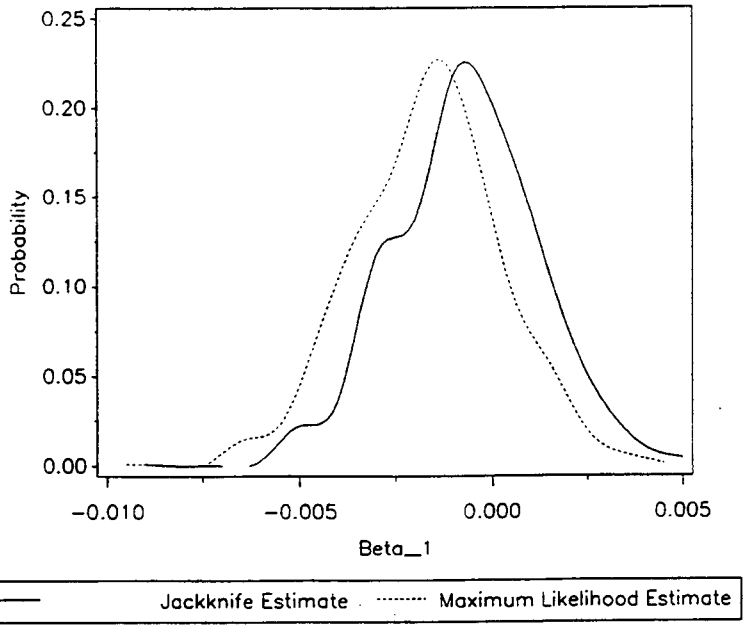


Figure 5.4

Maximum Likelihood and jackknifed Estimates of the
Age Parameter of the Lognormal Regression Model
for the SUDAAN First Order Birth Interval Data
SUDAAN DHS (1990)

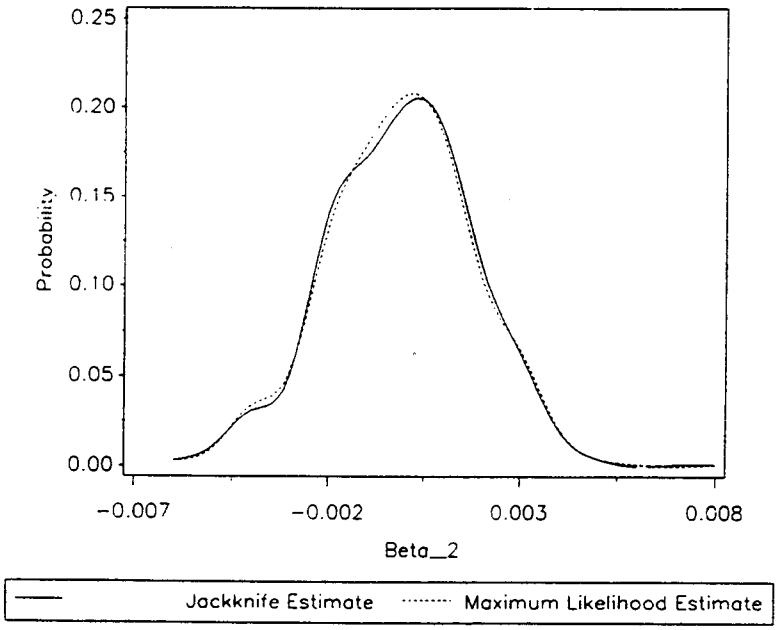
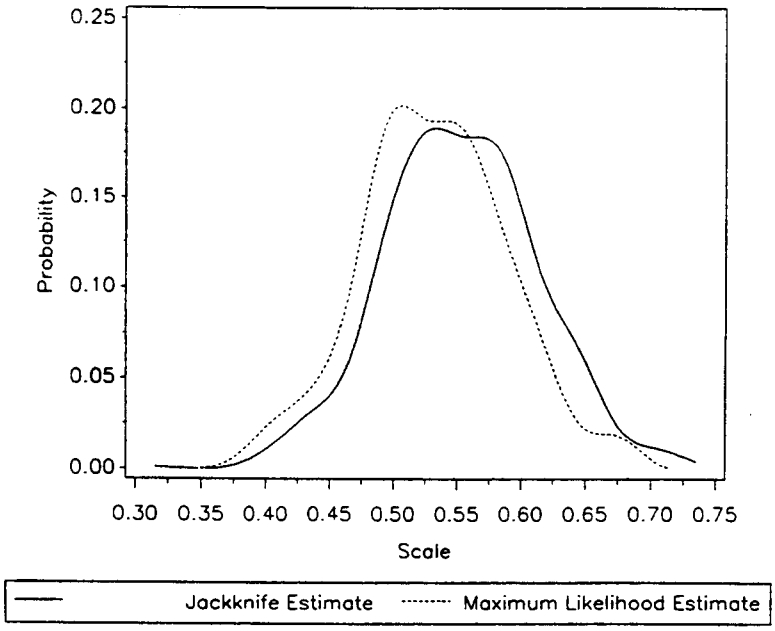


