Efficient Estimation of the Expectation of a Latent Variable in the Presence of Subject-Specific Ancillaries

Louis Mittel

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

Efficient Estimation of the Expectation of a Latent Variable in the Presence of Subject-Specific Ancillaries

Louis Mittel

Latent variables are often included in a model in order to capture the diversity among subjects in a population. Sometimes the distribution of these latent variables are of principle interest. In studies where sequences of observations are taken from subjects, ancillary variables, such as the number of observations provided by each subject, usually also vary between subjects. The goal here is to understand efficient estimation of the expectation of the latent variable in the presence of these subject-specific ancillaries.

Unbiased estimation and efficient estimation of the expectation of the latent parameter depend on the dependence structure of these three subject-specific components: latent variable, sequence of observations, and ancillary. This dissertation considers estimation under two dependence configurations. In Chapter 3, efficiency is studied under the model in which no assumptions are made about the joint distribution of the latent variable and the subject-specific ancillary. Chapter 4 treats the setting where the ancillary variable and the latent variable are independent.

Table of Contents

Acknowledgments

First and foremost, I want to thank my advisor, Professor Daniel Rabinowitz, for showing me how to be a better statistician, mathematician, and writer. His guidance on this dissertation has been indispensable, and having the opportunity to work with him has been both a pleasure and an honor.

I also would like to express my appreciation to Professors Mark Brown, Richard Davis, Zhiliang Ying, and Zhezhen Jin for their willingness to serve on my defense committee and for having provided me with helpful feedback and useful material. I would like to also acknowledge Professor Ingram Olkin who served as a great mentor to me from the first day I expressed an interest in statistics until he passed away in 2016.

Finally, I'd like to thank my parents for their abiding support and love.

Chapter 1

Introduction

The motivation for this dissertation was born out of a consulting experience in 2013 with a sociology doctoral student. The study aim was to estimate how the propensity for interracial romantic relationships differed by self-reported sexual orientation. Previous research on this topic was based on United States census records of cohabiting adults. For example, Rosenfeld and Kim [2005] conclude from an analysis of census data that same-sex couples are more likely than heterosexual married couples to be interracial, reporting an odds ratio of 2.82 in 1990 and 1.42 in 2000. However, census data is limited in that it provides only a current snapshot of information on individuals. As a response to previous work, in this study, data on individuals' partnership histories, not just current partnerships, from the National Longitudinal Study of Adolescent to Adult Health (Add Health) were to be analyzed.

Add Health is an ongoing longitudinal study that was initiated in 1994 for the purpose of examining a wide range of health and social topics. It claims to be the largest and most comprehensive survey of individuals from adolescence ever undertaken [Harris et. al, 2009]. The details pertaining to the study design and the sampling strategy used to achieve a na-

tionally representative sample can be found in Harris [2013]. Field interviewers conducted in-school and in-home interviews with over 15,000 adolescent participants in three waves over the first seven years of the study. At enrollment, background information on each participant was collected, including self-reported race and self-reported sexual orientation. At subsequent interviews, race and gender of participants' partners, for each romantic relationship experienced by participants was obtained from participants by questionnaire (See Wave III: In-Home Questionnaire, Public Use Sample in Harris and Udry [2016]). With this information, each study subject has a corresponding set of partnerships each of which can be classified as interracial or not.

Two features of the described project are particularly consequential. First, an individual's propensity for interracial partnerships varies between individuals, and it is the distribution of this characteristic that is intended to be studied. In particular, the sociologist wants to ultimately compare the distribution of this characteristic between the heterosexual male and the homosexual male populations. Second, not only do individual propensities to form relationships of a certain type vary, but the number of partners an individual has varies too. In the Add Health study, the total number of reported partners ranged from zero to forty-eight over the study's first seven years [Harris, 2009]. With such variation, some subjects present significantly more information about themselves than others.

These two features can be found together in other domains as well. In zoology, in the study of reptiles and birds, a biologist may be interested in maternal propensities to have offspring of a certain weight [Brown and Shine, 2009]. In addition to the distribution of this quality in a population, there also exists variation in the number of eggs laid by mothers. This variation is understood to be the result of genetic variation, variation in parental

behavior, and local environmental conditions like temperature [Pendlebury and Bryant, 2005], [Postma and van Noordwijk, 2005]. If measurements on hatchlings are collected and kept organized by clutch or nest, then this configuration parallels the Add Health study.

Another similar situation arises in the analysis of the sport of baseball. In baseball, a starting pitcher plays in about thirty games each season. Some days his pitching ability is sharper than on others. A pitcher's ability on a given day is reflected in the number of runs his opponent scores each inning that day. As ability varies from start to start, it is the distribution in ability in a given season that describes the pitcher that year. In addition, the starting pitcher pitches a random number of innings each game as the manager usually takes the pitcher out before the game ends.

Each of these examples contains a random number of observations attached to a latent characteristic of interest. This raises the question of how these two phenomena are related, if at all. In the sociology example, is the number of partners a person has independent of their likelihood of having a racially dissimilar partner? In the zoology example, number of eggs in a nest or clutch is known to be depend on many factors. These factors may influence the weight of the eggs, inducing a dependency between the number of observations and the characteristic. In baseball, the manager decides whether or not to take the pitcher out of the game based on the number of runs the pitcher has allowed thus far. Clearly, a pitcher's ability on a given day is not independent of the number of innings pitched.

Even though the interplay between these two features is not of primary importance to the researcher, it is nonetheless involved in studying what is — the latent characteristic in the population. The relationship between the characteristic and number of observations will determine how to perform optimal statistical inference on the primary scientific question.

In particular, the problem of estimating the expectation of the latent variable is sensitive to the dependence assumptions of these components. The work presented in this dissertation illustrates this, and develops unbiasedness and efficiency results under different dependence assumptions.

The types of models that are the basis for this dissertation have the structure of Empirical Bayes models, originally put forward by Herbert Robbins. As is the case here, Robbins [1985] considers a model with a latent mean parameter and a sequence of observations from each subject without making any parametric assumptions about the marginal or "prior" distribution of the latent parameters. Robbins' inferential objective, however, is to estimate the value of a particular latent parameter. Here, the object is to estimate the expectation of the latent parameters. Besides the difference in goals, Robbins treats the number of observations from each subject as fixed quantities, and therefore does not consider that the number of observations is random and possibly not independent of the other subject components, whereas this possibility is of concern here.

More recent developments in Empirical Bayes have focused on the goal of estimating the density of the prior distribution. This is sometimes referred to as the Bayes deconvolution problem. The name derives from the fact that the density of the observed data is a convolution of the assumed known conditional distribution with the unknown density of the latent parameter. Efron [2015] and Narasimhan and Efron [2016] present a solution to the deconvolution problem that operates by assuming a low-parameter exponential family model for the prior density and applying a penalized maximum likelihood procedure.

With an estimate of the entire prior density, a functional of the estimated prior can be computed in order to estimate a functional of the prior. However, if interest lies only in

estimating the particular functional of the prior, this is not necessarily the best approach. Indeed, Efron [2015] cautions that the estimator for the prior presented in the paper contains definitional bias. The upside is that one obtains smooth estimates of the prior with reduced variability, but this trade-off is only worth taking if an estimate of the entire prior density is required. The problem of efficient estimation of a functional of the prior distribution in a basic convolution model appears in Example 25.35 of van der Vaart [1998]. The mixture or convolution model studied is basic in the sense that it involves an identically distributed univariate component and latent parameter from each subject. The case of subjects having different numbers of observations is not treated. The result established is that for these kind of convolutions, empirical estimators are efficient estimators for the functional that is their expectation, at least asymptotically.

Interestingly, one of the demonstrated examples in Efron [2015] lends itself quite well to the considerations of this dissertation and has a similar structure to the sociologist's problem first discussed. The example comes from an intestinal surgery study on 800 cancer patients. During surgery to remove the primary tumor, surgeons also removed a number of "satellite" nodes that would each subsequently be tested to determine whether or not the node was malignant. The number of satellites removed from a patient ranged from one to forty. The number of malignant satellites from a patient were modeled as Binomial variates with the total number of draws being the total number of satellites removed during surgery each with a latent subject-specific probability of being malignant. The paper demonstrates the estimation of the distribution of the latent probability parameter. In the model, each patient's number of satellites removed is modeled as a fixed quantity – the relationship between the number of removed satellites and the patient's probability of a given satellite

being malignant is not considered.

Chapter 2

Notation

Let m denote the number of subjects from a given population of interest. In the sociology example, m would be the number of individuals included in the Add Health data set from a given population - either heterosexual or homosexual adults in the United States. Let i from 1 to m index subjects. Let N_i denote the number of observations made from the i^{th} subject. In the sociology example, N_i would be the number of recorded partners from the i^{th} subject over the period of the study. Let j index observations within subjects. Let Y_{ij} denote the j^{th} observation from the i^{th} subject. In the sociology example, Y_{ij} would be the indicator of racial dissimilarity for the j^{th} partner of the i^{th} subject. In the zoology example, Y_{ij} might be the weight of the j^{th} hatchling in the i^{th} clutch. In the baseball example, Y_{ij} would be a pitcher's number of runs allowed in the j^{th} inning of the pitcher's i^{th} game.

Let θ_i (i from 1 to m) denote a latent subject-specific parameter associated with the i^{th} subject. In the sociology example, θ_i would parameterize the likelihood that any given partner of the i^{th} subject is racially dissimilar. In the zoology example, θ_i would parameterize the mean weight of hatchlings born from the i^{th} clutch. In the baseball example, θ_i would

parameterize the mean number of runs allowed per inning in the pitcher's i^{th} game.

It is convenient to posit a sequence of potential observations, $Y_{i1}, Y_{i2}, \ldots$ from each subject, so that the observed outcomes from the i^{th} subject are the first N_i terms in the subject's sequence of potential observations. In the sociology example, data from a single subject might be modeled as Bernoulli trials with a probability parameter, particular to that subject. More generally, suppose that for each i, given θ_i, the $Y_{i1}, Y_{i2}, \ldots$ are independent and identically distributed with density or probability mass function of the form,

$$f_*^t(y) = f_0(y)e^{ty - A(t)} \tag{N.1}$$

for a specified f_0. Let $f^t(y; d)$ denote the density or probability mass function of the sum the first d of these variables given θ_i. In such families,

$$A''(\theta_i) = \int f_*^\theta(y)(y - A'(\theta_i))^2 d\, y$$

is the conditional variance of Y_{ij} given θ_i.

In order to reflect that study subjects are drawn from a population of interest, suppose that the θ_i , $Y_{i1}, Y_{i2}, \ldots$, N_i triplets are independent and identically distributed. In the sociology example, subjects can be treated as constituting a random sample by the design of the Add Health study. It is useful to let θ, $Y_{\cdot 1}, Y_{\cdot 2}, \ldots$, and N denote a generic latent parameter, sequence of potential observations, and number of subject-specific observations

drawn from the joint distribution of the triplets. Let

$$Y = \sum_{j=1}^{N} Y_{\cdot j} \, .$$

In the sociology example, Y is the total number of racially dissimilar partners recorded from a generic subject.

Let π denote the common density of θ, and ρ denote the common joint distribution of θ and N. Let μ denote the common expectation of $Y_{\cdot j}$:

$$\mu = \int \pi(\theta) A'(\theta) d\,\theta \, .$$

Of interest is estimation of μ. In the sociology example, μ would be the expected propensity for having interracial partners in the given population.

There are several possibilities for the dependence structure of N with respect to θ and the sequence of $Y_{\cdot 1}, Y_{\cdot 2}, \ldots$. These include:

1.

$$N \perp Y_{\cdot 1}, Y_{\cdot 2}, \ldots | \, \theta \tag{N.3}$$

2.

$$N \perp \theta, Y_{\cdot 1}, Y_{\cdot 2}, \ldots \tag{N.4}$$

3.

$$P\{N = k | \, \theta, Y_{\cdot 1}, Y_{\cdot 2}, \ldots\} = P\{N = k | \, \theta, Y_{\cdot 1}, Y_{\cdot 2}, \ldots, Y_{\cdot k}\} \tag{N.5}$$

The first case (N.3) is treated in Chapter 3, and the second case (N.4) is treated in Chapter 4. The third case, (N.5), captures the structure in the example of pitchers in baseball and is a topic of future research.

