Innovative Statistical Models for Inference from Complex Design Surveys and Longitudinal Studies

by

Xiaobi Huang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2011

Doctoral Committee:

Associate Professor Michael R. Elliott, Chair
Professor Fred M. Feinberg
Professor Siobán D. Harlow
Professor Roderick J. A. Little
Research Associate Professor Timothy D. Johnson

To my parents and grandparents

ACKNOWLEDGEMENTS

First and foremost, I owe my deepest gratitude to my doctoral and academic advisor Dr. Michael Elliott, who has supported me throughout my Ph.D study with his infinite patience and knowledge. This dissertation would not have come together without his constant guidance and encouragement. I learned a lot from him. To members of my doctoral committee: I would like to thank Dr. Timothy Johnson for sharing his rich knowledge in Bayesian statistics, and for being a perpetual source of confidence and inspiration as my academic advisor. I gratefully thank Dr. Roderick Little for his valuable advice and insights that he generously shared with me during group discussions. I am much indebted to Dr. Siobán Harlow for engaging me in Women's Reproductive Study that I have enjoyed so much. I have benefited a lot by her advice and insights in epidemiology. I gratefully thank Dr. Fred Feinberg for kindly granting me his time to read and give constructive advice to this thesis.

I would also like to thank my master advisor Dr. Thomas Nichols for guiding me into brain imaging study and for his constant support and encouragement during my first two years in USA.

I would regret my doctoral years at the University of Michigan if I did not join and presiding STATCOM@UMICH, the student community outreach group. I cherish this wonderful experience. I greatly thank Dr. Michael Elliott for his effort and support to this group. I thank Maria Larkina for working with me together in organizing the group and Dr. Kerby Shedden for his advice in our projects.

I have been blessed with many friendships which embraced my life in Michigan: My special thanks to Yue Chao for his valuable suggestions in computation and for always being such a great support. I am grateful to Jian Kang for valuable discussions in a broad range of statistical problems. My sincere thanks to Dr. Diane Marriott for her mentorship and spiritual support. I thank Dr. Peter Song for sharing and updating stories of our common hometown. I would also like to thank Xiaohong Liu, Desmond Lo, Rui Xiao, Laura Fernandes, Ying Guo and many others with whom I have shared our laughter and tears.

Words cannot express my gratitude to my family for their unflagging love and support throughout my life. My parents spared no effort to provide the best possible environment for me to grow up and attend school. My grandparents have persistent confidence in me and took care of me during my entire high school years. I cannot ask for more from them.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER I

Introduction

This thesis develops innovative applications for making inference from data from multilevel probability and from longitudinal sample designs. The first application focuses on the problem of "weight trimming", which is sometimes utilized in settings where highly disproportional probability of selection or inclusion lead to highly variable weights. Traditional methods utilize an ad-hoc cutpoint, whereas more principled model-based approaches attempt to use the data to balance the decrease in variance against the potential increase in bias caused by the trimming of the extreme weights. Here we extend these methods to multi-level complex sample designs. The second application focuses on developing changepoint or "stick-breaking" models for mean and variance of a large longitudinal data set, in this case for the specific application of modeling menstrual cycle lengths in the TREMIN cohort I data set which follows the cycle lengths of 2350 women across their reproductive lifespan. A multilevel model is used to develop both subject-specific and population-based measures of menopausal transition at their late reproductive life. Further analysis of the subject-level estimates using K-medoid clustering algorithm allows us to classify women into different typologies of menopausal transition, providing insights as to when existing menopausal staging criteria perform well and when they fail.

Broadly speaking, these applications can be seen as a general application of "missing data" methods in multilevel settings, in the first case to obtain "complete data" inference from a finite population in the survey sample setting, or more traditionally in the second case from a sample with missing components due to sporadic reporting, dropout, or censoring. Joint methods to obtain inferences from trends in both first and second moments are considered in the second application as well.

1.1 Model based method of weight trimming in sample surveys

In population-based sample surveys, each sampled unit will represent a portion of the total population of interest. In many surveys, units within the population have unequal probability of being included in the sample, so an estimator simply based on the values of the sampled units may be biased with respect to the corresponding population quantity. We generally calculate a weight for each included unit equal to the inverse of the probability of inclusion. The probability of inclusion may be a known quantity, such as a probability of selection determined by the investigator, or an estimated quantity, such as probability of response when non-response is present, or the product of two or more known estimated probabilities. We then use these weights to reduce bias in the estimation of population quantities of interest. For example, the arithmetic mean of sampled units $\bar{\boldsymbol{y}}$ may be biased in estimating the population mean $\bar{\boldsymbol{Y}}$ if there is an association between the probability of inclusion and the values of $\boldsymbol{y}$. Instead we may use the asymptotically unbiased Hájek estimator (Basu 1971) $\bar{\boldsymbol{y}}_w = \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i}$, where w_i is the weight, y_i is the value of the ith sampled subject and s denotes the subset of population units sampled.

The Hájek estimator is an example of a design-based approach to estimate a population quantity $Q(\boldsymbol{Y})$. Design-based approaches treat population values $\boldsymbol{Y} =$

$(Y_1, \ldots, Y_N)$ as fixed and sampling indicators $\boldsymbol{I} = (I_1, \ldots, I_N)$ as random. The goal is to account for the sampling mechanism $\boldsymbol{I}$ in estimating the population quantity $Q(\boldsymbol{Y})$, which exists independently of the data collected. Design-based approaches use observed data from sampled subjects along with information about the sampling mechanism to develop estimators $\hat{q}(\boldsymbol{y}, \boldsymbol{I})$ and $\hat{v}(\hat{q}(\boldsymbol{y}, \boldsymbol{I}))$ which are at least approximately unbiased with respect to $p(\boldsymbol{I} \mid \boldsymbol{Y})$ for $Q(\boldsymbol{Y})$ and $Var(Q(\boldsymbol{Y}))$ respectively.

An alternative form of finite population inference is the model-based approach, which uses parametric models $p(\boldsymbol{y}|\boldsymbol{\theta})$ to obtain inference about population quantities $Q(\boldsymbol{Y})$. These models can be embedded in the Bayesian framework to use the posterior predictive distribution to draw inference about unobserved (un-sampled or non-response) elements of the population: $p(\boldsymbol{Y}_{nobs}|\boldsymbol{y})$.

Design-based and model-based approaches have complementary strengths and weaknesses. Design-based approaches achieve better robustness but can reduce or remove bias at a cost of very large variances. This is particularly likely when weights are highly variable, when the sample size is small or when the correlation between inclusion probability and data is weak.

Various methods have been proposed to balance the bias-variance tradeoff. Perhaps the most popular method is weight trimming (Potter 1990), a design-based approach, in which a pre-determined value w_0 is chosen and weights larger than w_0 are set equal to w_0. Alternative approaches including weight smoothing models (Holt and Smith 1979, Ghosh and Meeden 1986, Little 1991, 1993, Lazzeroni and Little 1998, Elliott and Little 2000, Elliott 2007) and weight pooling (averaging) models (Elliott and Little 2000, Elliott 2008, 2009) have been developed to induce weight trimming through a statistical model. These model-based approaches stratify the data by the population of inclusion. These strata are called "weight strata".

Weight smoothing models treat underlying strata means as random effects and allow estimates of strata means to borrow strength from each other to achieve weight trimming. Simple weight pooling models assume distinct stratum means for smaller weight strata and a common mean for the larger weight strata. Compound weight pooling models are also called weight averaging models, which average the estimators of a set of weight pooling models across all possible pooling patterns. In Chapter II of this paper, we focus on developing model-based approaches for weight trimming for sample surveys with clustered sample designs.

1.2 Analyzing patterns of women's menstruation history

One goal for women's menstrual studies is to identify associations between women's menstrual characters and women's health. We study women's menstruation because it is the most easily observed event associated with women's reproduction health. Harlow (1995) suggested that menstrual cycles are the most easily observed marker of changes of ovarian functions. They can provide important information for ovarian aging, changes of endocrine system and endocrine risk factors for chronic diseases.

Studies have also suggested that alterations in menstrual bleeding are a significant source of gynecologic morbidity, especially in women's late reproductive life (Harlow 1995, Harlow and Campbell 2004, Mansfield and Voda 1997). The menopausal transition is a critical period in women's late reproductive life. The physiological changes and the timing, duration and characteristics during this transition may define in part women's long term chronic disease risk profile.

The biggest challenge in studying women's menstruation is lack of precision. Previous studies have depended mostly on visual examinations and simple quantitative criteria, thus motivating us to develop statistical models to better quantify and cap-

ture women's menstruation patterns.