Figure 5.5

Maximum Likelihood and jackknifed Estimates of the
Scale Parameter of the Lognormal Regression Model
for the SUDAAN First Order Birth Interval Data
SUDAAN DHS (1990)



5.3 Simulation Study

The performance of our estimates in moderate samples is studied through simulation studies. The main focus of our study is on the performance of the jackknife estimates compared to the MLE. Two distributions for the response variate are chosen for this study, and these are the lognormal and the gamma. For each distribution we consider estimating the parameters of the model without the covariate terms. The regression analysis is studied through simulation study where the response variable follows a lognormal distribution. In each case our simulations involve m=1000 data sets, each containing samples of size $n = 50, 100, \text{and } 200$.

Our first study concerns the estimation of the parameters μ and σ and obtains their root mean squared errors. In particular, we obtain samples from the following distributions:

(1) the lognormal, with location parameter $\mu = 0.5$ and scale parameter $\sigma = 1.5$; and

(2) the gamma with location parameter $\mu = 1.5$ and $\sigma^2 = 1.5$.

The baseline survival functions for the lognormal and gamma distribution models with parameters shape parameter μ and scale parameter σ^2 are given by:

(1) Lognormal:

$$S_W(w) = 1 - \Phi((w-\mu)/\sigma); \text{ and} \tag{5.5}$$

(2) Gamma:

$$S_X(x) = \int_x^{\infty} \frac{\sigma}{\Gamma(\mu)} (\sigma t)^{\mu - 1} \exp\{-\sigma t\} dt, \tag{5.6}$$

where w is the natural logarithm of the response variate and Φ is the cumulative distribution function of the normal distribution.

It easy to show that the length-biased version of the original distribution from the lognormal and the gamma distributions preserves their original distributional forms. The length-biased distribution with the original form being the lognormal with parameters μ and σ^2 is itself lognormal with the same parameter σ^2 and a shape parameter $\mu^* = \mu + \sigma^2$. For the gamma distribution the shape parameter for the length-biased density becomes $\mu^* = \mu + 1$, where μ is the shape parameter of the original distribution gamma and the scale parameter σ remains unchanged.

The lognormal random variate with mean $e^{\{\mu + \sigma^2/2\}}$ and variance $e^{\{2\mu + \sigma^2\}}(e^{\sigma^2} - 1)$ is generated using the SAS system's function NORMAL. The NORMAL function returns a random variate generated from a normal distribution with mean zero and variance 1. To generate the random variate following the gamma distribution with the shape parameter μ and scale parameter σ, we use the SAS RANGAM function. The RANGAM function returns a random variate distributed according to the gamma function with the specified parameter values.

The first simulation samples were drawn from the lognormal model with $\mu = 2.0$ and $\sigma^2 = 1.5$. The simulations from the gamma were obtained with $\mu = 1.5$ and $\sigma^2 = 1.5$. For each distribution $M = 1000$ samples of sizes $n = 50$, 100 and 200 were obtained. The MLE and the jackknifed estimates are shown in Tables 5.3 and 5.4 for the Lognormal and Gamma distributions, respectively.