A technical note: if the marginal distribution of the number of observations, N, has positive mass at zero (as is the case in the sociology example) then, under non-independence between N and θ, the expectation of $A'(\theta)$ is unestimatable because no data from this stratum of the population are ever observed. The natural estimable quantity under these circumstances is the conditional expectation of θ given $N \geq 1$ which can only be assumed to be a satisfactory replacement for the full unconditional expectation of θ. This assumption as needed is implicit in what follows.

Chapter 3

Ancillary Not Independent of Latent Parameter

Consider a model in which the subject-specific ancillary, N, and the subject sequence of potential observations, $Y_{\cdot 1}, Y_{\cdot 2}, \ldots$, are independent conditionally given the latent parameter:

$$N \perp Y_{\cdot 1}, Y_{\cdot 2}, \ldots \,|\, \theta \,. \tag{N.3}$$

As discussed in the previous chapter, it is assumed that

$$Y_{\cdot 1}, Y_{\cdot 2}, \ldots \,|\, \theta \overset{iid}{\sim} f_*^\theta(y) \,. \tag{N.1}$$

The subject-specific ancillary, N, and the the subject-specific latent parameter, θ, are drawn from the common joint distribution ρ, which will be left unspecified, so as not to suppose anything about the distribution of the latent parameter nor how θ and N are related:

$$\theta, N \sim \rho \text{ (unspecified)} \tag{3.1}$$

These three conditions define the model that all results and discussion in this chapter will concern. A subject's observed data is $Y_{\cdot 1}, \ldots, Y_{\cdot N}$. The model allows for some preliminary data reduction.

Lemma 3.1. *Y, N is sufficient for $Y_{\cdot 1}, \ldots, Y_{\cdot N}$.*

Proof. As a consequence of N.3 and N.1,

$$Y_{\cdot 1}, Y_{\cdot 2}, \ldots \,|\, \theta, N \overset{iid}{\sim} f_*^\theta(y)\,. \tag{3.2}$$

From (3.2),

$$Y_{\cdot 1}, \ldots, Y_{\cdot N} |\, \theta, N \overset{iid}{\sim} f_*^\theta(y)\,.$$

With the density $f_*^\theta(y)$ having exponential family form, Y is sufficient for $Y_{\cdot 1}, \ldots, Y_{\cdot N}$ conditionally given N and θ, and

$$Y |\, \theta, N \sim f^\theta(y; n)\,. \tag{3.3}$$

Thus, the likelihood of the observed data,

$$Y_{i1}, \ldots, Y_{iN_i}, N_i \quad i = 1, \ldots, m$$

can be written as a function of the Y_i, N_i pairs:

$$\prod_{i=1}^{m} \int f^\theta(Y_i; N_i) \rho(\theta, N_i)\; d\,\theta \tag{3.4}$$

□

Unbiasedness

There are two estimators that are particularly relevant to the discussion of unbiased estimation of μ:

$$\hat{\mu}_1 = \frac{1}{m}\sum_{i=1}^{m}\frac{Y_i}{N_i} = \frac{1}{m}\sum_{i=1}^{m}\bar{Y}_i$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{m} Y_i}{\sum_{i=1}^{m} N_i}\,.$$

These estimators are relevant because they are the efficient estimators in two related fully-parametric models. Imagine that instead of modeling $\theta_1, \ldots, \theta_m$ as random parameters drawn from a common distribution, they are treated as arbitrary and unrelated fixed parameters. In such a case, the parameter playing the role of μ would be

$$\frac{1}{m}\sum_{i=1}^{m} A'(\theta_i)\,.$$

In such a model, the efficient estimator would be $\hat{\mu}_1$, the maximum-likelihood estimator. If it were additionally assumed that all these fixed parameters were equal, $\theta_1 = \cdots = \theta_m = \theta$, the efficient estimator of μ would be $\hat{\mu}_2$, the maximum-likelihood estimator under this restricted model. (See Appendix A.1).

Returning now to the model of this chapter, consider $\hat{\mu}_1$, the uniformly weighted average

of subject-specific conditionally-unbiased estimators of $A'(\theta_i)$.

Theorem 3.2. *$\hat{\mu}_1$ is unbiased for μ.*

Proof.

$$\begin{aligned}
\mathbb{E}\{\hat{\mu}_1\} &= \mathbb{E}\left\{ \frac{1}{m} \sum_{i=1}^{m} \frac{Y_i}{N_i} \right\} \\
&= \mathbb{E}\left\{ \frac{1}{m} \sum_{i=1}^{m} \frac{Y}{N} \right\} \\
&= \mathbb{E}\left\{ \frac{Y}{N} \right\} \\
&= \mathbb{E}\left\{ \frac{1}{N} \mathbb{E}\Big\{ Y \Big| N, \theta \Big\} \right\} \\
&= \mathbb{E}\left\{ \frac{1}{N} N\, A'(\theta) \right\} && \text{(by (3.3))} \\
&= \mathbb{E}\Big\{ A'(\theta) \Big\} = \mu
\end{aligned}$$

□

The estimator $\hat{\mu}_2$ can be written as the weighted average of subject-specific conditionally-unbiased estimators of $A'(\theta_i)$, weighted by N_i:

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{m} Y_i}{\sum_{k=1}^{m} N_k} = \sum_{i=1}^{m} \frac{N_i}{\sum_{k=1}^{m} N_k} \frac{Y_i}{N_i}$$

Unlike $\hat{\mu}_1$, the estimator $\hat{\mu}_2$ is biased. This bias is critical in that it exists both in finite

samples and asymptotically. Indeed, the asymptotic bias is:

$$\frac{\mathrm{Cov}(A'(\theta), N)}{\mathbb{E}\{N\}}\,.$$

This is proved in the following theorem, relating the expectation of $\hat{\mu}_2$ to μ in finite samples.

Theorem 3.3.

$$\mathbb{E}\Big\{\hat{\mu}_2\Big\} = \mu + \frac{\mathrm{Cov}\big(A'(\theta), N\big)}{\mathbb{E}\{N\}} + \mathbb{E}\left\{\left(\frac{1}{\overline{N}} - \frac{1}{\mathbb{E}\{N\}}\right) N_1 A'(\theta_1)\right\}$$

where $\overline{N} = \frac{1}{m}\sum_{i=1}^{m} N_i\,.$

Proof.

$$\begin{aligned}
\mathbb{E}\Big\{\hat{\mu}_2\Big\} &= \mathbb{E}\left\{\sum_{i=1}^{m} \frac{N_i}{\sum_{k=1}^{m} N_k} \frac{Y_i}{N_i}\right\} \\
&= m\,\mathbb{E}\left\{\frac{N_1}{\sum_{k=1}^{m} N_k} \frac{Y_1}{N_1}\right\} \\
&= \mathbb{E}\left\{\frac{Y_1}{\overline{N}}\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{E}\{Y_1 | N_1, \theta_1\}}{\overline{N}}\right\} \\
&= \mathbb{E}\left\{\frac{N_1\, A'(\theta_1)}{\overline{N}}\right\} && \text{(by (3.3))} \\
&= \frac{\mathbb{E}\{N A'(\theta)\}}{\mathbb{E}\{N\}} + \mathbb{E}\left\{\left(\frac{1}{\overline{N}} - \frac{1}{\mathbb{E}\{N\}}\right) N_1 A'(\theta_1)\right\}.
\end{aligned}$$

The result follows from the fact that

$$\frac{\mathbb{E}\{NA'(\theta)\}}{\mathbb{E}\{N\}} = \frac{\mathrm{Cov}(A'(\theta), N)}{\mathbb{E}\{N\}} + \mu\,.$$

□

Corollary 3.4.

$$\lim_{m\to\infty} \mathbb{E}\Big\{\hat{\mu}_2\Big\} = \frac{\mathrm{Cov}(A'(\theta), N)}{\mathbb{E}\{N\}} + \mu$$

Proof. By Slutsky's theorem,

$$\lim_{m\to\infty} \mathbb{E}\left\{\left(\frac{1}{\overline{N}} - \frac{1}{\mathbb{E}\{N\}}\right) N_1 A'(\theta_1)\right\} = 0\,.$$

□

Efficiency

Minimum-Variance Bound of Unbiased Estimators

Under the model of this chapter, an unbiased estimator of μ from Y_{ij} , $j = 1, \ldots, N_i$, $i = 1, \ldots, m$ has variance no less than

$$\frac{1}{m}\left(Var\{A'(\theta)\} + E\Big\{\frac{A''(\theta)}{N}\Big\}\right).$$

The proof involves definitions and Lemmas.

A function is *permutation invariant* if it is invariant to permutations with respect to the subject arguments of the data. More formally, a function h on the support of the subjects'

triplets is *permutation invariant* if for every permutation σ of the indices i from 1 to m,

$$h\big((Y_{\sigma_1}, N_{\sigma_1}, \theta_{\sigma_1}), \ldots, (Y_{\sigma_m}, N_{\sigma_m}, \theta_{\sigma_m})\big) = h\big((Y_1, N_1, \theta_1), \ldots, (Y_m, N_m, \theta_m)\big).$$

Let $\tilde{\mu}$ be any unbiased estimator of μ, and let $\hat{\mu}$ denote the permutation invariant version of $\tilde{\mu}$:

$$\hat{\mu} = \frac{1}{m!} \sum_{\sigma \in \Pi_m} \tilde{\mu}\big((Y_{\sigma_1}, N_{\sigma_1}), \ldots, (Y_{\sigma_m}, N_{\sigma_m})\big)\,.$$

Lemma 3.5.

$$Var\{\tilde{\mu}\} \geq Var\{\hat{\mu}\}$$

Proof. As established in Lemma 3.1, Y_i, N_i is sufficient for the the observed sequence from subject i. Because all m observed pairs are independent and identically distributed, the unordered collection of pairs $S = \{(Y_1, N_1), \ldots, (Y_m, N_m)\}$ is a sufficient statistic for the observed data under the model.

By the Rao-Blackwell Theorem, the conditional expectation of any unbiased estimator with respect to any sufficient statistic creates a new estimator with no larger variance. Because $\hat{\mu} = \mathbb{E}\big\{\tilde{\mu}\big|S\big\}$, the result follows. □

Lemma 3.6. *If an estimator $\hat{\mu}$ is unbiased and permutation invariant then,*

$$\mathbb{E}\big\{\hat{\mu}\big|(N_1, \theta_1), \ldots, (N_m, \theta_m)\big\} = \frac{1}{m} \sum_{i=1}^{m} A'(\theta_i) \tag{3.5}$$

Proof. If $\hat{\mu}$ is permutation invariant, then $\mathbb{E}\big\{\hat{\mu}\big|(N_1, \theta_1), \ldots, (N_m, \theta_m)\big\}$ must also be permutation invariant. This is proved in Appendix A.2. Consequently, the left side of equa-

tion 3.5 is a function of the collection of (θ_i, N_i) pairs. On the right side of equation 3.5, $\frac{1}{m}\sum_{i=1}^{m} A'(\theta_i)$ is also a function of the collection of (θ_i, N_i) pairs.