1.2.1 Modeling women's menstruation patterns

In order to study women's menstruation patterns, it is important to consider both the menstrual cycle lengths and the variability of the lengths. Complexities of limited information as well as menstruation physiology, hormone use and gynecological surgeries make it difficult to study menstrual patterns across individual woman's reproductive life.

Luckily, two high quality data sets provide rich information about women's menstruation across their reproductive life. One of these data sets is the Tremin trust data (Treloar et al. 1967), which comes from a 70-year, two-cohort longitudinal study.

Based on previous studies and observations (Treloar et al. 1967, Harlow 1995, Harlow et al. 2007, Lisabeth et al. 2004a), it is known as women approach menopause, the patterns of their menstrual segment lengths change. In order to study when transitions in menstruation happen, in Chapter III, we develop a Bayesian change-point model for mean segment lengths as well as for the variability of the lengths for their late reproductive life (after age 35) based on the Tremin trust data. Our model provides a thorough study of menstrual characteristics of menopausal transition including its timing, duration and patterns of cycle variability. Use of a hierarchical model allows us to estimate both population-level and individual-level menstrual cycle characteristics. Furthermore, our model provides a method to assess associations of women's menstrual cycle pattern with subject-level covariates. For example, the model would enable comparisons between two Tremin cohorts representing two generations of American women. These comparisons can provide valuable information about whether secular trends exist in characteristics of menopausal transition. Also, our model addresses two important methodological problems in the analysis of

menstrual calendar data - how to account for intermittent missing data and how to account for use of exogenous hormones or truncation by surgeries.

1.2.2 Comprehensive study of women's menstruation patterns

The Bayesian changepoint model in Chapter III captures individual woman's menstrual patterns in their late reproductive life by eight characteristics: mean segment length at age 35, mean slope of segment length before changepoint, mean slope of segment length after changepoint, mean changepoint age, log-variance of segment length at age 35, slope of log-variance before variability changepoint, slope of log-variance after variability changepoint, and variability changepoint age. To summarize the rich information provided, it is helpful to identify subgroups of women with similar menstruation patterns. Such summary will also allow us to link these characteristics to previously defined transition markers.

It is known that women's menstrual patterns are related to age at menopause (Weinstein et al. 2003). In Chapter IV, we use accelerated failure time models to assess the associations between our eight summary measures of menstrual cycle patterns and age at menopause, and to consider the joint predictive ability of these measures for age at menopause.

In 2001, the Stages of Reproductive Aging Workshop (STRAW) proposed a staging system for women's menopausal transition, including different markers to characterize early and late transition stages (Soules et al. 2001, Harlow et al. 2007). These markers were mainly based on quantitative measures of sliding windows over women's menstruation history. Our model in Chapter III detects mean and variance changepoints for individual woman which can serve as alternative measures for early and late menopausal transition. Comparing the model-based changepoints with the descriptive markers in Chapter IV shows that the changepoints are more accurate in

describing menopausal transition because they use all of the data from a woman's late reproductive life, avoiding situations such as when a woman with highly variable cycles has an outlying cycle length large enough to serve as a transition marker years before onset of her transition.

Gorrindo et al. (2007) created a detailed classification system to categorize women's menstruation patterns to five types based on quantitative measures of cycle lengths and their variability of sliding windows over age. They defined the quantitative measures based on visual examination and prior epidemiology knowledge. It would be interesting to include less prior information and let pattern features define their own categories. Thus, in Chapter IV, we use a k-medoids algorithm to develop a six-category classification based on the eight characteristics detected by our changepoint model. These categories differ by baseline variances as well as by early and late changepoints, differences in the length of the early-to-late transition times, and in the deliniation of the changepoint signals. These categories are helpful in determining whether a woman's descriptive marker age corresponds well with her true menopausal transition age, versus a woman whose descriptive marker does not contain real transition information.

1.3 An outline of the dissertation

This dissertation is organized as follows. In Chapter II, an empirical Bayesian mixed-effect model is developed to smooth over weight strata via random effects in order to achieve estimators with smaller RMSEs and relatively small bias in sample surveys with cluster design and highly variable case weights. I further apply this method to NHANES III data to study associations between LDL cholesterol level and birth weight. This study extends weight smoothing method for population

based sample surveys to surveys with clustered or cross-classed sample surveys. In Chapter III, I develop a Bayesian changepoint model that includes separate mean and variance parameters to describe the menstruation trajectory for each woman, together with a hierarchical model to link them together. The model is designed for TREMIN data. I include multiple imputation in the Bayesian estimation procedure to deal with different forms of the missingness in the data set. Our method well quantifies individual women's mean and variability changes over their late reproductive life. Our imputation enables us to include many subjects which were not included in previous researches, thus provide more information. In Chapter IV, I develop an algorithm based classification to define subgroups of women's menstruation patterns based on the characteristics of women's menstruation patterns captured by the changepoint model. I further analyze these characteristics and link them with age of menopause and previously defined transition markers, altogether as well as by category. Our classification is the first algorithm based classification of women's menstruation patterns. Chapter IV provides a comprehensive analysis for understanding the associations between women's menstruation patterns and time of menopause as well as other reproductive characteristics. We also propose two transition markers, mean and variance changepoints, and compare them with previously defined transition markers.

CHAPTER II

Weight Smoothing Models in Clustered or Cross-Classed Sample Designs

2.1 Introduction

2.1.1 Design-Based vs. Bayesian Model-Based Inference for Complex Sample Designs

In contrast to most other areas of statistics, randomization or 'design-based' inference is standard for data from complex sample survey designs. Design-based approaches treat population values $\boldsymbol{Y} = (Y_1, ..., Y_N)$ as *fixed*, and sampling indicators $\boldsymbol{I} = (I_1, ..., I_N)$ as *random*. In population-based inference, the goal is to make inference about a population quantity $Q(\boldsymbol{Y})$.

In broad summary, design-based approaches use the observed data $\boldsymbol{y} = (y_1, ..., y_n)$ to develop estimators $\hat{q}(\boldsymbol{y}, \boldsymbol{I})$ that are at least approximately unbiased for the population quantity of interest:

$$E_{\boldsymbol{I}|\boldsymbol{Y}}(\hat{q}(\boldsymbol{y}, \boldsymbol{I})) \approx Q(\boldsymbol{Y})$$

Similarly, variance estimators of $\hat{q}(\boldsymbol{y}, \boldsymbol{I})$ given by $\hat{v}(\boldsymbol{Y}_{inc}, \boldsymbol{I})$ are obtained where

$$E_{\boldsymbol{I}|\boldsymbol{Y}}(\hat{v}(\boldsymbol{y}, \boldsymbol{I})) \approx Var_{\boldsymbol{I}|\boldsymbol{Y}}(\hat{q}(\boldsymbol{y}, \boldsymbol{I})).$$

In both cases expectation is taken with respect to the sampling mechanism that generates $\boldsymbol{I}$, and the approximation is $O(n^{-1})$ or better (Hansen and Hurwitz 1943)(Kish 1965)(Cochran 1977).

Bayesian population inference is an alternative form of finite population inference that also focuses on inference about population quantities $Q(\boldsymbol{Y})$, but allows for the use of parametric models $p(\boldsymbol{Y} \mid \boldsymbol{\theta})$ for population data based on the posterior predictive distribution for the unobserved elements of the population $p(\boldsymbol{Y}_{nobs} \mid \boldsymbol{y})$:

$$
\begin{aligned}
p(\boldsymbol{Y}_{nobs} \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{Y})}{p(\boldsymbol{y})} \\
&= \frac{\int p(\boldsymbol{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{p(\boldsymbol{y})} \\
&= \frac{\int p(\boldsymbol{Y}_{nobs} \mid \boldsymbol{y}, \boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{p(\boldsymbol{y})} \\
&= \int p(\boldsymbol{Y}_{nobs} \mid \boldsymbol{y}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}
\end{aligned}
\tag{2.1}
$$

(Ericson 1969, Holt and Smith 1979, Little 1993, Rubin 1987, Scott 1977, Skinner and Smith 1979). (We use the term "parametric models" loosely here to include not only standard parametric models such as fixed-effect regression models but also semi- or non-parametric methods that can also be construed as highly parameterized models [e.g., B-splines].) Here the difficult task of developing a prior for the entire population distribution $p(\boldsymbol{Y})$ is simplified by using the model $p(\boldsymbol{Y} \mid \boldsymbol{\theta})$ and averaging over the posterior distribution based on the sampled data $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.