Table 5.3
Mean, Bias and RMSE of the MLE and Jackknife Estimates
for M = 1000 samples of size n from a Lognormal Distribution.
Simulations with $\mu = 2.0$ and $\sigma^2 = 1.5$

	MLE						Jackknife Estimates					
n	$\widehat{\mu}$	Bias	RMSE	$\widehat{\sigma}$	Bias	RMSE	$\widehat{\mu}$	Bias	RMSE	σ	Bias	RMSE
50	1.9997	-0.0003	0.1795	1.2013	-0.0023	0.3220	1.9997	-0.0003	0.1795	1.2198	-0.0049	0.3057
100	1.9993	-0.0007	0.1244	1.2122	-0.0013	0.3000	1.9993	-0.0007	0.1244	1.2213	0.0034	0.2914
200	1.9983	-0.0017	0.0849	1.2214	-0.0033	0.2864	1.9983	-0.0017	0.0849	1.2260	0.0013	0.2809

The simulation results from the lognormal distribution confirm the findings in the real data example that there is no significant bias reduction by jackknifing, for the MLE of the shape parameter. However jackknifing had a positive result in reducing the bias of the MLE for the scale parameter.

Table 5.4
Mean, Bias and RMSE of the MLE and Jackknife Estimates
for M = 1000 samples of size n from a Gamma Distribution.
Simulations with $\mu = 1.5$ and $\sigma^2 = 1.5$

	MLE						Jackknife Estimates					
n	$\widehat{\mu}$	Bias	RMSE	$\widehat{\sigma}$	Bias	RMSE	$\widehat{\mu}$	Bias	RMSE	σ	Bias	RMSE
50	1.1601	-0.3399	0.3640	0.8703	-0.3544	0.3437	1.1668	-0.3332	0.3577	0.8853	-0.3394	0.3538
100	1.1654	-0.3328	0.3475	0.8805	-0.3459	0.3368	1.1688	0.3295	0.3443	0.8880	-0.3383	0.3441
200	1.1654	-0.3346	0.3407	0.8816	-0.3431	0.3380	1.1670	-0.3330	0.3391	0.8853	-0.3394	0.3428

The simulations from the gamma probability distribution indicate a sizable reduction in the bias of the MLE by jackknifing. The largest bias reduction by

jackknifing is observed for the scale parameter. The RMSE of the jackknifed estimate is slightly larger than the RMSE of the MLE of the scale parameter. So the bias reduction by jackknifing the MLE of the scale parameter is archieved at the expense of having a larger RMSE.

Next we consider fitting the regression model to each of the model lognormal model (5.5). The covariates Z_1 and Z_2 are generated from bivariate normal with parameters $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 1$, $\rho = 0.35$. The simulation study from the lognormal distributions should provide us with a typical applications for the proposed estimates of the length-biased data from exponential family of distributions.

For the linear regression model we assume

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_1 + \sigma\epsilon, \tag{5.8}$$

where ϵ have p.d.f. f(ϵ), β_j $(j = 0,1,2)$ are unkown regression coefficients to be estimated together with the scale parameter σ.

The sample sizes used in the simulations are n=50, 100, and 200. In each case we generate 1,000 samples of size n observations and compute $\widehat{\beta}_j$. The root mean squared error, is the criterion used to asses the performance of the estimates and is given by

$$\text{RMSE} = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\widehat{\beta}_{ji} - \beta_j)^2} \qquad (j = 0,\ 1,\ 2)$$

for the regression model parameter estimates. The results are shown in Table 5.5. The MLE and the jackknifed estimate for the intercept term are similar.

Jackknifing improved the bias of the MLE of the scale parameter. With respect to the RMSE the jackknifed estimates performed better than the MLE of the scale parameter, whereas for the intercept and regression coefficients the RMSE for the MLE and the jackknifed estimate are comparable.