Because it is assumed that the joint distribution of (θ_i, N_i) is fully unspecified (3.1), the collection of (θ_i, N_i) pairs is complete sufficient (Shao [2003], Example 2.17). Since both functions are functions of the complete sufficient statistic and have the same expectation, by completeness, the functions must be almost surely identical. □

Lemma 3.7. *Suppose $\hat{\mu}$ is unbiased for μ and permutation invariant, then*

$$Var\big\{\hat{\mu}\big|(N_1,\theta_1),\ldots,(N_m,\theta_m)\big\} \geq \frac{1}{m^2}\sum_{i=1}^{m}\frac{A''(\theta_i)}{N_i}\,.$$

The proof proceeds by showing first that under the conditional distribution, the projection of $\hat{\mu}$ onto the span of functions of the form

$$\sum_{i=1}^{m} h_i(Y_i, N_i)$$

may be expressed as

$$\frac{1}{m}\sum_{i=1}^{m} h^*(Y_i, N_i)$$

where

$$\mathbb{E}\big\{h^*(Y_i,N_i)\big|(N_1,\theta_1),\ldots,(N_m,\theta_m)\big\} = A'(\theta_i)\,.$$

It is then shown that such $h^*(Y_i, N_i)$ satisfy

$$Var\big\{h^*(Y_i,N_i)\big|(N_1,\theta_1),\ldots,(N_m,\theta_m)\big\} \geq \frac{A''(\theta_i)}{N_i}$$

from which the result follows directly.

Proof. Let $\mathcal{Z} = (N_1, \theta_1), \ldots, (N_m, \theta_m)$, and let

$$h^*(Y_i, N_i) = m\mathbb{E}\big\{\hat{\mu}\big|Y_i, N_i, \mathcal{Z}\big\} - \sum_{k\neq i} A'(\theta_k)$$

To prove that

$$\mathbb{E}\big\{h^*(Y_i, N_i)\big|\mathcal{Z}\big\} = A'(\theta_i)\,, \tag{3.6}$$

note that by taking expectations and applying Lemma 3.6:

$$\begin{aligned}\mathbb{E}\Big\{\mathbb{E}\big\{\hat{\mu}\big|Y_i, N_i, \mathcal{Z}\big\}\Big|\mathcal{Z}\Big\} &= \mathbb{E}\Big\{\hat{\mu}\Big|\mathcal{Z}\Big\} \\ &= \frac{1}{m}\sum_{i=1}^{m} A'(\theta_i)\,.\end{aligned}$$

To prove that

$$\frac{1}{m}\sum_{i=1}^{m} h^*(Y_i, N_i)$$

is the projection, it suffices to show that

$$\mathbb{E}\Big\{\sum_{i=1}^{m} f_i(Y_i, N_i)\Big[\hat{\mu} - \frac{1}{m}\sum_{j=1}^{m} h^*(Y_j, N_j)\Big]\Big|\mathcal{Z}\Big\} = 0$$

for any $\sum_{i=1}^{m} f_i(Y_i, N_i)$. That $h^*(Y_i, N_i)$ indeed satisfies the above equation is verified in Appendix A.3.

Finally, to show that

$$Var\big\{h_i^*(Y_i, N_i)\big|(N_1, \theta_1), \ldots, (N_m, \theta_m)\big\} \geq \frac{A''(\theta_i)}{N_i}$$

note that $h_i^*(Y_i, N_i)$ is an unbiased estimator of $A'(\theta_i)$ under the conditional distribution given $(N_1, \theta_1), \ldots, (N_m, \theta_m)$ (3.6). This suffices as a result of the Fisher Information theorem and the independence condition of this chapter (N.3).

The Fisher Information inequality requires certain regularity conditions on the relevant density to be met: the support must not depend on the parameter, the derivative of the density with respect to the parameter must exist for all values inside the parameter space and within the support, and integration and differentiation must be interchangable. Because it is assumed that $f_*^\theta(y)$ is an exponential family these conditions are all satisfied. A modified result that relaxes the exponential family assumption can also be obtained and is discussed after the proof of the theorem. □

Having established these two Lemmas, the efficiency theorem will now be proved.

Theorem 3.8. *Suppose that*

$$N \perp Y_{\cdot 1}, Y_{\cdot 2}, \ldots \,|\, \theta \tag{N.3}$$

$$\theta, N \sim \rho \text{ (unspecified)} \tag{3.1}$$

$$Y_{\cdot 1}, Y_{\cdot 2}, \ldots \,|\, \theta \overset{iid}{\sim} f_*^\theta(y) \tag{N.1}$$

Then,

$$Var\{\tilde{\mu}\} \geq \frac{1}{m}\left(Var\{A'(\theta)\} + E\Big\{\frac{A''(\theta)}{N}\Big\}\right)$$

Proof.

$$\begin{aligned}
Var\{\tilde{\mu}\} &\geq Var\{\hat{\mu}\} && (3.7)\\
&= Var\Big\{\mathbb{E}\{\hat{\mu}\Big|(N_1,\theta_1),\ldots,(N_m,\theta_m)\}\Big\} \\
&\quad + \mathbb{E}\Big\{Var\Big\{\hat{\mu}\Big|(N_1,\theta_1),\ldots,(N_m,\theta_m)\Big\}\Big\} && (3.8)\\
&= Var\Big\{\frac{1}{m}\sum_{i=1}^{m} A'(\theta_i)\Big\} + \mathbb{E}\Big\{Var\Big\{\hat{\mu}\Big|(N_1,\theta_1),\ldots,(N_m,\theta_m)\Big\}\Big\} && (3.9)\\
&\geq Var\Big\{\frac{1}{m}\sum_{i=1}^{m} A'(\theta_i)\Big\} + \mathbb{E}\Big\{\frac{1}{m^2}\sum_{i=1}^{m}\frac{A''(\theta_i)}{N_i}\Big\} && (3.10)\\
&= \frac{1}{m}\left(Var\{A'(\theta)\} + E\Big\{\frac{A''(\theta)}{N}\Big\}\right)
\end{aligned}$$

The first inequality (3.7) follows by Lemma 3.5. Line (3.8) follows from the usual conditional decomposition of variance. Line (3.9) applies Lemma 3.6. Line (3.10) follows from Lemma 3.7; if for all values of $(N_1,\theta_1),\ldots,(N_m,\theta_m)$,

$$Var\Big\{\hat{\mu}\Big|(N_1,\theta_1),\ldots,(N_m,\theta_m)\Big\} \geq \frac{1}{m^2}\sum_{i=1}^{m}\frac{A''(\theta_i)}{N_i} \quad ,$$

then their expectations maintain the inequality. □

Corollary 3.9. *$\hat{\mu}_1$ achieves the variance bound.*

The unbiasedness of $\hat{\mu}_1$ was proved in Lemma 3.2. The variance of $\hat{\mu}_1$ is

$$\frac{1}{m}\left(Var\big\{A'(\theta)\big\} + E\Big\{\frac{A''(\theta)}{N}\Big\}\right).$$

Example 3.10. *Suppose $Y|\theta, N \sim Binom(N, A'(\theta))$. Then, the efficient estimator of μ is*

$$\hat{\mu}_1 = \frac{1}{m}\sum_{i=1}^{m}\frac{Y_i}{N_i},$$

and

$$Var\Big\{\hat{\mu}_1\Big\} = \frac{1}{m}Var\Big\{\frac{Y}{N}\Big\} = \frac{1}{m}\left(Var\big\{A'(\theta)\big\} + E\Big\{\frac{A'(\theta)\big(1-A'(\theta)\big)}{N}\Big\}\right).$$

Extentions

This theorem can be extended to cover non-exponential family cases so long as subject-level conditional variance bounds for unbiased estimators of $\mathbb{E}(Y_{\cdot j}|\theta)$ are available. To maintain consistency of notation, let $A'(\theta)$ still refer to $\mathbb{E}(Y_{\cdot j}|\theta)$ even though $f_*^\theta(y)$ is no longer assumed to be an exponential family. If for any statistic ψ such that

$$\mathbb{E}\big\{\psi(Y,N)\big|N,\theta\big\} = A'(\theta)$$

for all values of θ and N implies that,

$$Var\big\{\psi(Y,N)\big|N,\theta\big\} \geq B(\theta,N),$$

then this proof method will show that the variance bound of any unbiased estimator of μ will be

$$\frac{1}{m}\left(Var\{A'(\theta)\} + E\Big\{B(\theta, N)\Big\}\right).$$

For example, suppose the distribution family $\{f_*^\theta(y) : \theta \in \Theta\}$ is a dominated family with Θ an open set in $\mathbb{R}$ and $f_*^\theta(y)$ differentiable with respect to θ. Under these conditions, the Fisher Information theorem can be applied to obtain a conditional variance bound, and through this bound the same proof methods will show that

$$Var\{\tilde{\mu}\} \geq \frac{1}{m}\left(Var\{A'(\theta)\} + E\Big\{B(\theta, N)\Big\}\right)$$

where

$$B(\theta, N) = \frac{A''(\theta)^2}{NI(\theta)} \quad \text{and} \quad I(\theta) = Var\Big\{\frac{\partial}{\partial\theta}\log f_*^\theta(y)\Big\}.$$

Discussion

If it was known that in addition to being sufficient, the collection of (Y_i, N_i) was complete under the model for the observed data, then completeness could be used directly to establish that $\hat{\mu}_1$ achieves the lowest possible variance among all unbiased estimators. However, it is not evident that the collection of (Y_i, N_i) is indeed complete. Recall that the likelihood for the observed data is

$$\prod_{i=1}^{m}\int f^\theta(Y_i|N_i)\rho(\theta, N_i)\; d\,\theta\,.$$

The observed data are drawn from a mixture of a nonparametric family (ρ) over a parametric

family (f^θ). This additional structure in the family of models from which the observed data are sampled from, differentiates it from the purely nonparametric case in which the collection of observations would be complete.

Furthermore, if f^θ is not an exponential family such that Y, N is not complete conditionally, then certainly the collection of (Y_i, N_i) will not be complete under the model for the observed data. Hence, a proof method for Theorem 3.8 that does not rely on the completeness of the collection of (Y_i, N_i) is advantageous.

The model for the observed data in this chapter is different from a model in which the observed data follow a completely unspecified distribution. This difference is discussed in the context of a simpler mixture model (without the aspect of ancillary variables) in Example 25.35 of van der Vaart [1998]. As van der Vaart explains:

> Nonparametric mixtures over an exponential family form very large models, which are only slightly smaller than the nonparametric model. For estimating a functional such as the mean of the observations, it is of relatively little use to know that the underlying distribution is a mixture. More precisely, the additional information does not decrease the asymptotic variance, although there may be an advantage for finite [sample size]. On the other hand, the mixture structure may express a structure in reality and the mixing distribution may define the functional of interest.

Connection with the Inspection Paradox

In this section, it is demonstrated how the model in this chapter is a generalization of the Inspection Paradox. As will be shown, in the Inspection Paradox model, ρ is partially — rather than completely — unspecified. The degree of additional specification depends on the version of the Inspection Paradox.

There are several versions of the Inspection Paradox. Consider first perhaps the most basic form. Supposes that there are m families in a population, arbitrarily numbered [Ross, 2003]. Let N_i denote the number of people in family i. It is assumed that the N_i, $i = 1, \ldots, m$ are independent and follow a common unspecified distribution. Let Y_{ij} be the family size of the j^{th} person in the i^{th} family. The Inspection Paradox gets its name from the fact that the distribution of family size if people are sampled differs from the distribution of family size if households are sampled. This will be shown to be analogous to the difference between the limiting distributions of the estimators $\hat{\mu}_1$ and $\hat{\mu}_1$ under the model of this chapter. As a result, the relationship between the expectations of these two distributions can be attained via Theorem 3.3.

Members of the same household each live with the same number of people, so all Y_{ij} are identical and equal to N_i within household i. To embed this model inside the one for this chapter, let $A'(\theta_i)$ be the common value of Y_{ij} in household i. Now, instead of ρ being completely unspecified, the relationship between N_i and $A'(\theta_i)$ is deterministic and known: $A'(\theta_i) = N_i$. The distribution of N_i remains unspecified.

Another version of the Inspection Paradox concerns waiting times for trains. (The Inspection Paradox is also sometimes referred to as the Waiting Time Paradox.) In this version,

one supposes that riders arrive to a station according to a Poisson Process with rate λ. Let the train interarrival times be $2A'(\theta_1), 2A'(\theta_2), \ldots$. It is assumed that these interarrival times are independent and identically distributed from an unspecified distribution. Let N_i be the number of riders that arrive during the i^{th} interarrival period. All these riders take the same train. Arbitrarily number the riders in each train, and let Y_{ij} be the amount of time the j^{th} rider in the i^{th} train had to wait for the trains arrival.