Design and Bayesian model-based approaches have complementary strengths and weaknesses, and debate about the merits of each approach continues. In the design-based approach $\boldsymbol{Y}$ is treated as fixed, which has the advantage of robustness, as no parametric assumptions are made about the data. Also, in probability samples the distribution of $\boldsymbol{I}$ is under the control of the investigator (at least to a large degree). Thus sample design is automatically accounted in inference, since the design determines the repeated sampling characteristics of $\boldsymbol{I}$.

But the design-based approach does not always work well (Kalton 2002). As Basu (1971)'s jocular example of the circus statistician estimating the weight of an elephant

troupe shows, unbiasedness of design-based estimators is sometimes purchased at the cost of very large variances, leading to unacceptably large mean square errors. Small area estimation (Ghosh and Lahiri 1988) typically requires some sort of sharing of information across subdomains of the population, which can only be accomplished with some form of modelling. A consistent reference distribution is lacking in the design-based approach, as illustrated by the fact that a poststratified estimator of a population mean from a simple random sample needs to condition on the sample sizes n_j within the $j = 1, ..., J$ poststrata to obtain a non-infinite variance estimate over the distribution of $\boldsymbol{I}$ (since there is a non-zero probability that some $n_j = 0$), whereas the sample mean that ignores poststratification is design-unbiased only if the n_j are treated as random (Little 2004).

In general, model-based approaches have the advantage of efficiency if the model reasonably approximates the data. However, implicit in model-based inference based on (2.1) is that the sampling indicator $\boldsymbol{I}$ need not be modeled. This requires:

1. $p(\boldsymbol{I} \mid \boldsymbol{Y}) = p(\boldsymbol{I} \mid \boldsymbol{Y}_{obs})$ and

2. $p(\boldsymbol{Y}_{nob} \mid \boldsymbol{Y}_{obs}, \boldsymbol{I}, \boldsymbol{\theta}) = p(\boldsymbol{Y}_{nob} \mid \boldsymbol{Y}_{obs}, \boldsymbol{\theta})$.

Assumption (1) requires the distribution of I to to depend only on Y_{obs} (note that independence from Y is a common special case) and is termed an "ignorable" sampling design (Rubin 1987), and is usually satisfied in probability samples. Assumption (2) requires a model for the data $p(\boldsymbol{Y} \mid \boldsymbol{\theta})$ that is attentive to design features and robust enough to sufficiently capture all aspects of the distribution of $\boldsymbol{Y}$, or, at a minimum, those relevant to $Q(\boldsymbol{Y})$. Hence Little (1983, 1991) argues that design-based properties should be taken into account in model formulation. Models should be restricted to the class that give rise to approximately *design consistent* estimators (An estima-

tor $\hat{\theta}_v$ of true parameter θ_v is consistent for a finite population under a given class of designs if for any fixed $\varepsilon > 0$, $\lim_{v\to\infty} Pr(|\hat{\theta}_v - \theta_v| > 0) = 0$ with v indicating the size of the population.). This approach provides optimal estimators if the model is true, and provides protection against model misspecification to the extent that the model fails.

2.1.2 Weight Trimming and Weight Smoothing

Analysis of survey data from unequal probability samples typically use case weights to provide unbiased linear estimators of population values such as population means or totals, or asymptotically unbiased non-linear estimators of population values such as population regression parameters (Binder 1983). Case weights are set equal to the inverse of the probability of selection: $w_i = 1/p(I_i = 1)$. Case weights may be generalized to be the inverse of an estimated probability of inclusion by incorporating non-response adjustments, which are equal to the inverse of the probability of response (Gelman and Carlin 2002, Oh and Scheuren 1983), or calibration adjustments, which constrain case weights so that weighted sample totals equal known population totals, either jointly, as in poststratification or generalized regression estimation, or marginally, as in generalized raking estimation (Deville and Sarndal 1992, Isaki and Fuller 1982). The Horvitz-Thompson estimator (Horvitz and Thompson 1952) of a population total is given by $\hat{T} = \sum_{i\in s} w_i y_i$; other estimators of population quantities (means, quantiles, regression slopes, etc.) are typically obtained by replacing sample totals with their Horvitz-Thompson estimators.

However, these fully-weighted estimators reduce or remove bias at a cost of increasing the variance. This increase can overwhelm the reduction in bias, so that the mean square error actually increases under a weighted analysis. This is particularly likely when the weights are highly variable, when the correlation between

the probability of selection and the data is weak, or when the sample size is small. The most common approach to deal with this problem is know as *weight trimming* (Potter 1990), in which weights larger than some pre-determined value ω_0 are set as ω_0 and the remaining weights adjusted upward by a constant so that the weighted sample size remains unchanged. This manipulation of weights reflects a traditional design-based approach to survey inference.

To accommodate unequal probability of selection in the model-based approach, one option is to treat the weights as stratifying variables, where strata are defined by the probability of inclusion (Little 1983, 1991, Rubin 1983). This yields a model that satisfies assumption (2) in Section 2.1.1, since the population distribution is independent of the sampling indicator distribution conditional on the stratum means. Standard weighted estimates are then obtained by treating the stratum means of survey outcomes as fixed effects. Weight trimming can be effectively achieved by treating the underlying stratum means as random effects (Holt and Smith 1979, Ghosh and Meeden 1986, Little 1991, 1993, Lazzeroni and Little 1998, Elliott and Little 2000, Elliott 2007). By treating the strata means as random variables, the model allows estimates of strata means to borrow strength from each other to obtain what Elliott (2007) termed *weight smoothing*. This approach is consistent with the Bayesian model emphasis on prediction. If there are outlying data elements with large weights that are poorly estimated, the hierarchial model will borrow strength from the rest of the data to reduce unnecessary variability; if these outlying data elements are well-estimated, there will be little smoothing of the means to preserve bias reduction.

A more general form of the Bayesian weight smoothing model is

$$y_{hi}|\mu_h \sim f(y_{hi}|\mu_h, \sigma^2)$$

$$\mu \sim N_h(\phi, G)$$

$$\phi, \mathbf{G} \sim \Pi$$

where $f(y_{hi}|\mu_h, \sigma^2)$ is an exponential family distribution with mean μ_h and scale parameter σ^2. Here h indexes the "weight strata" with constant or nearly constant inclusion probability, $\mu = (\mu_1, ..., \mu_H)$, $\phi = (\phi_1, ..., \phi_H)$. G is the covariance matrix for random effects of weight strata, and Π is the hyperprior distribution (Elliott 2007, Elliott and Little 2000, Lazzeroni and Little 1998). Under this model, the posterior mean of the population mean is given by

$$E(\bar{Y}|y) = \sum_h [n_h \bar{y}_h + (N_h - n_h)\hat{\mu}_h]/N_+$$

where n_h is the number of subjects of each strata in the sampled data, N_h is the number of subjects of each strata in the total population, N_+ is the size of the total population and $\hat{\mu}_h = E(\mu_h|y)$ is the posterior mean of the weight strata (Elliott and Little 2000). When $G \to 0$, borrowing across the weight strata is large and $E(\bar{Y}|y)$ tends toward the unweighted mean $\overline{y} = \frac{\sum_{i=1}^n y_i}{n}$. When $G \to \infty$, there is little borrowing across the weight strata and $E(\bar{Y}|y)$ tends toward the fully weighted mean $\overline{y}_w = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}$. We can adjust the extent of weight smoothing effect by fitting models with different structure of G. We can also adjust our modelling for $\hat{\mu}_h$ to make the fixed effects fit the true data better. It was found in Elliott and Little (2000) that when the variance of the data is large, weight smoothing models have smaller RMSEs compared to design based methods. The models with autoregressive or exchangeable covariance structures tend to reduce variances more than the non-parametric model or the model with a fixed linear effect. However, the non-parametric model was the most robust to a variety of mean structures, followed by the linear model.

Elliott (2007) generalized the findings of Elliott and Little (2000) to linear and

generalized linear models. For linear models, the posterior mean of the population regression parameter $\boldsymbol{B}$ is available in closed form and in negligible sampling fractions, which is given by

$$\hat{\boldsymbol{B}} = E(\boldsymbol{B}|y, X) = \left[\sum_h W_h \sum_{i=1}^{n_h} \boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T\right]^{-1} \left[\sum_h W_h \sum_{i=1}^{n_h} \left(\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T\right) \hat{\boldsymbol{\beta}}_h\right]$$

where $W_h = N_h/n_h$ and $\hat{\boldsymbol{\beta}}_h = E(\boldsymbol{\beta}_h \mid \boldsymbol{y})$. Elliott (2007) found that the exchangeable model and autoregressive model tends to be more biased when the variance of the data increases, while the linear model and nonparametric model are approximately unbiased. Similar to Elliott and Little (2000), Elliott (2007) also found that the exchangeable model and autoregressive model have larger reductions of RMSE than the linear model and the nonparametric model, but the linear model and the nonparametric model have better robustness.