Table 5.5

Mean, Bias and RMSE of the MLE and Jackknife Estimates of the Parameters of the Lognormal Regression Model, M = 1000 simulations of samples of size n = 50, 100, 200

		MLE			Jackknife Estimates		
n		Mean	Bias	RMSE	Mean	Bias	RMSE
50	β_0	2.2496	0.7405	0.7698	2.2496	0.7405	0.7698
	β_1	1.4921	-0.0079	0.1853	1.4915	-0.0085	0.1870
	β_2	1.8041	0.0041	0.2027	1.8041	0.0041	0.2031
	σ	1.1685	-0.0563	0.1394	1.2125	-0.0122	0.1329
100	β_0	2.2506	0.7506	0.7605	2.2506	0.7506	0.7605
	β_1	1.4936	-0.0064	0.1289	1.4933	-0.0067	0.1291
	β_2	1.8048	0.0048	0.1398	1.8047	0.0047	0.1398
	σ	1.2005	-0.0247	0.0913	1.2222	-0.0025	0.0896
200	β_0	2.2502	0.7502	0.7546	2.2502	0.7502	0.7546
	β_1	1.4919	-0.0081	0.0897	1.4918	-0.0082	0.0896
	β_2	1.8053	0.0053	0.0955	1.8052	0.0052	0.0954
	σ	1.2092	-0.0155	0.0637	1.2201	-0.0046	0.0625

True parameters $\beta_0 = 1.5$, $\beta_1 = 1.5$, $\beta_2 = 1.8$ and $\sigma^2 = 1.5$

The distribution of the parameter estimates from the $M = 1000$ samples of size $n = 50$ are shown in Figures 5.6, 5.7, 5.8 and 5.9 for the intercept term, X1 and X2 regression coefficients and the scale parameters, respectively. The graphs suggest that the distribution of both the MLE and the jackknifed MLE are approximately normal. Figure 5.9 further illustrate the shift in the scale distribution resulting from a sizable bias reduction by jackknifing the MLE of the scale parameter.

Figure 5.6

Maximum Likelihood and Jackknifed Estimates of the
Intercept Parameter of the Lognormal Regression Model
The mean estimates from M=1000 samples of size n=50

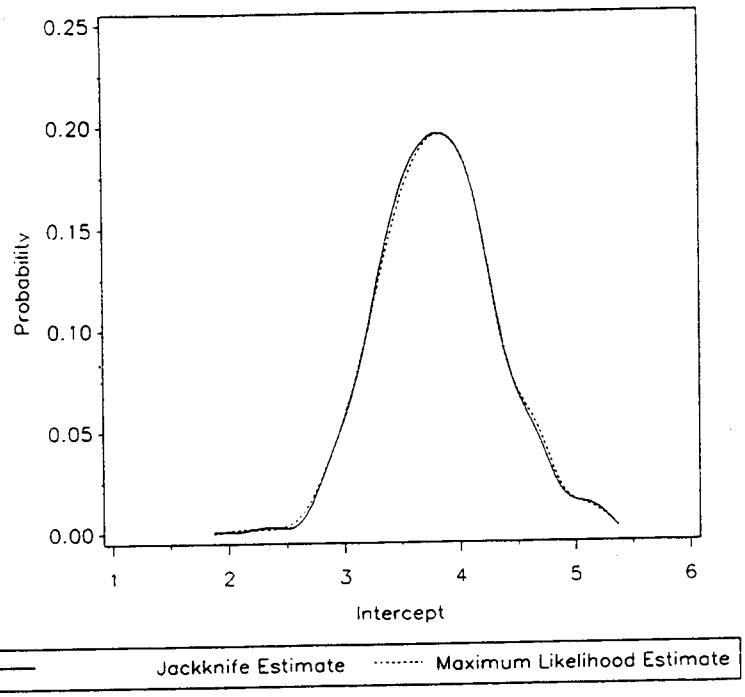


Figure 5.7

Maximum Likelihood and Jackknifed Estimates of the
Beta_1 Parameter of the Lognormal Regression Model
The mean estimates from M=1000 samples of size n=50