Unlike in the previous case, here Y_{ij}, $j = 1, \ldots, N_i$, are not identical — riders on the same train have not all waited the same amount of time. Instead it follows from the Poisson Process that

$$Y_{i1}, Y_{i2}, \ldots \big| \theta_i \overset{iid}{\sim} \text{Uniform}\big(0, 2A'(\theta_i)\big)\,,$$

and

$$N_i \big| \theta_i \sim \text{Poisson}\big(2\lambda A'(\theta_i)\big)\,. \tag{3.11}$$

Hence,

$$\mathbb{E}\big\{N_i \big| \theta_i\big\} = 2\lambda A'(\theta_i)\,. \tag{3.12}$$

This all can be framed in terms of the model of this chapter. Take $f_*^\theta(y)$ to be the $\text{Uniform}\big(0, 2A'(\theta)\big)$ density, and instead of ρ, the joint distribution of θ and N being left unspecified, let the marginal distribution of θ be unspecified, while $N|\theta$ follows the distribution (3.11).

Because of the relationship in expectation between $2A'(\theta)$ and N from (3.12),

$$\frac{\text{Cov}(2A'(\theta), N)}{\mathbb{E}\big\{N\big\}} = \frac{2\lambda \text{Cov}(A'(\theta), A'(\theta))}{2\lambda \mathbb{E}\big\{A'(\theta)\big\}} = \frac{\text{Var}\big(A'(\theta)\big)}{\mathbb{E}\big\{A'(\theta)\big\}}\,.$$

This then can be applied in Theorem 3.3 to obtain the relationship, in the limit, between the mean rider waiting time, $\widetilde{\mathbb{E}}(Y_{ij})$, and the mean midpoint interarrival time, $\mathbb{E}\{A'(\theta)\}$:

$$\widetilde{\mathbb{E}}(Y_{ij}) = \frac{\mathrm{Var}(A'(\theta))}{\mathbb{E}(A'(\theta))} + \mathbb{E}\{A'(\theta)\} \tag{3.13}$$

This result would otherwise be derived through the size-biasing transformation [Brown, 2006] which relates the interarrival distribution to the distribution of the interarrival length associated with a rider who arrives at any given point. If g is the probability density function of $2A'(\theta)$, and $\tilde{g}$ is the probability density function of $2A'(\theta)$ covering any fixed point, then

$$\tilde{g}(\theta) = \frac{\theta g(\theta)}{\mu}\,.$$

Taking expectations and rewriting the expression yields Equation 3.13.

Chapter 4

Ancillary Independent of Latent Parameter

In the previous chapter, the joint distribution of N and θ was left completely unspecified. In this chapter, the model is modified by adding the restriction that these two components be independent, while their marginal distributions remain completely unspecified:

$$N \perp \theta, Y_{\cdot 1}, Y_{\cdot 2}, \ldots \tag{N.4}$$

$$\theta \sim \pi \text{ (unspecified)} \tag{4.1}$$

The parameter space in this model is the collection of all distributions on θ. As before, it is assumed that the functional form of the conditional density of subject observations is an

exponential family specified up to θ:

$$Y_{\cdot 1}, Y_{\cdot 2}, \ldots \,|\, \theta \overset{iid}{\sim} f_*^\theta(y)\,. \tag{N.1}$$

With the restriction of independence, this new family of models is a subclass of the ones considered in the first chapter, and as such there will be a larger set of unbiased estimators with a different efficiency bound.

The likelihood function of the observed data can again be written as a function of the Y_i, N_i pairs:

$$\prod_{i=1}^{m} \Pr(N_i) \prod_{i=1}^{m} \int f^\theta(Y_i; N_i)\pi(\theta)\, d\theta\ .$$

Because $N_1, \ldots, N_m$ are independent of the rest of the model (N.4), and the rest of the model contains the parameter of interest, by the Conditionality principle, statistical inference can be performed without loss under the conditional distribution of the observations given the particular $N_1, \ldots, N_m$ that are observed. Note that after conditioning, observations are no longer identically distributed. In what follows, all expectations are taken conditional on $N_1, \ldots, N_m$, though the notation will be suppressed.

Unbiasedness

The class of unbiased estimators of μ for the model of this chapter includes all weighted averages of subject-specific averages of the form

$$\sum_{i=1}^{m} a_i \bar{Y}_i$$
$$\text{such that } \sum_{i=1}^{m} a_i = 1\,.$$

where $a_i = f(N_1, \ldots, N_m)$. This class includes both estimators discussed in the previous section, (and any linear combination of them):

$$\hat{\mu}_1 = \frac{1}{m}\sum_{i=1}^{m} \bar{Y}_i \qquad \hat{\mu}_2 = \sum_{i=1}^{m} \frac{N_i}{\sum_k N_k} \bar{Y}_i\,.$$

These estimators are identical if $N_1, \ldots, N_m$ are all identical. Assuming $N_1, \ldots, N_m$ are not identical, neither of these estimators dominates the other (in so far as having smaller variance) under all distributions π of θ. This is demonstrated in the following Lemma and example. The joint asymptotic distribution of the estimators can be found in Appendix A.4.

Lemma 4.1. *Let* $f_{-1}(t) = t^{-1}$ *and* $\overline{f(Y)} = \frac{1}{m}\sum_{i=1}^{m} f(Y_i)$.

$$Var\Big\{\hat{\mu}_1\Big\} = \frac{1}{m}\left(Var\big\{A'(\theta)\big\} + \mathbb{E}\big\{A''(\theta)\big\}\overline{f_{-1}(N)}\right)$$
$$Var\Big\{\hat{\mu}_2\Big\} = \frac{1}{m}\left(Var\big\{A'(\theta)\big\}\frac{\overline{N^2}}{\overline{N}^2} + \mathbb{E}\big\{A''(\theta)\big\}f_{-1}(\overline{N})\right).$$

Proof. From

$$Var\Big\{\frac{Y_i}{N_i}\Big\} = Var\Big\{A'(\theta)\Big\} + \frac{\mathbb{E}\big\{A''(\theta)\big\}}{N_i}\,,$$

after calculation, the variance of the estimators can be expressed as

$$Var\Big\{\hat{\mu}_1\Big\} = \frac{1}{m}\left(Var\big\{A'(\theta)\big\} + \mathbb{E}\big\{A''(\theta)\big\}\frac{1}{m}\sum_{i=1}^{m}\frac{1}{N_i}\right)$$

$$Var\Big\{\hat{\mu}_2\Big\} = \frac{1}{m}\left(Var\big\{A'(\theta)\big\}\frac{m\sum_{i=1}^{m}N_i^2}{(\sum_{i=1}^{m}N_i)^2} + \mathbb{E}\big\{A''(\theta)\big\}\frac{m}{\sum_{i=1}^{m}N_i}\right).$$

Noting that

$$\frac{m\sum_{i=1}^{m}N_i^2}{(\sum_{i=1}^{m}N_i)^2} = \frac{\overline{N^2}}{\overline{N}^2} \qquad \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N_i} = \overline{f_{-1}(N)}$$

$$\frac{m}{\sum_{i=1}^{m}N_i} = f_{-1}(\overline{N})$$

□

Remark. By Jensen's inequality, assuming that $N_1, \ldots, N_m$ are not all identical,

$$\overline{f_{-1}(N)} > f_{-1}(\overline{N}) \qquad \text{and} \qquad \frac{\overline{N^2}}{\overline{N}^2} > 1$$

Thus, if the variance expressions are compared term by term, the first term is smaller in $Var\big\{\hat{\mu}_1\big\}$ whereas the second is smaller in $Var\big\{\hat{\mu}_2\big\}$. Therefore, in general, which estimator has the smaller variance will depend on the relative magnitudes of $Var\big\{A'(\theta)\big\}$ and $\mathbb{E}\big\{A''(\theta)\big\}$ under π, and neither estimator has smaller variance than the other under all π. This is demonstrated in the following example.

Example 4.2. *Suppose* $Y|\theta, N \sim Binom(N, A'(\theta))$. *Note that*

$$A''(\theta) = A'(\theta)\big(1 - A'(\theta)\big).$$

If π is a dirac function at $A'(\theta^*)$, then

$$Var\{A'(\theta)\} = 0 \qquad and \qquad \mathbb{E}\{A''(\theta)\} = \theta^*(1-\theta^*)\,.$$

In this case, $Var\{\hat{\mu}_2\} < Var\{\hat{\mu}_1\}$.

If π is the distribution such that $P_\pi(A'(\theta) = 0) = P_\pi(A'(\theta) = 1) = \frac{1}{2}$ (the distribution of maximal variance), then

$$Var\{A'(\theta)\} = \frac{1}{4} \qquad and \qquad \mathbb{E}\{A''(\theta)\} = 0\,.$$

In this case, $Var\{\hat{\mu}_1\} < Var\{\hat{\mu}_2\}$.

Efficiency

One possibility is that efficient estimator for μ in the model of this chapter is

$$\hat{\mu}_3 = \left(\sum_{i=1}^{m} \frac{1}{Var\{\overline{Y}_i\}} \right)^{-1} \sum_{i=1}^{m} \frac{\bar{Y}_i}{Var\{\overline{Y}_i\}}$$

where

$$\frac{1}{Var\{\overline{Y}_i\}} = \frac{N_i}{N_i \cdot Var\{A'(\theta)\} + \mathbb{E}\{A''(\theta)\}}\,.$$

Of course, this estimator is only theoretical in that it is a function of the model nuisance parameters $Var\{A'(\theta)\}$ and $\mathbb{E}\{A''(\theta)\}$. Either these parameters or $Var\{\overline{Y}_i\}$ must, in turn, be sufficiently estimated. This chapter concludes with a presentation of a method to estimate these parameters, and then it is demonstrated via simulation that the resulting estimator

performs nearly as well as if the nuisance parameters were known. Under the model of this chapter, the efficient score is

$$\sum_{i=1}^{m} \mathbb{E}\big\{\psi(\theta_i)\big|\bar{Y}_i\big\}$$

where $\psi(\theta)$ is a mean zero function that satisfies, for all θ^*:

$$\sum_{i=1}^{m} \mathbb{E}\Big\{\mathbb{E}\big\{\psi(\theta_i)\big|\bar{Y}_i\big\}\Big|\theta_i = \theta^*\Big\} = c\big(A'(\theta^*) - \mu\big)\,.$$

When $N_1, \ldots N_m$ are identical, this equation is solved, and the efficient score corresponds to the estimator $\hat{\mu}_3 = \hat{\mu}_1 = \hat{\mu}_2$. When $N_1, \ldots N_m$ are not identical, we do not have a general solution to the equation for all π. A solution is found for π in only a subset of the parameter space — the conjugate family of distributions. In this region, the efficient score corresponds to the estimator $\hat{\mu}_3$, the weighted average of $\bar{Y}_i$ with weights equal to the inverse of the variance of $\bar{Y}_i$. This may be the efficient estimator outside of this region as well.

The efficient score can be obtained by projection of any mean-centered unbiased estimator, such as $\frac{1}{m}\sum_{i=1}^{m}(\bar{Y}_i - \mu)$, onto the tangent space. First the tangent space is established, and then the efficient score is found via projection.

Theorem 4.3. *Under the model of this chapter (Assumptions 4.1, N.4, and N.1), the tangent space is:*

$$\Big\{\sum_{i=1}^{m} \mathbb{E}\big\{\alpha(\theta_i)\big|\bar{Y}_i\big\} \;\; s.t. \;\; \mathbb{E}\{\alpha(\theta)\} = 0\Big\}$$

Proof. See Appendix A.7. □

Theorem 4.4. *The efficient score is the element* $\sum_{i=1}^{m} \mathbb{E}\big\{\psi(\theta_i)\big|\bar{Y}_i\big\}$ *where* $\psi(\theta)$ *is a mean zero function that satisfies, for all* θ^**,*

$$\sum_{i=1}^{m} \mathbb{E}\Big\{ \mathbb{E}\big\{ \psi(\theta_i) \big| \bar{Y}_i \big\} \Big| \theta_i = \theta^* \Big\} = c\big(A'(\theta^*) - \mu \big) \tag{4.2}$$

for some constant, c.