In this paper, we extend the weight smoothing models of Elliott and Little (2000) and Elliott (2007) to accommodate clustered sample designs. As in other settings with clustered designs, we utilize random effects to induce correlation between subjects sampled within a cluster. A delicate aspect of this extension is that the random effects of the weight strata will often cross the random effects of the clusters, yielding "cross-classified" random effects models (Rasbash and Goldstein 1994). We develop specific models for population means and population linear regression parameters. Section 2.2 develops models for population means. Section 2.3 considers the empirical behavior of these models via simulation studies under a variety of scenarios that closely mimic the full complexity of all probability sample designs. Section 2.4 extends the clustered weight smoothing models to the linear model setting and considers an application to assess Barker's hypothesis (Barker et al. 1993) that cardiovascular disease development is associated with low birth weight using data from

the National Health and Nutrition Examination Survey (NHANES) III (U.S. Department of Health and Human Services 1997). Section 2.5 summarizes the results of the simulations and application and considers how these findings relate to previous work.

2.2 Weight Smoothing Models for Cluster Sample Designs

Previous work using random effects models to implement weight trimming have focused on models that are strictly applicable only to disproportionately stratified or post stratified samples without clustering. However, many sample surveys commonly have more complex sample designs, such as single or multi-stage cluster samples or strata that cross the weight strata. Here we develop models that accommodate both case weights and clustered sample designs. For estimation of the population means, the general form of the models we consider is

$$\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{a}^H, \boldsymbol{a}^Q, \boldsymbol{a}^{HQ} \sim N(X^H\boldsymbol{\mu} + Z^H\boldsymbol{a}^H + Z^Q\boldsymbol{a}_q^Q + Z^{HQ}\boldsymbol{a}_{hq}^{HQ}, \sigma^2 I_n)$$

$$\boldsymbol{a}^H \sim N(0, \sigma_h^2 I_H),\ \boldsymbol{a}^Q \sim N(0, \sigma_q^2 I_Q),\ \boldsymbol{a}^{HQ} \sim N(0, \sigma_{hq}^2 I_{HQ})$$

where $\boldsymbol{\mu}$ are fixed effects of population mean, $\boldsymbol{a}^H = (a_1^H, ..., a_H^H)^T$ are the random effects associated with the weight strata, $\boldsymbol{a}^Q = (a_1^Q, ..., a_Q^Q)^T$ are the random effects associated with the sample design clusters, and $\boldsymbol{a}^{HQ} = (a_{11}^{HQ}, ..., a_{HQ}^{HQ})^T$ are random effects associated with the cross section of clusters and strata to account for possible correlation between weight stratum and design cluster random effects. We consider $\boldsymbol{a}^H$, $\boldsymbol{a}^Q$, and $\boldsymbol{a}^{HQ}$ to be mutually independent, while Z^H, Z^Q, Z^{HQ} are index matrixes which indicate whether one subject belongs to a specific stratum, cluster and cross section of stratums and cluster. Under this model, the posterior mean of the

population mean is

$$\begin{aligned}\hat{\bar{Y}} = E(\bar{Y}|y) &= E(E(\bar{Y}|\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{a}^H, \boldsymbol{a}^Q, \boldsymbol{a}^{HQ})) \\ &= \sum_{h=1}^{H}\sum_{q=1}^{Q} P_{hq}(X_h^H \hat{\boldsymbol{\mu}} + \hat{a}_h^H + \hat{a}_q^Q + \hat{a}_{hq}^{HQ})\end{aligned}$$

where $P_{hq} = N_{hq}/N$, $\hat{\boldsymbol{\mu}} = E(\boldsymbol{\mu} \mid \boldsymbol{y})$, $\hat{a}_h^H = E(a_h^H \mid \boldsymbol{y})$,$\hat{a}_h^Q = E(a_q^Q \mid \boldsymbol{y})$, and $\hat{a}_h^{HQ} = E(a_{hq}^{HQ} \mid \boldsymbol{y})$

We consider the following special cases of the model by varying the form of the weight stratum fixed effects X^H:

Exchangeable Stratum Effect(XSE):

$$X^H = I$$

Linear Stratum Effect(LSE):

$$X^H = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & H \end{pmatrix}$$

Spline Linear Stratum Effect(SLSE):

$$X^H = (\underline{1}, \underline{h}, I_1 \times \underline{h}, \ldots, I_k \times \underline{h}) = \begin{pmatrix} 1 & 1 & i_{11} \times 1 & \ldots & i_{k1} \times 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & H & i_{1H} \times H & \ldots & i_{kH} \times H \end{pmatrix}$$

The XSE model assumes that the means for the weight strata deviate around a common overall mean (no trend relating the probability of selection of the mean of the outcome). The LSE model assumes that there exists a linear trend of weights for the observations, so we added weight strata as a fixed effect μ^H. The SLSE model assumes the observed values are associated with weights in a linear spline pattern

with $k-2$ knots at weight strata. Thus, the coefficients of linear effects varies between different weight strata. In the design matrix of fixed effects X^{SH}, $I_1, \ldots, I_k$ are the indicator vectors corresponding to k different weights intervals which indicate whether observations belong to corresponding weight intervals.

All of these models can be written in the mixed-effect form

$$\boldsymbol{y} = X\boldsymbol{\mu} + Z\boldsymbol{a} + \boldsymbol{\varepsilon}$$

where $X = N^H X^H$, $Z = (N^H Z^H \ N^Q Z^Q \ N^{HQ} Z^{HQ})$, $\boldsymbol{a} = (\boldsymbol{a}^{H^T}, \boldsymbol{a}^{Q^T}, \boldsymbol{a}^{HQ^T})^T \sim N(0, G))$ for

$$G = \begin{pmatrix} \sigma_h^2 I_H & 0 & 0 \\ 0 & \sigma_q^2 I_Q & 0 \\ 0 & 0 & \sigma_{hq}^2 I_{HQ} \end{pmatrix}, \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n),$$

and N^H, N^Q, and N^{HQ} are $n \times H$, $n \times Q$, and $n \times HQ$ "incidence" matrices relating the distinct weight strata, clusters, and strata-by-clusters to the data ($n_{jk}^H = 1$ if y_j is in weight stratum k and 0 otherwise). For the model parameter estimation, we can use either a fully Bayesian approach to infer about the posterior predictive distribution, or use empirical Bayesian (EB) methods obtained via ML or REML estimation from standard linear mixed model methods (Laird and Ware 1982, Carlin and Louis 1996). We pursue the empirical Bayesian approach in this manuscript. The estimates of G, σ^2, μ, a can be obtained by maximum likelihood (ML) or restricted maximum likelihood (REML) methods. The corresponding log-likelihood functions are as follows:

$$\text{ML:} \quad l(G, \sigma^2) = -\frac{1}{2} log|V| - \frac{1}{2} \boldsymbol{r}^T V^{-1} \boldsymbol{r} - \frac{n}{2} log(2\pi)$$

$$\text{REML:} \quad l_R(G, \sigma^2) = -\frac{1}{2} log|V| - \frac{1}{2} log|X^T V^{-1} X| - \frac{1}{2} \boldsymbol{r}^T V^{-1} \boldsymbol{r} - \frac{n-p}{2} log(2\pi)$$

where $V = ZGZ^T + \sigma^2 I$ and $\boldsymbol{r} = \boldsymbol{y} - X(X^T V^{-1} X)^{-1} X^T V^{-1} \boldsymbol{y}$.

When the variance components G and σ^2 are known, both the ML and REML estimator of the overall mean is $\hat{\boldsymbol{\mu}} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}\boldsymbol{y}$, and the estimator of the random effects is $\hat{\boldsymbol{a}} = \hat{G}Z'\hat{V}^{-1}(\boldsymbol{y} - X\hat{\boldsymbol{\mu}})$. Estimates of G and σ^2 can also be obtained by EM algorithm or Newton-Raphson algorithm (Lindstrom and Bates 1988).