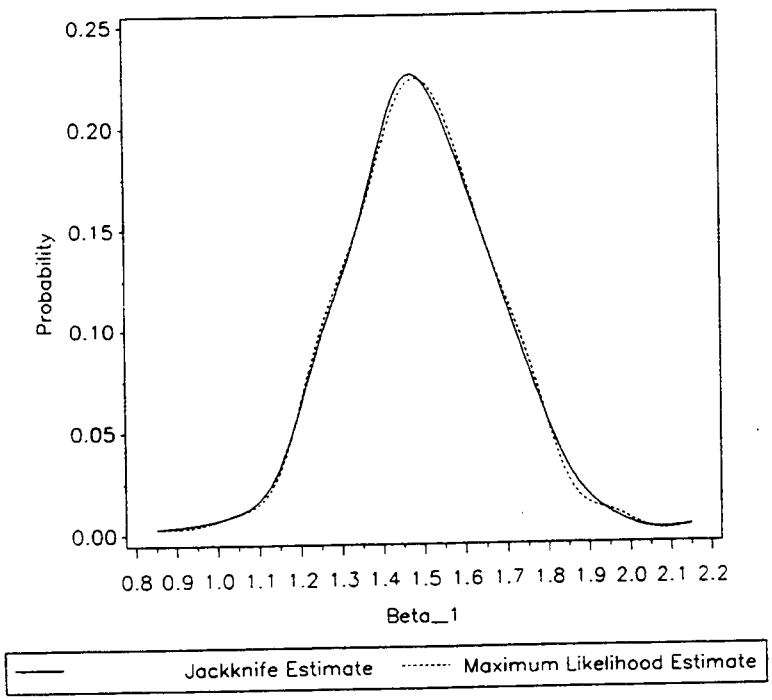


Figure 5.8

Maximum Likelihood and Jackknifed Estimates of the
Beta_2 Parameter of the Lognormal Regression Model
The mean estimates from M=1000 samples of size n=50

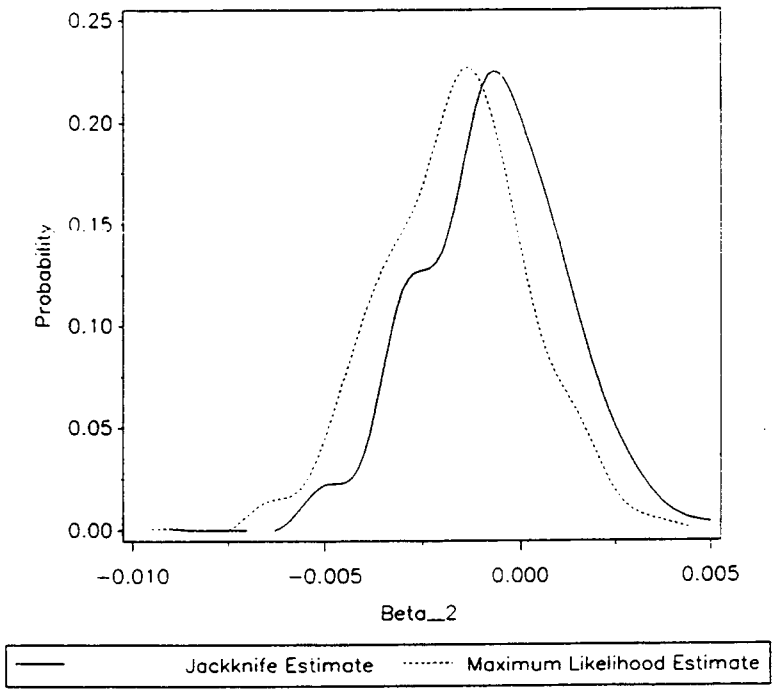
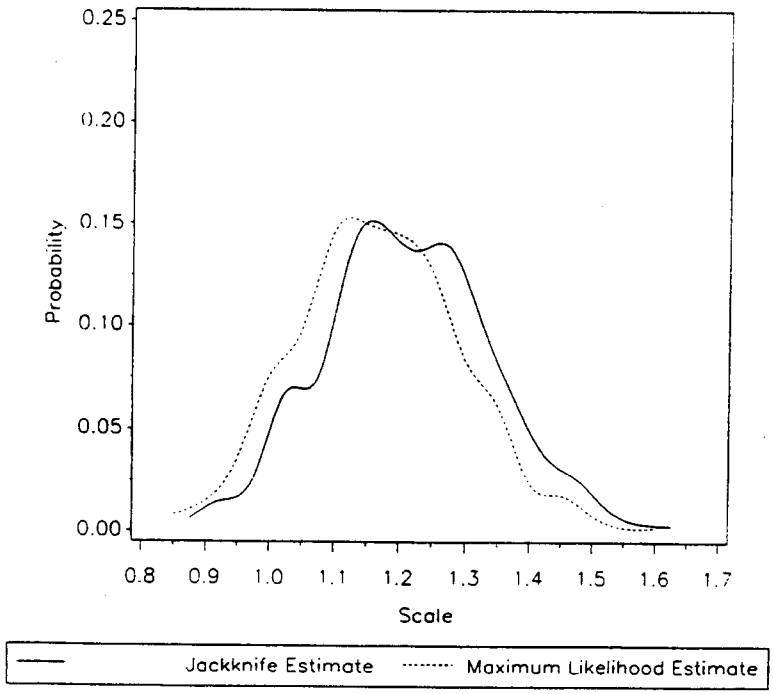