Proof. See Appendix A.8. □

The characterization of the efficient score in (4.2) involves means, averages and conditional expectations. Several results that relate these elements together are established next. These results lay the groundwork in order to solve the equation in certain cases, though they also serve as standalone results.

Averages and Conditional Expectation of the Mean

The conditional expectation of the mean parameter given $\overline{Y}$, $\mathbb{E}\big\{ A'(\theta) \big| \bar{Y} \big\}$, will sometimes (depending on π) be of the form $a\bar{Y} + b$ (where a and b aren't functions of $\bar{Y}$). Ericson [1969] showed that if $\mathbb{E}\big\{ A'(\theta) \big| \bar{Y} \big\}$ has this form, a and b are inflexible. In other words, if the "posterior" expectation of the mean parameter is to be linear in $\overline{Y}$, there is only one line it can be. In a subsequent paper, he showed that if π is a member of the usual conjugate prior family with respect to the distribution of $\bar{Y}$ given θ, then the posterior expectation will be of this linear form [Ericson, 1970]. This result was later formalized and strengthened by Diaconis and Ylvisaker [1979] to include the converse. Together these results prove that

$$\mathbb{E}\Big\{ A'(\theta) - \mu \Big| \bar{Y} \Big\} = \frac{Var\big\{ A'(\theta) \big\}}{Var\big\{ \overline{Y} \big\}} \Big(\bar{Y} - \mu \Big) \tag{4.3}$$

if and *only if* the distribution on θ is in the conjugate family with respect to the distribution of $\bar{Y}$ given θ. The family of conjugate distributions, $\{\pi(\theta)_{n^*,\tau}\}$, parameterized by τ and n^* is

$$\pi(\theta)_{n^*,\tau} = H(\tau, n^*)e^{n^*\tau\cdot\theta - n^*A(\theta)}\,d\theta\,. \tag{4.4}$$

where $n^* \in \mathbb{R}^+$ and $\tau \in \text{Interior}\{\mathcal{Y}\}$ where $\mathcal{Y}$ is the convex hull of the support set of $Y_{\cdot j}$. Diaconis and Ylvisaker [1979] show that these densities are indeed proper and that in this family,

$$\mu = \mathbb{E}\big\{A'(\theta)\big\} = \tau\,.$$

The following Lemma expresses $\mathbb{E}\big\{A'(\theta)\big|\bar{Y}\big\}$ in terms of $\bar{Y}$ for a wider class of densities on θ. This represents a generalization of Equation (4.3) which is true only in the conjugate prior case.

Lemma 4.5. *For any differentiable density $\pi(\theta)$ such that $\int \frac{d}{d\theta}(e^{y\cdot\theta - nA(\theta)}\pi(\theta))\,d\theta = 0$,*

$$\mathbb{E}\Big\{A'(\theta) - \mu\Big|\bar{Y}\Big\} = (\bar{Y} - \mu) + \frac{1}{N}\mathbb{E}\Big\{\frac{\pi'(\theta)}{\pi(\theta)}\Big|\bar{Y}\Big\}\,. \tag{4.5}$$

Proof. The result follows from an application of integration by parts. (See Appendix A.5). □

It will now be verified that Lemma 4.5 is consistent with Equation 4.3; that is, when π is in the conjugate family, Equation 4.5 becomes Equation 4.3. Note that in the conjugate family (4.4),

$$\frac{\partial}{\partial\theta}\pi(\theta) = n^*\big(\tau - A'(\theta)\big)\pi(\theta)\,.$$

Hence,

$$\frac{\pi'(\theta)}{\pi(\theta)} = -n^*\Big(A'(\theta) - \mu\Big)\,. \tag{4.6}$$

Because $\pi'(\theta)/\pi(\theta)$ is a multiple of $A'(\theta) - \mu$, Equation (4.5) can be re-written as

$$\mathbb{E}\Big\{A'(\theta) - \mu\Big|\bar{Y}\Big\} = \frac{N}{(N+n^*)}\Big(\bar{Y} - \mu\Big)\,.$$

It is left to show that

$$\frac{N}{(N+n^*)} = \frac{Var\big\{A'(\theta)\big\}}{Var\big\{\overline{Y}\big\}}\,.$$

For any differentiable density π on θ such that $\int \frac{d}{d\theta}(A'(\theta)\pi(\theta))\,d\theta = 0$, by integration by parts (Appendix A.6) it follows that

$$\mathbb{E}\big\{A'(\theta)\frac{\pi'}{\pi}(\theta)\big\} = -\mathbb{E}\big\{A''(\theta)\big\}\,. \tag{4.7}$$

In the conjugate family, this regularity condition holds [Diaconis and Ylvisaker, 1979], and also, $\pi'(\theta)/\pi(\theta) = -n^*\big(A'(\theta) - \mu\big)$. Hence, for π in the conjugate family,

$$n^* Var\big\{A'(\theta)\big\} = \mathbb{E}\big\{A''(\theta)\big\}\,. \tag{4.8}$$

Therefore,

$$\begin{aligned}\frac{Var\Big\{A'(\theta)\Big\}}{Var\Big\{\overline{Y}\Big\}} &= \frac{Var\Big\{A'(\theta)\Big\}}{Var\Big\{A'(\theta)\Big\} + \frac{1}{N}\mathbb{E}\big\{A''(\theta)\big\}} \\ &= \frac{1}{1+\frac{n^*}{N}} \qquad \text{(by 4.8)} \\ &= \frac{N}{N+n^*}\end{aligned}$$

as required.

Lemma 4.6. *For any distribution π on θ, if any function $g(\theta) \in L^2$ has the property that*

$$\mathbb{E}\big\{g(\theta) - \mathbb{E}\{g(\theta)\}\big|\bar{Y}\big\} = a\big(\bar{Y}-\mu\big)\,, \tag{4.9}$$

then

$$a = \frac{\mathrm{Cov}\big(g(\theta), A'(\theta)\big)}{Var\big\{\overline{Y}\big\}}\,.$$

Proof.

$$\begin{aligned}\mathbb{E}\Big\{\big(g(\theta) - \mathbb{E}\{g(\theta)\}\big)\big(\bar{Y}-\mu\big)\Big\} &= \mathbb{E}\Big\{\big(g(\theta) - \mathbb{E}\{g(\theta)\}\big)\mathbb{E}\big\{\bar{Y}-\mu\big|\theta\big\}\Big\} \\ &= \mathbb{E}\Big\{\big(g(\theta) - \mathbb{E}\{g(\theta)\}\big)\big(A'(\theta)-\mu\big)\Big\} \\ &= \mathrm{Cov}\big(g(\theta), A'(\theta)\big)\,.\end{aligned}$$

But also,

$$\begin{aligned}\mathbb{E}\Big\{\big(g(\theta)-\mathbb{E}\{g(\theta)\}\big)\big(\bar{Y}-\mu\big)\Big\} &= \mathbb{E}\Big\{\mathbb{E}\big\{g(\theta)-\mathbb{E}\{g(\theta)\}\big|\bar{Y}\big\}\big(\bar{Y}-\mu\big)\Big\} \\ &= \mathbb{E}\Big\{a\big(\bar{Y}-\mu\big)\big(\bar{Y}-\mu\big)\Big\} \qquad \text{(Assumption (4.9))} \\ &= aVar\big\{\overline{Y}\big\}\,.\end{aligned}$$

The result follows from equating these two expressions of $\mathrm{Cov}\big(g(\theta),\bar{Y}\big)$. □

With these results available, we now return to the efficient score equation (4.2), and consider its solution by cases.

All N_i are the same

Suppose all N_i are equal to N. Lemma 4.5 can be rewritten as

$$\mathbb{E}\Big\{A'(\theta)-\mu-\frac{1}{N}\frac{\pi'}{\pi}\big(\theta\big)\Big|\bar{Y}\Big\}=\bar{Y}-\mu\,.$$

The function

$$\psi(\theta)=\Big(A'(\theta)-\mu-\frac{1}{N}\frac{\pi'}{\pi}\big(\theta\big)\Big)$$

solves equation (4.2) for any π that satisfy the regularity conditions of Lemma 4.5. Note that because N does not vary by subject, the function ψ can contain N. The resulting efficient score is

$$\frac{1}{Var\big\{\overline{Y}\big\}}\sum_{i=1}^{m}\big(\bar{Y}_i-\mu\big)\,.$$

Thus, conforming to intuition, the efficient estimator when all N_i are equal is the average of the $\bar{Y}_i$, $i = 1, \ldots, m$ with an equal weighting across subjects, reflecting the now symmetric nature of the problem.

Conjugate Distribution

Suppose all $N_1, \ldots, N_m$ are not necessarily equal, but π is contained in the subset of the parameter space that is the conjugate family $\{\pi(\theta)_{n^*,\tau}\}$ where

$$\pi(\theta)_{n^*,\tau} = H(\tau, n^*) e^{n^*\tau\cdot\theta - n^* A(\theta)}\, d\theta\,.$$

From Equation (4.3) we have that

$$\mathbb{E}\Big\{ A'(\theta) - \mu \Big| \bar{Y} \Big\} = \frac{Var\big\{A'(\theta)\big\}}{Var\big\{\bar{Y}\big\}} \Big(\bar{Y} - \mu\Big)\,.$$

From this, it follows that

$$\psi(\theta) = \frac{A'(\theta) - \mu}{Var\big\{A'(\theta)\big\}}$$

solves the score equation (4.2). The resulting efficient score is

$$\sum_{i=1}^{m} \frac{1}{Var\big\{\bar{Y}_i\big\}} \Big(\bar{Y}_i - \mu\Big)$$

which produces the inverse-variance weighted estimator, $\hat{\mu}_3$.

General Distributions

We do not have a solution to the score equation (4.2) for any distribution π when $N_1, \ldots, N_m$ are nonidentical. However, suppose that a function $g_\pi(\theta)$ exists that does not depend on N such that

$$\mathbb{E}\big\{g(\theta) - \mathbb{E}\{g(\theta)\}\big|\bar{Y}\big\} = a\big(\bar{Y} - \mu\big)\,.$$

Then,

$$\psi(\theta) = \frac{A'(\theta) - \mu}{\mathrm{Cov}\big(g(\theta), A'(\theta)\big)}$$

would solve the score equation (4.2), and by Lemma 4.6 the corresponding estimator would be the inverse-variance weighted estimator. Note that $g(\theta)$ need not be identified. It would only need to be established that such a $g(\theta)$ exists.

Chapter 5

Simulation

The previous chapter provided efficiency results, under independence, for the theoretical estimator that is the inverse-variance weighted average of subject averages, $\overline{Y}_i$:

$$\hat{\mu}_3 = \Big(\sum_{i=1}^{m} w_i\Big)^{-1} \sum_{i=1}^{m} w_i \overline{Y}_i$$

where

$$w_i = Var\big\{\overline{Y}_i \big| N_i\big\}^{-1} = \Big(Var\big\{A'(\theta)\big\} + \frac{1}{N_i}\mathbb{E}\big\{A''(\theta)\big\}\Big)^{-1}. \tag{5.1}$$

This estimator is theoretical because it is a function of two unknown nuisance parameters that depend on π: $Var\big\{A'(\theta)\big\}$ and $\mathbb{E}\big\{A''(\theta)\big\}$. In order for this estimator to be implemented, the weights, w_i, $i = 1, \ldots, m$, must be replaced with estimated weights $\hat{w}_i$, $i = 1, \ldots, m$. This chapter addresses the matter of estimating the weights in the inverse-variance estimator. A method of estimating the weights is provided, and results on the efficiency of the

corresponding estimator are obtained through simulation. The main result of the simulation is that using estimated weights performs nearly as well as if the true weights are known.