With regard to the variance estimation of $\hat{\bar{\mathbf{Y}}} = \mathbf{E}(\bar{\mathbf{Y}}|\boldsymbol{y})$, the empirical Bayesian approach yields:

$$\begin{aligned}\mathrm{Var}(\hat{\bar{Y}}) &= (\boldsymbol{N_{hq}} - \boldsymbol{n_{hq}})^T \mathrm{Var}(\hat{\boldsymbol{\mu}} - \bar{\boldsymbol{Y}}_{\boldsymbol{nob}})(\boldsymbol{N_{hq}} - \boldsymbol{n_{hq}}) \\ &= (\boldsymbol{N_{hq}} - \boldsymbol{n_{hq}})^T(\sigma^2\Lambda + ZGZ^T + \Gamma V \Gamma^T - 2\Gamma ZGZ^T)(\boldsymbol{N_{hq}} - \boldsymbol{n_{hq}})/N_+^2\end{aligned}$$

(Holt and Smith 1979, Lazzeroni and Little 1998, Elliott and Little 2000), where $\Lambda = diag(N_{hq} - n_{hq})^{-1}$, $\Gamma = (I - ZGZ^TV^{-1})X(X^TV^{-1}X)^{-1}X^TV^{-1} + ZGZ^TV^{-1}$, $\mathbf{N_{hq}} - \mathbf{n_{hq}}$ is a $(H + Q + H \times Q) \times 1$ vector of counts of unobserved population of each cluster-strata cross section.

Another approach to variance estimation is the jackknife replication method (Korn and Graubard 1999). Let $\hat{\theta}$ represent the estimate from the model based on all the data, and $\hat{\theta}_{(i)}$ be an estimate from the model excluded the data from ith cluster. The sample weights of the remaining data are multiplied by a factor of $Q/(Q-1)$ when calculating $\hat{\theta}_{(i)}$. The jackknife variance estimator is given by $\widehat{\mathrm{VAR}}_{JK}(\hat{\theta}) = Q/(Q-1)\sum_{i=1}^{Q}(\hat{\theta}_{(i)} - \hat{\theta})^2$. The jackknife estimator is more robust than the model based estimator since it does not assume normally distributed data. However, simulation studies have shown that jackknife estimators are often conservative, yielding confidence intervals that are too wide (Kish and Frankel 1974). In this work, for robust property and convenience of computing, we estimate the variance using the jackknife estimator.

2.3 Simulation Study

2.3.1 Design of the simulation study

To study the performance of the proposed models, we develop a simulation that approximates the design common in many household area probability samples. We first generated $Q = 1000$ clusters, with the size of each drawn from a geometric distribution $S_q \sim \text{GEO}(0.002)$. We then generated a random cluster effect $a_q \sim N(0, \xi^2)$. Each cluster can be thought of as primary sampling unit (commonly a county or other geographic unit), and each element in the cluster can be thought of as corresponding to a household. A household size variable $s_{qi}, i = 1, \ldots, S_q$ was generated from Poisson distribution:

$$\underset{\sim}{s} = \text{Poission}(\lambda_q), \lambda_q = 8\frac{\exp(k_1 a_q/\text{range}(a_q))}{1 + \exp(k_1 a_q/\text{range}(a_q))}$$

We generated a binary covariate associated with each household: $x_{qi} \sim \text{Bin}(0.2)$.

The outcome for each member in the household was then generated from

$$Y_{qi} \sim N(10 + k_2 s_{qi} + a_q, \sigma^2) \tag{2.2}$$

$$a_q \sim N(0, \xi^2), \xi^2 = c\sigma^2$$

A sample was drawn from the generated population using a common three stage design. The first stage sampled 20 clusters with probability proportional to cluster size (PPS). The second stage sampled $s_q = 50$ households within each sampled cluster with probability $p_{qi} = \frac{3x_{qi}+1}{\sum_i (3x_{qi}+1)}$. If there were less than 50 households in a sampled cluster, all of the households in this cluster are sampled. The third stage samples one element (*i.e.* one adult) from each sampled household.

Weights are calculated as the inverse of the inclusion probability.

$$\begin{aligned} P(I_{qi}=1) &= \frac{\text{Cluster size}}{\text{Population size}} \times \frac{\text{Sampled households within cluster}}{\text{Cluster size}} \\ &\quad \times \frac{1}{\text{Household size}} \\ &= \frac{s_q \times p_{qi}}{S} \times \frac{1}{s_{qi}} \end{aligned}$$

where I_{qi} is the inclusion indicator and $S = \sum_q S_q$. The weight w_{qi} is then given by

$$w_{qi} = \frac{1}{P(I_{qi})=1} = S \times \frac{s_{qi}}{s_q \times p_{qi}}$$

The constant c tunes the size of the cluster effect. The cluster effect increases as c increases. The constants k_1 and k_2 tune the correlation between outcomes and weights. As k_1 increases, the correlation between weight strata and cluster size increases inducing correlation between outcomes and weights. As k_2 increases, the correlations between outcomes and weights increase directly. As σ^2 increases, the signal-to-noise ratio for the association between the probability of inclusion and the outcome decreases. We conduct simulations under different cluster effects, when $c = 0.001, c = 0.01, c = 0.1$. Under each cluster effect, we study the situations when $k_1 = k_2 = 0$, $k_1 = 1, k_2 = 2$. Under each situation, we study population variance $\sigma^2 = 0.1, 10, 1000$.

In order to study the performance when the data is generated with a non-linear relationship between the probability of selection and the outcome, we simulated a set of quadratic population data and applied the same models. The model for the data will be given by

$$Y_{qi} \sim N(10 + k_2 s_{qi} + k_3 s_{qi}^2 + a_q, \sigma^2) \tag{2.3}$$

$$a_q \sim N(0, \xi^2), \xi^2 = c\sigma^2$$

We set $k_3 = 0.5$ to represent the nonlinearity and studied the situations when $k_1 = 0, k_2 = 0$ and $k_1 = 1, k_2 = 2$. Similarly with simulations without model misspecification, under each situation we study population variance $\sigma^2 = 0.1, 10, 1000$.

Two hundred samples are drawn from each population data set. The properties we are interested in are root mean squared error (RMSE), bias, and coverage of nominal 95% confidence intervals. RMSE is estimated as $\sqrt{\sum_{i=1}^{200}(\hat{\mu}_i - \mu)^2/200}$, where μ is the population mean and $\hat{\mu}_i$ is the estimate from the ith of the 200 samples.

To better describe the characteristics of weights and weight strata, we take one sample data set of the population data with cluster effect parameter $c = 0.001$, variance equal to 10 and correlation parameters $k_1 = 0, k_2 = 0$ for example. The sample size is 1000 which represents a population of 2,361,115. We set separate weight stratum according to every 0.05 quintile of weights, thus 20 strata are created. The characteristics are depicted in Figure 2.1.

We can see from the figure more than 80% sample weights are less than 5,000, while less than 20% weights ranged from 5,000 to 16,000.

2.3.2 Simulation Results

Simulation by linear model

Table 2.1, Table 2.2 and Table 2.3 summarize the relative bias, RMSE, and coverage for the three model-based estimators described in Section 2.2, using the 18 simulated populations described in the tables. The data are generated under linear model (2). For the weight smoothing models, we use weights to estimate the proportion of each cross section to the whole population, *i.e.*, $\hat{P_{hq}} = \hat{N_{hq}}/\hat{N} = \sum_i w_{hqi} / \sum_{h,q,i} w_{hqi}$. For the Spline Linear Stratum Effect model (SLSE), we use 40 and 70 percentiles of the weights as our cut points. The variances are estimated by the jackknife method. For comparison, we also consider the unweighted estimator of

the mean, and two design-based estimators: a fully weighted estimator, an a trimmed weight estimator. Weights are trimmed at three times the average weights for the "crude" weight trimming method. We use the Taylor series expansion theory to estimate the variance of weighted estimator $\hat{\bar{Y}} = (\sum_{q=1}^{Q}\sum_{i=1}^{n_q} w_{qi}y_{qi})/(\sum_{q=1}^{Q}\sum_{i=1}^{n_q} w_{qi})$, where n_q denotes number of subjects sampled in cluster q. Specifically, the procedure computes the estimated variance as $\widehat{\mathrm{Var}}(\hat{\bar{Y}}) = \frac{n(1-f)}{n-1}\sum_{q=1}^{Q}(e_{q.} - \bar{e}_{..})^2$, where $e_{q.} = (\sum_{i=1}^{n_q} w_{qi}(y_{qi} - \hat{\bar{Y}}))/(\sum_{q=1}^{Q}\sum_{i=1}^{n_q} w_{qi})$ and $\bar{e}_{..} = \sum_{q=1}^{Q} e_{q.}/n_q$. $f = n/N$ indicates the sampling fraction.