Figure 5.9

Maximum Likelihood and Jackknifed Estimates of the
Scale Parameter of the Lognormal Regression Model
The mean estimates from M=1000 samples of size n=50



Chapter 6

Discussion and Suggestions for Future Research

The main goal of this research work has been to develop the form of bias of the maximum likelihood estimators for the parameters of the length-biased distribution from an exponential family, and jackknife the MLE to reduce their bias. The classical jackknife estimation technique eliminates the bias of order n^{-1}, in the MLE. Both the MLE and the Jackknifed estimators are shown analytically to have the same asymptotic normal distribution.

The Fisher information matrix from the original distribution and its length-biased version are compared by D-optimality and A-optimality criteria. The Fisher information from the length-biased density g of an exponential family is the sum of the Fisher information from the original density f and a matrix of the second partial derivative, with respect to the parameter vector of the $\log E_f[Y]$. It is shown that there are situations when the D-optimality criterion may be inconclusive in determining which design provides more information about the parameters of the model.

Numerical studies are carried out to evaluate the theoretical results given in this research work. The results indicate the MLE are generally biased. However,

in the cases studied, especially for the lognormal distribution, the reduction in the bias of the MLE by jackknifing is minimal. In particular MLE and jackknifed estimates for the shape parameter are identical, indicating no bias of order n^{-1} in the MLE. The jackknife technique performed as well as the MLE in terms of the RMSE. The results also suggest both the MLE and the Jackknifed estimators have the same asymptotic normal distribution, as claimed in our analytical work.

There are issues for further research to be resolved. This work assumes prior knowledge of the distributional form of the response variable. There are situations when this prior information may not be available. In this case some other estimation techniques, such as semi-parametric or other nonparametric approaches, would be the appropriate, and their jackknifed estimators can be analyzed. Some other family of distributions, other than the exponential family for the original distribution would provide more insight into the study of the optimality of the length-biased design compared to sampling from the original design.

Another area of interest for future research would be to consider the length-biased distribution in the case of multivariate response. In this case marginal length-biased distributions may be developed.

BIBLIOGRAPHY

Bayarri, M.J. and De Groot, M.H. (1989). Comparison of Experiments with Weighted Distributions. Statistical Data Analysis and Inference (ed Yadolah Dodge), North Holland Publishing Co.,185-196.

Bayarri, M.J. and De Groot, M.H. (1987). Information in selection models. In *Probability and Bayesian Statistics (ed. R.Viertl*), 39-51.

Bhattacharyya, B.B. Franklin, L.A. and Richardson, G.D. (1988) A comparison of nonparametric unweighted and length-biased density of fibres. *Comm. Statist. A 17* , 3629-44.

Blackwell, D. (1951). Comparison of experiments. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability,*93-102. Berkeley, California: University of California Press.

Brillinger, D.R. (1964). The Asymptotic behavior of Tukey's general method of setting approximate confidence limits (the jackknife) when applied to maximum likelihood estimates. *Review of the International Statistical Institute.* **32,** 202-206.

Cox, D.R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling, (ed.* N.L. Johnson and H. Smith), pp. 506-27. New York: Wiley.

Cramer, H. (1946). Mathematical Methods of Statistics. Princeton, NJ: Princeton University Press.