The following approach to estimating $Var\{A'(\theta)\}$ and $\mathbb{E}\{A''(\theta)\}$ is similar to Robbins [1985] where estimates of these parameters are also required. Let

$$s_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(Y_{ij} - \overline{Y}_i)^2\,, \qquad s^2 = \frac{1}{\tilde{m}}\sum_{\{i:N_i>1\}} s_i^2$$

$$u^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\overline{Y}_i - \hat{\mu}_1)^2\,, \qquad l = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N_i}\,.$$

Note that in the definition of s^2, subjects with only one observation are not included in the average ($\tilde{m}$ is the the number of remaining subjects). Under independence between $A'(\theta)$ and N, dropping these subjects does not introduce any bias, and it follows that $\mathbb{E}\{s^2\} = \mathbb{E}\{A''(\theta)\}$. Furthermore, because $\mathbb{E}\{u^2\} = Var\{\overline{Y}\}$ and $\mathbb{E}\{l\} = \mathbb{E}\{1/N\}$, we have that

$$\mathbb{E}\{u^2 - ls^2\} = Var\{A'(\theta)\}\,.$$

Regardless of the conditional distribution of subject observations, $u^2 - ls^2$ and s^2 are unbiased estimators of the two parameters involved in the weights. It's reasonable then to employ these estimators to estimate the weights in (5.1). Let

$$\hat{w}_i = \left((u^2 - ls^2)^+ + \frac{1}{N_i}s^2\right)^{-1}$$

where $(\,\cdot\,)^+$ returns zero if its contents are negative, and is used because $u^2 - ls^2$ is estimating a non-negative parameter. Let $\hat{\mu}_3^*$ be the inverse-variance estimator with estimated

weights $\hat{w}_i$.

Results

In the simulation, we study the estimated inverse-variance estimator ($\hat{\mu}_3^*$), the true inverse-variance estimator ($\hat{\mu}_3^{tr}$), the estimator weighted proportionally to N_i ($\hat{\mu}_2$), and the uniformly weighted estimator ($\hat{\mu}_1$). The simulation was performed under different combinations of the distribution π, the distribution of N, and the sample size m. As in Chapter 4, N and θ are assumed independent. Conditionally given θ, subject sequences are taken to be Bernoulli trials with probability $A'(\theta)$.

The distribution of N is taken to put equal mass on the integers from 1 to an upper bound, inclusive. The upper bound is tested at the levels of 5 and 20. Five different distributions of $A'(\theta)$ are tested that have different shapes. Four of these are Beta distributions with different shapes: weakly unimodal, strongly unimodal, U-shaped, J-shaped. In addition, a bimodal distribution outside of the Beta family is also tested. The sample size is tested at the levels $m = 25$, $m = 100$, and $m = 400$.

For each of the 30 ($2 \times 5 \times 3$) combinations of model settings, a simulation was performed. At each simulation iteration, m independent and identically distributed realizations of subject triplets ($A'(\theta)$, N, $Y_{\cdot 1} \ldots Y_{\cdot N}$) were obtained. The estimators $\hat{\mu}_3^*$, $\hat{\mu}_3^{tr}$, and $\hat{\mu}_2$ were computed on the basis of the m pairs of (N, $Y_{\cdot 1} \ldots Y_{\cdot N}$). One million simulations were run at each combination. The simulation was coded in R.

The estimated bias of $\hat{\mu}_3^{tr}$ and $\hat{\mu}_3^*$ is presented in Table 5.1 . We know from theory that the true bias of $\hat{\mu}_3^{tr}$ is zero. The estimated bias of $\hat{\mu}_3^{tr}$ is helpful, however, to check the order of the error in the simulation-based estimates. As can be seen looking down the rows

of m, the simulation provides strong confirmation that the estimator $\hat{\mu}_3^*$ is asymptotically unbiased. This is to be expected as the employed estimators in the weights for $Var\{A'(\theta)\}$ and $\mathbb{E}\{A''(\theta)\}$ are consistent.

For smaller m, the simulation shows that $\hat{\mu}_3^*$ contains a minor amount of bias. Furthermore, it also reveals the direction of the bias is toward 0.5. Looking at the simulations we see that $\hat{\mathbb{E}}\{\hat{\mu}_3^*\} > \mu$ for $\mu < 0.5$, and $\hat{\mathbb{E}}\{\hat{\mu}_3^*\} < \mu$ for $\mu > 0.5$. By $m = 400$, the estimated bias and estimated standard deviation of $\hat{\mu}_3^*$ and $\hat{\mu}_3^{tr}$ coincide in the simulations, at least up to four decimal places. For $m = 400$ or larger, there is no practical difference between $\hat{\mu}_3^*$ and $\hat{\mu}_3^{tr}$.

For small m, the simulation reveals the existence of a bias-variance trade off in the estimator $\hat{\mu}_3^*$. This is apparent in the simulations at $m = 25$. At this sample-size, at each of the 30 combinations, the estimated standard deviation of $\hat{\mu}_3^*$ is always no greater than the estimated standard deviation of $\hat{\mu}_3^{tr}$ despite $\hat{\mu}_3^{tr}$ having weights the are unbeatable among exactly unbiased estimators.

Note we see in the table the property mentioned in Chapter 4: for a fixed ratio of α and β (which is to say a fixed μ) as $\alpha, \beta \to \infty$ (as the variance of π goes to zero), the standard deviation of $\hat{\mu}_3^{tr}$ converges to the standard deviation of $\hat{\mu}_2$. Likewise, for a fixed ratio of α and β (which is to say a fixed μ) as $\alpha, \beta \to 0$ (as the variance of π is maximized holding μ constant), the standard deviation of $\hat{\mu}_3^{tr}$ converges to the standard deviation of $\hat{\mu}_1$.

In conclusion, if π is unknown, for large m, $\hat{\mu}_3^*$ behaves nearly identically to $\hat{\mu}_3^{tr}$ and is therefore demonstrably better than $\hat{\mu}_1$ and $\hat{\mu}_2$ under any π. Even for m as small as 25, the degree of bias present in $\hat{\mu}_3^*$ is likely on a scale that is unimportant to the scientific question. Here too, however, the simulation shows that $\hat{\mu}_3^*$ has smaller standard deviation than either

$\hat{\mu}_1$ and $\hat{\mu}_2$. In concert, these results support using $\hat{\mu}_3^*$ to estimate μ if N and $A'(\theta)$ are independent.

Table 5.1: Simulation Results

		$\widehat{\text{Bias}}(\hat{\mu}_3^*)$	$\widehat{\text{Bias}}(\hat{\mu}_3^{tr})$	$\widehat{\text{sd}}(\hat{\mu}_3^*)$	$\widehat{\text{sd}}(\hat{\mu}_3^{tr})$	$\widehat{\text{sd}}(\hat{\mu}_2)$	$\text{sd}(\hat{\mu}_1)$
$A'(\theta) \sim \text{Beta}(2,3)$							
	$m = 25$	0.0010	0.0000	0.0670	0.0673	0.0681	0.0725
$N \sim U(1,5)$	$m = 100$	0.0003	0.0000	0.0336	0.0336	0.0340	0.0362
	$m = 400$	0.0001	0.0000	0.0168	0.0168	0.0170	0.0181
	$m = 25$	0.0005	−0.0000	0.0509	0.0510	0.0534	0.0551
$N \sim U(1,20)$	$m = 100$	0.0002	0.0000	0.0255	0.0255	0.0267	0.0276
	$m = 400$	0.0001	0.0000	0.0127	0.0127	0.0133	0.0138
$\text{Beta}(10,15)$							
	$m = 25$	0.0008	−0.0000	0.0594	0.0596	0.0596	0.0677
$N \sim U(1,5)$	$m = 100$	0.0003	0.0001	0.0297	0.0297	0.0297	0.0339
	$m = 400$	0.0001	0.0000	0.0148	0.0148	0.0149	0.0169
	$m = 25$	0.0006	0.0001	0.0368	0.0367	0.0370	0.0451
$N \sim U(1,20)$	$m = 100$	0.0002	0.0000	0.0183	0.0183	0.0185	0.0225
	$m = 400$	0.0000	−0.0000	0.0091	0.0091	0.0092	0.0113
$\text{Beta}(3,1)$							
	$m = 25$	−0.0023	0.0000	0.0610	0.0611	0.0620	0.0651
$N \sim U(1,5)$	$m = 100$	−0.0006	0.0000	0.0305	0.0305	0.0310	0.0326
	$m = 400$	−0.0002	−0.0000	0.0152	0.0152	0.0155	0.0163
	$m = 25$	−0.0011	0.0000	0.0477	0.0477	0.0503	0.0508
$N \sim U(1,20)$	$m = 100$	−0.0003	−0.0000	0.0238	0.0238	0.0251	0.0254
	$m = 400$	−0.0001	−0.0000	0.0119	0.0119	0.0126	0.0127
$\text{Beta}(0.2,0.6)$							
	$m = 25$	0.0011	−0.0000	0.0743	0.0745	0.0787	0.0754
$N \sim U(1,5)$	$m = 100$	0.0003	0.0000	0.0373	0.0373	0.0394	0.0377
	$m = 400$	0.0001	0.0000	0.0186	0.0186	0.0197	0.0189
	$m = 25$	0.0006	0.0001	0.0684	0.0685	0.0757	0.0690
$N \sim U(1,20)$	$m = 100$	0.0001	−0.0000	0.0342	0.0342	0.0379	0.0345
	$m = 400$	0.0000	−0.0000	0.0171	0.0171	0.0189	0.0173
$.5\text{Beta}(8,24) + .5\text{Beta}(24,8)$							
	$m = 25$	−0.0001	−0.0001	0.0739	0.0742	0.0760	0.0778
$N \sim U(1,5)$	$m = 100$	−0.0000	−0.0000	0.0370	0.0371	0.0379	0.0389
	$m = 400$	−0.0000	−0.0000	0.0185	0.0185	0.0190	0.0194
	$m = 25$	0.0000	0.0000	0.0609	0.0610	0.0652	0.0635
$N \sim U(1,20)$	$m = 100$	−0.0000	−0.0000	0.0304	0.0304	0.0325	0.0318
	$m = 400$	−0.0000	−0.0000	0.0152	0.0152	0.0163	0.0159

Bibliography

Brown, M. (2006). Exploiting The Waiting Time Paradox: Applications Of The Size-Biasing Transformation. Probab. Eng. Inf. Sci. 20, 2, 195-230.

Brown, G. P. and Shine, R. (2009). Beyond size-number trade-offs: clutch size as a maternal effect. Philosophical Transactions of the Royal Society B: Biological Sciences. 364(1520): 1097-1106.

Cook, R. and Lawless, J. (2007). The Statistical Analysis of Recurrent Events. Springer. New York, NY, USA.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate Priors for Exponential Families. Annals of Statistics. 7:269 - 281.

Efron, B. (2010). Large-Scale Inference Cambridge University Press. Cambridge, UK.

Efron, B. (2015). The Bayes Deconvolution Problem. Technical report (Stanford University. Department of Statistics). URL http://statweb.stanford.edu/ ckirby/brad/papers/2015BayesDeconvolutionProblem.pdf

Ericson, W. A. (1969) A Note on the Posterior Mean of a Population Mean. Journal of the Royal Statistical Society 31, no. 2: 332 - 334.

Ericson, W. A. (1970). On the Posterior Mean and Variance of a Population Mean. Journal of the American Statistical Association. 65.330: 649 - 652.

Pendlebury C. J. and Bryant, D. M. (2005). Effects of Temperature Variability on Egg Mass and Clutch Size in Great Tits. The Condor, vol. 107, no. 3, 710-714.

Postma, E. and van Noordwijk, A. J. (2005). Genetic Variation for Clutch Size in Natural Populations of Birds from a Reaction Norm Perspective. Ecology, 86: 2344-2357.

Harris, K. M. (2009). The National Longitudinal Study of Adolescent to Adult Health (Add Health), Waves I & II, 1994 – 1996; Wave III, 2001 – 2002; Wave IV, 2007 – 2009 [machine-readable data file and documentation]. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.