When $k_1 = k_2 = 0$, the outcomes are not associated with weights and weight strata. The weights are unnecessary, which is the favorable situation for unweighted estimators. All of the estimators have good bias properties. Unweighted estimators have best RMSE properties. In general, model based estimators have better performance when cluster effect is small. XSE estimators tend to simulate unweighted estimators. When cluster effect is small or moderate ($c = 0.001$ or $c = 0.01$), XSE estimators have around 30% reduction in RMSEs relative to fully-weighted estimators. When cluster effect goes larger ($c = 0.1$), the reduction of RMSEs of XSE estimator relative to fully-weighted estimators is around 7-8%. LSE and SLSE estimators are more close to fully-weighted estimators, with around (0-4%) reduction in RMSEs when cluster effect is small or moderate ($c = 0.001$ or $c = 0.01$) and around (0-2%) reduction when cluster effect is large ($c = 0.1$). The coverage properties of the estimators are generally good.

When $k_1 = 1$ and $k_2 = 2$,the weights help to correct for the underrepresentation of smaller observation in the sample, and the relationship between the mean and the probability of selection is enhanced by inducing a relationship between the size of the unobserved cluster component a_q and the probability of selection. Thus the

unweighted estimator suffers from bias and, when the variance is small, substantial increases in RMSE over the weighted estimator. The trimmed weight estimators have improved performance than unweighted estimators, but generally larger RMSEs than model based estimators. XSE estimators here are not performing as well as fully-weighted estimators. LSE and SLSE estimators are robust enough to capture the relationship between the probability of selection and the mean, and yield RMSE savings over the fully weighted estimators for 0-4%. Results does not differ much under different cluster effects.

Simulation by Quadratic Model

Table 2.4, Table 2.5 and Table 2.6 summarized the relative bias, RMSE, and coverage for the unweighted, two design-based, and three model based mean estimators, with the 18 simulated population data sets described in tables. The data are generated under the quadratic model (3). The linear model here is misspecified. Under each cluster effect ($c = 0.001, c = 0.01, c = 0.1$),we consider the simulated populations from $k_1 = k_2 = 0$, and $k_1 = 1, k_2 = 2$.

Under all simulation settings, the unweighted estimators perform poorly with respect to biases, RMSE and coverage, reflecting the substantial relationship between the probability of selection and the mean. The fully-weighted estimator largely eliminated this bias and restored approximately correct coverage.

When $k_1 = 0$ and $k_2 = 0$, XSE estimators have larger RMSEs than fully-weighted estimators. LSE and SLSE estimators had similar results or some slightly reduction in RMSE relative to the fully-weighted estimator (up to 2%) with modest and large variances. Coverage was approximately correct for LSE and SLSE estimators for smaller variances. The trimmed weight estimators have more bias and larger RMSE relative to the model-based estimators. There is not much differences under different

cluster effects.

Results were similar for the simulations with $k_1 = 1$ and $k_2 = 2$. RMSE savings for model-based estimators relative to fully-weighted estimators decreased a little when cluster effect is small and variance is large.

2.4 Application: Assessing Associations between LDL Cholesterol Level and Birth Weight in a Population-Based Sample

Previous studies have assessed the relationship between low birth weight and increased risk for cardiovascular disease and non-insulin-dependent diabetes later in life. The proposed biological mechanism is that fetal malnourishment leads to changes in fetal and placental hormones which will in turn lead to later life increased expression of cardiovascular disease risk factors, such as increasing insulin resistance, altering liver function and lowering blood vessel elasticity. However, the findings in previous studies showed contradictory results (Forrester et al. 1996, Huxley and Law 2000, Huxley and A. Collins 2002, Owen et al. 2003, Matthes et al. 1994). Moreover, few of these studies used well-defined national population. Few studies used data from children and few of them treated blood lipid level directly as an outcome.

In this application, we assess the association between blood lipid levels and birth weight (BW) in a multi-ethnic, population-based sample of 4-12 year old children. Data were obtained from National Health and Nutrition Examination Survey III, a cross-sectional probability sample of the United States population. Analysis was restricted to the 4,151 subjects aged 4-12 for whom birth weight and lipid profiles were available. Some of the measurements are invalid, we consider them as missing, so the sample is restricted to 867 children. The data set has 98 pseudo PSUs, which can be considered as clusters. Because the weights were nearly unique to each subject, we stratified the data based on percentiles of the case weights. We compare three

stratifications: 85 weight strata, 50 weight strata and 20 weight strata.

We assess the relationship between LDL cholesterol (LDL-C) level and birth weight using a quadratic effect, to allow for the possibility that both low and high birth weight may be associated with LDL-C level. Unweighted (UWT), fully weighted (FWT), weight trimmed to three times the average weight (TWT), exchangeable strata effect model (XSE), linear strata effect model (LSE), spline linear strata effect model(SLSE) estimators of the slopes relating linear and quadratic terms of birth weight to LDL-C are obtained. One method to account for unequal probability of selection in linear regression is to include an interaction term between the weight strata indicators and the covariates (just as the inclusion of the weight strata themselves in the mean model above can be thought of as in interaction between the intercept term and the weight strata). The model then estimates a separate slope for each probability of selection stratum, and averages these slopes in proportion to their (known or estimated from case weights) fraction in the population. Little sharing of information across the weight strata yields a fully-weighted estimator of the population slope; larger degrees of sharing yield estimators that approximate weight trimming Elliott (2007).

Similar to section 2.2, we can write these models in the mixed-effect model form; notice that we are interested in estimating the coefficients for both intercept and covariates of linear and quadratic 'birth weight':

$$\boldsymbol{LDL} = X\boldsymbol{\beta} + Z\boldsymbol{a} + \boldsymbol{\varepsilon}$$

where

$$X = X^H, Z = (Z^H \ Z^{H*bw} \ Z^{H*bw^2} \ Z^Q \ Z^{Q*bw} \ Z^{Q*bw^2}),$$

$$\boldsymbol{a} = (\boldsymbol{a}^{H^T}, \boldsymbol{a}^{H*bw^T}, \boldsymbol{a}^{H*bw^{2T}}, \boldsymbol{a}^{Q^T}, \boldsymbol{a}^{Q*bw^T}, \boldsymbol{a}^{Q*bw^{2T}})^T \sim N(0, G))$$

$$G = \begin{pmatrix} \sigma_h^2 I_H & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_{h*Bw}^2 I_H & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{h*Bw^2}^2 I_H & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_q^2 I_Q & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{q*Bw}^2 I_Q & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{q*Bw^2}^2 I_Q \end{pmatrix}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$$

LDL indicates the LDL cholesterol level and bw denotes the birth weight. We assume the all random effects are independent of each other, i.e. exchangeable. Here, Z^H, Z^Q are index matrices which indicate whether one subject belongs to specific stratum, cluster. Z^{H*bw}, Z^{H*bw^2} and Z^{Q*bw}, Z^{Q*bw^2} are matrices of corresponding covariates of linear and quadratic birth weights for each subject. Here we assume the correlations between random effect of intercept and random effect of slope to be zero, while Elliott (2007) assume they are correlated in the exchangeable model, but not in the linear model.

We consider three models with different structure X^H:

Exchangeable Stratum Effect(XSE):

$$X^H = (\underline{1}, \underline{bw}, \underline{bw}^2) = \begin{pmatrix} 1 & bw & bw^2 \\ \vdots & \vdots & \vdots \\ 1 & bw & bw^2 \end{pmatrix}, \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$$

The fixed effects for intercept and birth weight are:

$$\beta_{I,h}^* = \beta_0, \beta_{bw,h}^* = \beta_1, \beta_{bw^2,h}^* = \beta_2$$

Linear Stratum Effect(LSE):

$$X^H = (\underline{1}, \underline{bw}, \underline{bw}^2, \underline{h}, \underline{h} \times \underline{bw}, \underline{h} \times \underline{bw}^2) = \begin{pmatrix} 1 & bw & bw^2 & 1 & 1 \times bw & 1 \times bw^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & bw & bw^2 & H & H \times bw & H \times bw^2 \end{pmatrix}$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$$

The fixed effects for intercept and birth weight are:

$$\beta^*_{I,h} = \beta_0 + \beta_3 h, \beta^*_{bw,h} = \beta_1 + \beta_4 h, \beta^*_{bw^2,h} = \beta_2 + \beta_5 h$$

Spline Linear Stratum Effect(SLSE):

In this application, SLSE used 40% and 70% weight strata as splines, which is indicated by vectors of $I1$ and $I2$ correspondingly.