Godambe, V.P. and Rajarhi, M.B. (1989). Optimal Estimation for Weighted Distributions: Semi-Parametric Model. Statistical Data Analysis and Inference (ed Yadolah Dodge), North Holland Publishing Co.,199-208.

Gray, H.L., Watkins, T.A. and Adams, J.E. (1972). On the jackknife statistic, its extensions and its relation to e_η-transformations. Ann. Math. Statist. **43,** 1-30.

Gupta, R.C. and Keating, J.P. (1986). Relations for Reliability Measures Under Length-Biased Sampling. *Scand. J. Statist.* **13,** 49-56 .

Gupta, R.C. and Kirmani S.N.U.A. (1990). The role of weighted distributions in stochastic modeling. *Comm. Statist. Theory math., 19,* 3147-3162.

Huber, P.J. (1964). Robust estimation of a location parameter. Ann. math. and statist., **35,** 73-101.

Jones, M.C. (1991). Kernel density estimation for length biased data. *Biometrics,* **78,** 3, pp. 511-519.

Kendal, M.G. and Stuart, A. (1958).The Advanced Theory of Statistics, Vol. 2 (1958). Charles Griffin, London.

Krieger, A.M. and Pfeffermann, D. (1992). Maximum Likelihood Estimation from Complex Sample Surveys. *Survey Methodology, 18, 225-239.*

Lindley, D.V. (1956). On measure of Information provided by an experiment. *Ann. Math. Statist., 27,* 968-1005.

LeCam, L. (1956). On the asymptotic theory of estimation and testing hypotheses. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol, **1**, 129-156. Berkeley: University of California Press.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, 2nd edition. London; Chapman and Hall.

Miller, R.G. (1974). The Jackknife - a review. *Biometrika,* **61,** 1, 1-15.

Miller, R.G. (1974). An unbalanced Jackknife. Ann. Statist., **2**, 880-891.

Patil, G.P. and Rao, C.R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics* **34**, 179-89 .

Patil, G.P. and Rao, C.R. (1977). Weighted distributions: A Survey of Their Applications. In *Applications of Statistics,* (ed. P.R. Krishaiah), North Holland Publishing Co., 383-405.

Rao, C.R. (1985). Weighted Distributions Arising out of Methods of Ascertainment: What Population Does a Sample represent? *A celebration of Statistics, the ISI Centenary Volume.* (eds. Anthony C. Aitkinson and S.E. Fienberg), Springer-Verlag, 543-569.

Rao, C.R. (1965). On Discrete Distributions Arising out of Methods of Ascertainment. In *Classical and Contagious Discrete Distributions,* (ed. G.P. Patil), Pergamon Press and Statistical Publishing Society, Calcutta, 320-332.

Rao, C.R. (1973). Linear Statistical Inference and its Applications. 2nd ed. Wiley, New York.

Schucany, W.R., Gray, H.L. and Owen, D.B. (1971). Bias reduction in estimation. J. Amer. Statist. **66** 524-533.

Shah, K.R. and Sinha, B.K. (1989). *Theory of Optimal Designs. Lecture notes in Statistics, 54 (eds. J. berger, S,.Fienberg, J. Gani, K.Krickeberg, and B. Singer). Springer- Verlag.*

Sen, P.K. (1987). What do the Arithmetic, Geometric and Harmonic means tell us in Length-Biased Sampling. *Statistics & Probability Letters,* 5 95-98.

Sen, P.K. and Singer, J.M. (1993). Large Sample Methods in Statistics. An Introduction with Applications, Chapman & Hall, London.

Sen, P.K. (1977). Some Invariance Principles to Jackknifing and their Role in Sequential Analysis. Ann. Statist. **2** 316-329.

Silvey, S.D. (1980). Optimal Design. An introduction to the theory for parameter estimation. *University Press, Cambidge.*

Zelen, M. (1974). Problems in Cell Kinetics and Early Detection of Disease. *Reliability and Biometry,* SIAM, Philadelphia, 701-726.

Wilks, S.S., (1962). Mathematical Statistics. New York: John Wiley.