Harris, K.M., C.T. Halpern, E. Whitsel, J. Hussey, J. Tabor, P. Entzel, and J.R. Udry. (2009). The National Longitudinal Study of Adolescent to Adult Health: Research Design [WWW document]. URL: http://www.cpc.unc.edu/projects/addhealth/design.

Harris, K. (2013). Design features of Add Health. [WWW document]. http://www.cpc.unc.edu/projects/addhealth/data/guides/design%20paper%20WI-IV.pdf

Harris, K. M. and Udry, J. R. (2016). National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2008 [Public Use]. Chapel Hill, NC: Carolina Population Center, University of North Carolina-Chapel Hill/Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributors].

Narasimhan, B and Efron, B (2016). A g-modeling Program for Deconvolution and Empirical

Bayes Estimation. Technical report (Stanford University. Department of Statistics). URL http://statweb.stanford.edu/ ckirby/brad/papers/2016G-ModelingProgram.pdf

Robbins, H. (1985). Linear Empirical Bayes Estimation of Means and Variances. Proc. Nail. Acad. Sci. USA 82(6): 1571-1574.

Rosenfeld, M. and Kim, B. (2005). The Independence of Young Adults and the Rise of Interracial and Same-Sex Unions. American Sociological Review. 70:541-62.

Ross, S. M. (2003). The Inspection Paradox. Probab. Eng. Inf. Sci. 17(1): 47-51.

Shao, J. (2003). Mathematical Statistics. Springer Texts in Statistics. New York, NY, USA.

Song, P. X. (2007). Correlated Data Analysis: Modeling, Analytics, and Applications Springer. New York, NY, USA.

Tsiatis, A. (2006). Semiparametric Statistics and Missing Data Springer. New York, NY, USA.

van der Vaart, A. W. (1998). Asymptotic Statistics. Cambridge University Press, Cambridge, UK.

Appendix

A.1 Efficiency in Fixed and Common Parameter Model

In this section, the estimator

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{m} Y_i}{\sum_{i=1}^{m} N_i}$$

is proved to be efficient for estimating μ in a model where $\theta_1, \ldots, \theta_m$ are not subject-specific random effects but instead fixed parameters that are assumed to be all equal:

$$\theta_1 = \cdots = \theta_m = \theta\,.$$

In the framework of this dissertation, this can be thought of as a model in which $\pi(\theta)$ is a dirac density function at some point θ.

As before,

$$Y_{\cdot 1}, Y_{\cdot 2}, \ldots \,|\, \theta \overset{iid}{\sim} f_*^{\theta}(y)\,. \tag{N.1}$$

but here let $f_*^\theta(y)$ be the wider class of distributions, the exponential-dispersion family:

$$f_*^\theta(y) = c(y, \sigma^{-2}) \exp\left[\frac{1}{\sigma^2}\{\theta y - A(\theta)\}\right].$$

Note that in these models,

$$\mu = A'(\theta) \qquad \text{and} \qquad Var\{Y_{ij}\} = \sigma^2 A''(\theta).$$

Lemma A.1. *The estimator $\hat{\mu}_2$ is equal to the inverse-variance weighted average of the subject-specific averages.*

Proof. Let V_i be the variance of $\bar{Y}_i$.

$$V_i = \frac{\sigma^2 A''(\theta)}{N_i} \qquad \text{therefore,} \qquad \frac{1}{V_i} \propto N_i.$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{m} Y_i}{\sum_{k=1}^{m} N_k} = \Big(\sum_{k=1}^{m} N_k\Big)^{-1} \sum_{i=1}^{m} N_i \frac{Y_i}{N_i} = \Big(\sum_{k=1}^{m} \frac{1}{V_k}\Big)^{-1} \sum_{i=1}^{m} \frac{1}{V_i}\frac{Y_i}{N_i}$$

□

Lemma A.2. *The estimator $\hat{\mu}_2$ is efficient for estimating μ.*

Proof. The likelihood of the observed sequence from the i^{th} subject is

$$c(Y_{i1}, \ldots, Y_{iN_i}, \sigma^{-2}) \exp\left[\frac{N_i}{\sigma^2}\{\theta \bar{Y}_i - A(\theta)\}\right].$$

It follows that the likelihood of the entire data, Y_{ij} , $j = 1, \ldots, N_i$, $i = 1, \ldots, m$ can be

written as,

$$c(Y, \sigma^{-2}) \exp\left[w_+\{\theta\hat{\mu}_2 - A(\theta)\}\right].$$

where $w_i = N_i/\sigma^2$, and $w_+ = w_1 + \ldots w_m$.

That $\hat{\mu}_2$ attains the Information Bound is a consequence of the fact that

$$\frac{\partial}{\partial\theta} \log L = \frac{\hat{\mu}_2 - A'(\theta)}{w_+}.$$

□

A.2 Lemma 3.6

The proof of Lemma 3.6 relied on the following result: If h is permutation invariant, then,

$$\mathbb{E}\Big\{h\big((Y_1, N_1)\ldots(Y_m, N_m)\big)\Big|(N_1, \theta_1), (N_2, \theta_2)\ldots(N_m, \theta_m)\Big\}$$

depends on $(N_1, \theta_1), (N_2, \theta_2)\ldots(N_m, \theta_m)$ through the collection,

$$\Big\{(N_1, \theta_1), (N_2, \theta_2)\ldots(N_m, \theta_m)\Big\}.$$

To prove this result, it suffices to show that for any permutations σ^* and σ' of the indices i, from 1 to m,

$$\mathbb{E}\Big\{h\big((Y_{\sigma_1^*}, N_{\sigma_1^*})\ldots,(Y_{\sigma_m^*}, N_{\sigma_m^*})\big)\Big\} = \mathbb{E}\Big\{h\big((Y_{\sigma_1'}, N_{\sigma_1'})\ldots,(Y_{\sigma_m'}, N_{\sigma_m'})\big)\Big\}$$

Proof. Let σ^* be a permutation of the indices.

$$\begin{aligned}
&\mathbb{E}\Big\{h\big((Y_{\sigma_1^*}, N_{\sigma_1^*}) \ldots, (Y_{\sigma_m^*}, N_{\sigma_m^*})\big)\Big\} \\
=&\mathbb{E}\Big\{\frac{1}{m!}\sum_{\sigma} h\big((Y_{\sigma(\sigma_1^*)}, N_{\sigma(\sigma_1^*)}) \ldots, (Y_{\sigma(\sigma_m^*)}, N_{\sigma(\sigma_m^*)})\big)\Big\} \\
=&\mathbb{E}\Big\{\frac{1}{m!}\sum_{\sigma} h\big((Y_{\sigma_1}, N_{\sigma_1}) \ldots, (Y_{\sigma_m}, N_{\sigma_m})\big)\Big\}
\end{aligned}$$

where the first equality follows from the assumption that h is permutation invariant, and the second equality follows from the fact that addition is commutative and permutations are bijections. □

A.3 Lemma 3.7

Here it is verified that

$$h^*(Y_i, N_i) = m\mathbb{E}\big\{\hat{\mu}\big|Y_i, N_i, \mathcal{Z}\big\} - \sum_{k\neq i} A'(\theta_k)$$

satisfies

$$\mathbb{E}\Big\{\sum_{i=1}^{m} f_i(Y_i, N_i)\Big[\hat{\mu} - \frac{1}{m}\sum_{j=1}^{m} h^*(Y_j, N_j)\Big]\Big|\mathcal{Z}\Big\} = 0\,.$$

This is equivalent to satisfying

$$\sum_{i=1}^{m} \mathbb{E}\Big\{ f_i(Y_i, N_i)\hat{\mu} \Big| \mathcal{Z} \Big\} = \sum_{i=1}^{m}\sum_{j=1}^{m} \frac{1}{m} \mathbb{E}\Big\{ f_i(Y_i, N_i) h^*(Y_j, N_j) \Big| \mathcal{Z} \Big\} \tag{2}$$

$$= \sum_{i=j=1}^{m} \frac{1}{m} \mathbb{E}\Big\{ f_i(Y_i, N_i) h^*(Y_j, N_j) \Big| \mathcal{Z} \Big\} \tag{3}$$

$$+ \sum\sum_{i \neq j} \frac{1}{m} \mathbb{E}\Big\{ f_i(Y_i, N_i) h^*(Y_j, N_j) \Big| \mathcal{Z} \Big\} . \tag{4}$$

If $i = j$,

$$\mathbb{E}\Big\{ f_i(Y_i, N_i) \mathbb{E}(\hat{\mu} | Y_i, N_i, \mathcal{Z}) \Big| \mathcal{Z} \Big\} = \mathbb{E}\big\{ f_i(Y_i, N_i)\hat{\mu} \big| \mathcal{Z} \big\}$$

and,

$$\frac{1}{m} \mathbb{E}\Big\{ f_i(Y_i, N_i) h^*(Y_i, N_i) \Big| \mathcal{Z} \Big\} = \mathbb{E}\big\{ f_i(Y_i, N_i)\hat{\mu} \big| \mathcal{Z} \big\} - \mathbb{E}\big\{ f_i(Y_i, N_i) \big| \mathcal{Z} \big\} \sum_{k \neq i} \frac{1}{m} A'(\theta_k)$$

Hence, the summation in (3) is equal to

$$\sum_{i=1}^{m} \mathbb{E}\Big\{ f_i(Y_i, N_i)\hat{\mu} \Big| \mathcal{Z} \Big\} - \sum\sum_{i \neq j} \mathbb{E}\big\{ f_i(Y_i, N_i) \big| \mathcal{Z} \big\} \frac{1}{m} A'(\theta_j) .$$

If $i \neq j$, then by the independence between subjects and the conditional expectation of $h_j^*(Y_j, N_j)$ (3.6), it follows that

$$\frac{1}{m} \mathbb{E}\Big\{ f_i(Y_i, N_i) h^*(Y_j, N_j) \Big| \mathcal{Z} \Big\} = \mathbb{E}\big\{ f_i(Y_i, N_i) \big| \mathcal{Z} \big\} \frac{1}{m} A'(\theta_j) \ ,$$

and the summation in line 4 is equal to

$$\sum_{i \neq j} \sum \mathbb{E}\big\{ f_i(Y_i, N_i) \big| \mathcal{Z} \big\} \frac{1}{m} A'(\theta_j) \,.$$

This establishes that the left side and right side of (2) are equal, and that h^* as defined solves the projection equation.