$$\begin{aligned} X^H = (&\underline{1}, \underline{bw}, \underline{bw}^2, \underline{h}, I_1 \times \underline{h}, I_2 \times \underline{h}, \underline{bw} \times \underline{h}, \underline{bw} \times I_1 \times \underline{h}, \underline{bw} \times I_2 \times \underline{h}, \\ &\underline{bw}^2 \times \underline{h}, \underline{bw}^2 \times I_1 \times \underline{h}, \underline{bw}^2 \times I_2 \times \underline{h}) \\ = &\begin{pmatrix} 1 & bw & 1 & i_{11} \times 1 & i_{21} \times 1 & 1 \times bw & i_{11} \times 1 \times bw, & i_{21} \times 1 \times bw \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & bw & H & i_{1H} \times H & i_{2H} \times H & H \times bw & i_{1H} \times H \times bw, & i_{2H} \times H \times bw \end{pmatrix} \\ &\begin{pmatrix} 1 \times bw^2 & i_{11} \times 1 \times bw^2, & i_{21} \times 1 \times bw^2 \\ \vdots & \vdots & \vdots \\ H \times bw^2 & i_{1H} \times H \times bw^2, & i_{2H} \times H \times bw^2 \end{pmatrix} \end{aligned}$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8)$$

The fixed effects for intercept and birth weight are:

$$\beta^*_{I,h} = \beta_0 + \beta_2 h + \beta_3 i_{h1} h + \beta_4 i_{h2} h$$

$$\beta^*_{bw,h} = \beta_1 + \beta_2 h + \beta_5 i_{h1} h + \beta_6 i_{h2} h$$

$$\beta^*_{bw^2,h} = \beta_1 + \beta_2 h + \beta_7 i_{h1} h + \beta_8 i_{h2} h$$

We will get estimators for each section $\hat{\boldsymbol{\beta}_{\boldsymbol{hq}}} = (\hat{\beta^0_{hq}}, \hat{\beta^1_{hq}})^T$ after adding corresponding random effects:

$$\hat{\beta^0_{hq}} = \hat{\beta^*_{I,h}} + \hat{a}^H_h + \hat{a}^Q_q$$

$$\hat{\beta^1_{hq}} = \hat{\beta^*_{bw,h}} + \hat{a}^{H*bw}_h + \hat{a}^{Q*bw}_q$$

$$\hat{\beta^2_{hq}} = \hat{\beta^*_{bw^2,h}} + \hat{a}^{H*bw^2}_h + \hat{a}^{Q*bw^2}_q$$

The estimators for the population are then given by

$$\hat{B} = E(B|y,X) = [\sum_h \sum_q W_{hq} \sum_{i=1}^{n_{hq}} x_{hqi} x^T_{hqi}]^{-1} [\sum_h \sum_q W_{hq} (\sum_{i=1}^{n_{hq}} x_{hqi} x^T_{hqi}) \hat{\boldsymbol{\beta}_{\boldsymbol{hq}}}]$$

Here, $\hat{\beta}_{hq}$ is the target estimator for each cross-section which combined the fixed effect and random effects.

The results are summarized in Table 2.7. By assuming the fully weighted estimator as unbiased, the MSEs for model based methods are calculated using method described in Little et al. (1997), which is a refined MSE estimator. $\widehat{\text{MSE}} = \hat{V} + \max\{\hat{B}^2 - \hat{V}_{01}, 0\}$, where $\hat{V}$ is the jacknife variance estimate for model based parameters, $\hat{B}$ estimates bias between the model based estimator and the fully weighted estimator, $\hat{V}_{01}$ is the jacknife variance for the bias estimator $\hat{B}$ which corrects for upward bias of $\hat{B}^2$ as an estimate of the squared bias.

Generally, UWT, TWT and all model based estimators reduced estimates' standard errors compared to FWT estimators. Because we assume the FWT estimator

is the unbiased estimator, estimators more biased from FWT estimators will have larger RMSEs. Since all estimators for quadratic birth weight term are similar, all other estimators have much smaller standard error, thus much smaller RMSEs than FWT estimators. For intercept and linear birth weight term, FWT and UWT estimators are different, thus RMSEs of UWT estimators for these two terms are larger compared with FWT estimator. As we look into the model based estimators, the XSE estimates of these two terms are closer to UWT estimators. For intercept term, the bias is not large enough to inflate RMSEs, thus XSE estimators for intercept have smaller RMSEs compared with FWT estimators. For linear birth weight term, estimated biases and standard errors are large, which inflated RMSEs of XSE estimators. LSE and SLSE estimators for these two terms in the models with 20 and 50 weight strata are very close to FWT estimators, thus have reduced RMSEs compared to FWT estimators. LSE and SLSE estimators of 85 weight strata model for these two terms are between FWT and UWT estimators, but they have large reduction in estimated standard error, thus they also have large reduction in RMSEs. As we noticed, LSE and SLSE estimators in 85 weight strata model have larger bias toward FWT estimators than these estimators in 20 and 50 weight strata. This difference might be caused by the fact that there are too many parameters to estimate in 85 weight strata model. In this data set, the correlation between outcome and weights are around 0.04. Consistent with our simulation results, LSE and SLSE model based methods have superior performance with respect to variations when the outcome and weights are correlated. FWT and TWT estimators suggest that low birth weight has slight linear association with high LDL-C level, while other estimators do not show significant linear association. For quadratic birth weight effect, the upper bound of confidence intervals of all estimators are very close to zero, suggesting a possible

weak association. To further assess whether there is an association between birth weight and LDL-C level, we further conducted wald test for linear and quadratic birth weight coefficients jointly. Results suggest that generally there is a significant association of low birth weight and high LDL-C level. Thus in our study, we conclude that there is a significant relationship between low birth weight and increased risk for cardiovascular disease and non-insulin-dependent diabetes in one's later childhood, although this relationship is weak.

2.5 Discussion

The model developed in this paper extends the previous works of Elliott and Little (2000), Lazzeroni and Little (1998), Elliott (2007) and others that build weight smoothing models only for weight strata. This manuscript extends the weight smoothing models to survey data with cluster designs.

As we can see from simulation, the model-based methods proposed here outperform design-based methods overall. When cluster effect is small and outcomes are not associated with weights, XSE estimators largely reduced RMSEs relative to FWT estimators. When the correlation is larger or model are misspecified, the RMSEs of XSE estimators go up. LSE and SLSE estimators have slightly reductions in RMSEs relative to FWT estimators while generally maintaining good coverage. Performances of LSE ad SLSE estimators are similar under different cluster effects and correlations.

As we can see from the simulation and application results, the performance of the weight smoothing models relative to the design-based estimators will depend on the population structure. Smoothing by weight strata decreases MSE, but clustering inflates MSE. If the variability of the weights is not large compared to the variability

of clustering, the model-based estimators do not substantially outperform the design-based estimators. When weights are unique or nearly unique and thus need to be pooled to provide stable estimates, it is important to ensure that we have a sufficient number of weight strata.

In this paper, we have pursued the empirical Bayesian approach because of the memory limit of software R and because of time-saving considerations. Under this approach, the variance estimation of $\mathrm{Var}(\bar{Y}|y)$ will be biased downward since it ignores uncertainty in the estimates of G and σ^2. The advantage of fully Bayesian method is that the hyperprior of the random effect parameters will account for the uncertainty in the prior parameters. This effect will be most important in smaller samples than have been considered in this manuscript.

While applying the method to linear regression models, as in our application part to assess associations between LDL cholesterol level and birth weight, it is necessary to smooth over weight strata to estimate both intercept and slope coefficients. It is also necessary to adjust for cluster effects for both intercept and slope coefficients. Adjusting the cluster or weight strata effect only for the intercept coefficient may not account for model misspecification or non-ignorable sampling and thus lead to biased estimates of the slope coefficient.