A.4 Asymptotic Distribution of Estimators

Lemma A.3. *Suppose the model of Chapter 4 under which $\theta \perp N$, then*

$$\sqrt{m} \begin{pmatrix} \hat{\mu}_1 - \mu \\ \\ \hat{\mu}_2 - \mu \end{pmatrix} \xrightarrow{\mathcal{L}} N\big(0, \Sigma\big)$$

where

$$\Sigma_{11} = \mathbb{E}\Big\{ \frac{1}{N} \Big\} \mathbb{E}\{A''(\theta)\} + Var\Big\{ A'(\theta) \Big\}$$

$$\Sigma_{22} = \frac{\mathbb{E}\{A''(\theta)\}}{\mathbb{E}\{N\}} + \frac{\mathbb{E}\{N^2\}}{\mathbb{E}\{N\}^2} Var\Big\{ A'(\theta) \Big\}$$

$$\Sigma_{12} = \frac{\mathbb{E}\{A''(\theta)\}}{\mathbb{E}\{N\}} + Var\Big\{ A'(\theta) \Big\} \,.$$

Proof. Let

$$\bar{X} = \begin{pmatrix} \frac{1}{m} \sum Y_i \\ \\ \frac{1}{m} \sum N_i \\ \\ \frac{1}{m} \sum \bar{Y}_i \end{pmatrix} \qquad g(s,t,u) = \begin{pmatrix} u \\ \\ s/t \end{pmatrix} \qquad \beta = \begin{pmatrix} \mu \mathbb{E}\{N\} \\ \\ \mathbb{E}\{N\} \\ \\ \mu \end{pmatrix}$$

By the mulivariate central limit theorem, $\sqrt{m}(\bar{X} - \beta) \xrightarrow{\mathcal{L}} N\big(0, \Sigma^*\big)$. Note that

$$g(\bar{X}) = \begin{pmatrix} \hat{\mu}_1 \\ \\ \hat{\mu}_2 \end{pmatrix}, \qquad \text{and} \qquad g(\beta) = \begin{pmatrix} \mu \\ \\ \mu \end{pmatrix}.$$

By the multivariate delta method,

$$\sqrt{m}\big(g(\bar{X}) - g(\beta)\big) \xrightarrow{\mathcal{L}} N\big(0, \nabla g(\beta)^T \Sigma^* \nabla g(\beta)\big).$$

which establishes the result, after calculation. □

A.5 Lemma 4.5

If $\pi(\theta)$ is any differentiable density with support (a, b) for $a, b \in \{\mathbb{R}, +\infty, -\infty\}$ such that

$$\lim_{\theta \to b^-} e^{y\cdot\theta - NA(\theta)}\pi(\theta) = \lim_{\theta \to a^+} e^{y\cdot\theta - NA(\theta)}\pi(\theta) \qquad \forall\ y, N$$

then,

$$\mathbb{E}\Big\{ A'(\theta) - \mu \Big| \bar{Y} \Big\} = (\bar{Y} - \mu) + \frac{1}{N}\mathbb{E}\Big\{ \frac{\pi'(\theta)}{\pi(\theta)} \Big| \bar{Y} \Big\}.$$

Proof. Let $p(\theta, \bar{y})$ be the joint density of $\bar{Y}$ and θ;

$$p(\theta, \bar{y}) = \pi(\theta) f_0(\bar{y}) e^{N\,[\theta\cdot\bar{y} - A(\theta)]},$$

and let $m(\bar{y})$ be the marginal density of $\bar{Y}$. By integration by parts,

$$\int \pi'(\theta) f_0(\bar{y}) e^{N\,[\theta\cdot\bar{y}-A(\theta)]}\,d\theta = p(\theta,\bar{y})\Big|_{a^+}^{b^-} - \int N\big(\bar{y}-A'(\theta)\big)p(\theta,\bar{y})\,d\theta\,. \tag{5}$$

Because it is assumed that π is such that the first term on the right side of (5) vanishes, and also

$$\int \pi'(\theta) f_0(\bar{y}) e^{N\,[\theta\cdot\bar{y}-A(\theta)]}\,d\theta = \int \frac{\pi'}{\pi}(\theta)p(\theta,\bar{y})\,d\theta\,,$$

it follows that

$$\int \frac{\pi'}{\pi}(\theta)p(\theta,\bar{y})\,d\theta = -\int N(\bar{y}-A'(\theta))p(\theta,\bar{y})\,d\theta$$

or,

$$\int \frac{\pi'}{\pi}(\theta)p(\theta,\bar{y})\,d\theta = N\int A'(\theta)p(\theta,\bar{y})\,d\theta - N\,\bar{y}\,m(\bar{y})\,.$$

Diving through by $m(\bar{y})$ and N yields,

$$\frac{1}{N}\mathbb{E}\big\{\frac{\pi'(\theta)}{\pi(\theta)}\big|\bar{Y}\big\} = \mathbb{E}\big\{A'(\theta)\big|\bar{Y}\big\} - \bar{Y}\,.$$

The result follows from rearranging terms and subtracting μ on both sides.

□

A.6 Proof of Equation 4.7

Lemma .1. *Suppose the support of θ is the set (a,b) where for $a,b\in\{\mathbb{R},+\infty,-\infty\}$. If $\pi(\theta)$ is differentiable, and*

$$\lim_{\theta\to b^-} A'(\theta)\pi(\theta) = \lim_{\theta\to a^+} A'(\theta)\pi(\theta)$$

then,

$$\mathbb{E}\{A'(\theta)\frac{\pi'}{\pi}(\theta)\} = -\mathbb{E}\{A''(\theta)\}.$$

Proof. Using integration by parts,

$$\int A''(\theta)\pi(\theta)\,d\theta = A'(\theta)\pi(\theta)\Big|_{a^+}^{b^-} - \int A'(\theta)\frac{\pi'}{\pi}(\theta)\pi(\theta)\,d\theta$$

As long as the first term on the right side vanishes (which has been assumed), then

$$\mathbb{E}\{A'(\theta)\frac{\pi'}{\pi}(\theta)\} = -\mathbb{E}\{A''(\theta)\}.$$

□

A.7 Theorem 4.3: Tangent Space

The semiparametric tangent space is the mean-square closure of all scores for parametric submodels. The log-likelihood of observed data is

$$\sum_{i=1}^{m} \log\left\{\int f(\bar{y}_i|n_i,\theta)\pi(\theta,\gamma)\,d\theta\right\}.$$

The score with respect to an arbitrary submodel parameterized by γ at the true model, located where $\gamma = \gamma_0$, is

$$\sum_{i=1}^{m} \frac{\partial}{\partial\gamma}\log\left[\left\{\int f(\bar{y}_i|n_i,\theta)\pi(\theta,\gamma)\,d\theta\right\}\right]\Bigg|_{\gamma=\gamma_0}$$

$$= \sum_{i=1}^{m} \frac{\int f(\bar{y}_i|n_i,\theta)\,\dfrac{\partial}{\partial\gamma}\pi(\theta,\gamma_0)\,d\theta}{\int f(\bar{y}_i|n_i,\theta)\pi(\theta,\gamma_0)\,d\theta}\,.$$

Dividing and multiplying by $\pi(\theta,\gamma_0)$ in the integral in the numerator

$$= \sum_{i=1}^{m} \frac{\int S_\gamma(\theta,\gamma_0) f(\bar{y}_i|n_i,\theta)\pi(\theta,\gamma_0)\,d\theta}{\int f(\bar{y}_i|n_i,\theta)\pi(\theta,\gamma_0)\,d\theta} = \sum_{i=1}^{m} \mathbb{E}\big\{S_\gamma(\theta_i)\big|\bar{Y}_i\big\}$$

Any mean zero function $\alpha(\theta)$ with finite variance is a score $S_\gamma(\theta)$ for some parametric submodel $\pi(\theta,\gamma)$. Therefore the entire tangent space is:

$$\Big\{\sum_{i=1}^{m} \mathbb{E}\big\{\alpha(\theta_i)\big|\bar{Y}_i\big\}\ \ \forall \text{ mean zero } \alpha(\theta) \in L^2\Big\}\,.$$

A.8 Theorem 4.4: Efficient Score

Unbiased estimators of μ all have the same projection onto the tangent space. The efficient score is a multiple of this projection. Here we prove that the projection of $\hat{\mu}_1 - \mu$ onto the tangent space is

$$\sum_{i=1}^{m} \mathbb{E}\big\{\psi(\theta_i)\big|\bar{Y}_i\big\}$$

where $\psi(\theta)$ is a mean zero function that satisfies, for all θ^*, that

$$\sum_{i=1}^{m} \mathbb{E}\Big\{\mathbb{E}\big\{\psi(\theta_i)\big|\bar{Y}_i\big\}\Big|\theta_i = \theta^*\Big\} = A'(\theta^*) - \mu\,.$$

Proof. Let

$$\sum_{i=1}^{m} \mathbb{E}\big\{\psi(\theta_i)\big|\bar{Y}_i\big\}$$

be the projection of $\hat{\mu}_1 - \mu$ onto the tangent space. That the projection is of this form is a consequence of Theorem 4.3. The projection also satisfies, by definition,

$$\mathbb{E}\Big\{\Big[(\hat{\mu}_1 - \mu) - \sum_{i=1}^{m} \mathbb{E}\big\{\psi(\theta_i)\big|\bar{Y}_i\big\}\Big] \sum_{i=1}^{m} \mathbb{E}\big\{\alpha(\theta_i)\big|\bar{Y}_i\big\}\Big\} \;\; = \;\; 0 \;\; \forall \text{ mean zero } \alpha(\theta) \in L^2.$$

Rewritten:

$$\frac{1}{m}\mathbb{E}\Big\{\sum_{i=1}^{m}(\bar{Y}_i - \mu) \sum_{i=1}^{m} \mathbb{E}\big\{\alpha(\theta_i)\big|\bar{Y}_i\big\}\Big\} = \mathbb{E}\Big\{\sum_{i=1}^{m} \mathbb{E}\big\{\psi(\theta_i)\big|\bar{Y}_i\big\} \sum_{i=1}^{m} \mathbb{E}\big\{\alpha(\theta_i)\big|\bar{Y}_i\big\}\Big\}. \qquad (6)$$

Note that

$$\begin{aligned}
\frac{1}{m}\mathbb{E}\Big\{\sum_{i=1}^{m}(\bar{Y}_i - \mu) \sum_{i=1}^{m} \mathbb{E}\big\{\alpha(\theta_i)\big|\bar{Y}_i\big\}\Big\} &= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\Big\{(\bar{Y}_i - \mu)\mathbb{E}\big\{\alpha(\theta_i)\big|\bar{Y}_i\big\}\Big\} \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\Big\{(\bar{Y}_i - \mu)\,\alpha(\theta_i)\Big\} \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\Big\{\mathbb{E}\big\{(\bar{Y}_i - \mu)\big|\theta_i\big\}\alpha(\theta_i)\Big\} \\
&= \mathbb{E}\Big\{(A'(\theta) - \mu)\,\alpha(\theta)\Big\} \;\; .
\end{aligned}$$

On the right side of (6),

$$\mathbb{E}\Big\{\sum_{i=1}^{m} \mathbb{E}\big\{\psi(\theta_i)\big|\bar{Y}_i\big\} \sum_{i=1}^{m} \mathbb{E}\big\{\alpha(\theta_i)\big|\bar{Y}_i\big\}\Big\} \;\; = \;\; \sum_{i=1}^{m} \mathbb{E}\Big\{\mathbb{E}\big\{\psi(\theta_i)\big|\bar{Y}_i\big\}\mathbb{E}\big\{\alpha(\theta_i)\big|\bar{Y}_i\big\}\Big\}.$$

Since,

$$\begin{aligned}\mathbb{E}\Big\{\mathbb{E}\big\{\psi(\theta)\big|\bar{Y}\big\}\mathbb{E}\big\{\alpha(\theta)\big|\bar{Y}\big\}\Big\} &= \mathbb{E}\Big\{\mathbb{E}\big\{\mathbb{E}\big\{\psi(\theta)\big|\bar{Y}\big\}\alpha(\theta)\big|\bar{Y}\big\}\Big\} \\ &= \mathbb{E}\Big\{\mathbb{E}\big\{\psi(\theta)\big|\bar{Y}\big\}\alpha(\theta)\Big\},\end{aligned}$$

$$\mathbb{E}\Big\{\sum_{i=1}^{m}\mathbb{E}\big\{\psi(\theta_i)\big|\bar{Y}_i\big\}\sum_{i=1}^{m}\mathbb{E}\big\{\alpha(\theta_i)\big|\bar{Y}_i\big\}\Big\} = \sum_{i=1}^{m}\mathbb{E}\Big\{\mathbb{E}\big\{\psi(\theta_i)\big|\bar{Y}_i\big\}\alpha(\theta_i)\Big\}.$$

Having re-written the two sides of the projection condition (6), ψ must satisfy, for every mean zero $\alpha(\theta) \in L^2$:

$$\mathbb{E}\Big\{\big(A'(\theta)-\mu\big)\,\alpha(\theta)\Big\} = \sum_{i=1}^{m}\mathbb{E}\Big\{\mathbb{E}\big\{\psi(\theta_i)\big|\bar{Y}_i\big\}\alpha(\theta_i)\Big\} \tag{7}$$

Taking $\alpha(\theta)$ to be indicator functions $\mathbb{1}_{\theta=\theta^*}$, this equation will be satisfied for all α if

$$A'(\theta^*)-\mu = \sum_{i=1}^{m}\mathbb{E}\big\{\mathbb{E}\big\{\psi(\theta_i)\big|\bar{Y}_i\big\}\big|\theta_i=\theta^*\big\} \qquad \forall\,\theta^*. \tag{8}$$

□