$k_1 = k_2 = 0$	$Var = 0.1$			$Var = 10$			$Var = 1000$		
Estimator	%RB	RMSE	Cvr	%RB	RMSE	Cvr	%RB	RMSE	Cvr
UNWT	0.00	0.65	94.5	0.04	0.65	95.5	-0.63	0.68	95
FWT	0.01	1	92	0.08	1	94.5	-0.87	1	95.5
TWT	0.01	0.92	92	0.09	0.89	95	-0.86	0.89	95
XRE	0.00	0.68	93.5	0.03	0.68	95.5	-0.67	0.71	96
LSE	0.10	0.98	92	0.09	0.96	94	-0.80	0.97	96
SLSE	0.10	1.00	93	0.11	0.98	95	-0.62	0.96	95.5
$k_1 = 1, k_2 = 2$	$Var = 0.1$			$Var = 10$			$Var = 1000$		
UNWT	-7.58	5.96	0	7.71	4.90	0	-8.63	1.41	60
FWT	0.09	1	94	0.24	1	93	-1.29	1	95.5
TWT	-1.20	1.31	82.5	-1.21	1.14	93	-2.61	0.97	94
XRE	-0.39	1.04	93	-0.74	1.07	88.5	-6.94	1.24	79.5
LSE	-0.14	0.99	94.5	-0.04	0.98	92	-1.72	0.99	94.5
SLSE	-0.13	1.00	94	-0.09	0.98	91	-1.92	1.00	95

Table 2.1: Simulation results: Population generated under linear model (2.2), small cluster effect ($c = 0.001$). Results are based on 200 simulations. Relative bias to the true population mean(%RB), square root of mean square error (RMSE) relative to RMSE of fully weighted estimator, and true coverage of the nominal 95% confidence or credible interval of population mean estimator (CVR). Population means are estimated via design based unweighted (UNWT), fully weighted (FWT), and weight trimmed estimators (TWT), and under the exchangeable (XSE), linear (LSE), and linear spline (SLSE) models.

$k_1 = k_2 = 0$	$Var = 0.1$			$Var = 10$			$Var = 1000$		
Estimator	%RB	RMSE	Cvr	%RB	RMSE	Cvr	%RB	RMSE	Cvr
UNWT	0.02	0.72	95	0.02	0.70	94.5	-0.12	0.72	95
FWT	0.00	1	92.5	0.07	1	96	0.20	1	95
TWT	0.00	0.91	94.5	0.08	0.94	94.5	0.61	0.91	96
XSE	0.01	0.74	94	0.00	0.77	95.5	-0.28	0.76	94.5
LSE	0.00	0.98	93	0.08	0.99	95.5	0.38	0.96	95.5
SLSE	0.00	0.98	94.5	0.08	0.98	95.5	0.18	0.99	95
$k_1 = 1, k_2 = 2$	$Var = 0.1$			$Var = 10$			$Var = 1000$		
UNWT	-7.46	5.67	0	-7.70	4.93	0	-8.20	1.13	76.5
FWT	0.03	1	93	-0.40	1	94	0.88	1	93
TWT	-1.35	1.35	84	-1.79	1.44	84.5	-1.23	0.90	94.5
XSE	-0.50	1.05	91	-1.41	1.31	86.5	-5.69	0.99	84.5
LSE	-0.20	1.00	92.5	-0.65	0.99	93	0.37	0.96	94.5
SLSE	-0.18	1.01	94	-0.72	0.98	91	0.19	0.97	95

Table 2.2: Simulation results: Population generated under linear model (2.2), moderate cluster effect ($c = 0.01$). Results are based on 200 simulations. Relative bias to the true population mean(%RB), square root of mean square error (RMSE) relative to RMSE of fully weighted estimator, and true coverage of the nominal 95% confidence or credible interval of population mean estimator (CVR). Population means are estimated via design based unweighted (UNWT), fully weighted (FWT), and weight trimmed estimators (TWT), and under the exchangeable (XSE), linear (LSE), and linear spline (SLSE) models.

$k_1 = k_2 = 0$	$Var = 0.1$			$Var = 10$			$Var = 1000$		
Estimator	%RB	RMSE	Cvr	%RB	RMSE	Cvr	%RB	RMSE	Cvr
UNWT	-0.01	0.76	95	0.05	0.82	94.5	-1.40	0.79	94
FWT	-0.01	1	93	1.18	1	93	-3.75	1	92.5
TWT	-0.02	0.95	93	0.97	0.96	94	-2.26	0.93	92
XSE	-0.01	0.92	94.5	0.93	0.93	94	-3.09	0.93	93
LSE	-0.01	1.00	93.5	1.16	0.98	93.5	-3.76	0.98	92
SLSE	-0.01	1.00	93.5	1.17	0.99	93.5	-3.74	0.99	92.5
$k_1 = 1, k_2 = 2$	$Var = 0.1$			$Var = 10$			$Var = 1000$		
UNWT	-7.57	5.61	0	-7.86	3.66	2	-10.84	1.13	87
FWT	-0.04	1	92	-0.44	1	92.5	-0.60	1	94.5
TWT	-1.41	1.38	77	-1.75	1.23	91.5	-2.61	0.99	93
XSE	-0.57	1.07	88.5	-1.36	1.16	93	-6.69	1.06	88.5
LSE	-0.27	1.01	91.5	-0.70	0.99	92	-1.08	0.98	93.5
SLSE	-0.24	1.00	92	-0.76	1.01	92	-1.24	0.99	93

Table 2.3: Simulation results: Population generated under linear model (2.2), large cluster effect ($c = 0.1$). Results are based on 200 simulations. Relative bias to the true population mean(%RB), square root of mean square error (RMSE) relative to RMSE of fully weighted estimator, and true coverage of the nominal 95% confidence or credible interval of population mean estimator (CVR). Population means are estimated via design based unweighted (UNWT), fully weighted (FWT), and weight trimmed estimators (TWT), and under the exchangeable (XSE), linear (LSE), and linear spline (SLSE) models.

$k_1 = k_2 = 0$	$Var = 0.1$			$Var = 10$			$Var = 1000$		
Estimator	%RB	RMSE	Cvr	%RB	RMSE	Cvr	%RB	RMSE	Cvr
UNWT	-15.22	6.20	0	-15.12	5.68	0	-15.14	2.63	4
FWT	-0.05	1	93	0.19	1	93.5	0.78	1	96.5
TWT	-3.29	1.60	65.5	-2.91	1.38	70.5	-2.60	0.92	96
XSE	-0.99	1.07	90	-0.90	1.04	88.5	-7.35	1.49	75
LSE	-0.54	1.00	91	-0.33	0.97	92.5	-0.13	0.92	97.5
SLSE	-0.41	1.01	91.5	-0.17	0.98	92	-0.35	0.93	98
$k_1 = 1, k_2 = 2$	$Var = 0.1$			$Var = 10$			$Var = 1000$		
UNWT	-15.32	5.65	0.00	-15.12	4.87	0	-8.45	3.08	1.5
FWT	0.04	1	91.5	0.50	1	95	0.24	1	95.5
TWT	-3.06	1.41	76.0	-2.71	1.20	83	-1.55	1.03	90
XSE	-0.96	1.05	89.0	-0.53	0.99	96	-3.03	1.42	74.5
LSE	-0.47	0.99	91.5	0.02	0.97	96	-0.25	0.99	97
SLSE	-0.36	1.01	91.0	0.10	0.99	95	-0.40	0.98	96.5

Table 2.4: Simulation results: Population generated under quadratic model (2.3), small cluster effect ($c = 0.001$). Results are based on 200 simulations. Relative bias to the true population mean(%RB), Square root of mean square error (RMSE) relative to RMSE of fully weighted estimator, and true coverage of the nominal 95% confidence or credible interval of population mean estimator (CVR). Population means are estimated via design based unweighted (UNWT), fully weighted (FWT), and weight trimmed estimators (TWT), and under the exchangeable (XSE), linear (LSE), and linear spline (SLSE) models.

$k_1 = k_2 = 0$	$Var = 0.1$			$Var = 10$			$Var = 1000$		
Estimator	%RB	RMSE	Cvr	%RB	RMSE	Cvr	%RB	RMSE	Cvr
UNWT	-15.34	6.18	0	-15.54	5.95	0	-15.27	2.59	8.5
FWT	-0.31	1	94	-0.40	1	92.5	-0.31	1	94.5
TWT	-3.30	1.56	68	-3.73	1.64	65.5	-3.41	1.06	92.5
XSE	-1.23	1.09	90	-1.52	1.14	87.5	-8.16	1.62	73.5
LSE	-0.79	0.98	92	-0.93	0.98	90	-1.38	0.98	95.5
SLSE	-0.59	1.00	95	-0.75	0.99	90	-1.49	0.99	95.5
$k_1 = 1, k_2 = 2$	$Var = 0.1$			$Var = 10$			$Var = 1000$		
UNWT	-15.07	5.69	0	-15.19	4.84	0	-15.66	2.70	4.5
FWT	0.15	1	95	-0.36	1	91.5	-0.45	1	92
TWT	-2.96	1.41	77.5	-3.35	1.36	77.5	-3.63	1.08	89.5
XSE	-0.80	1.03	93	-1.41	1.09	90.5	-6.07	1.39	80
LSE	-0.34	0.99	95	-0.85	0.99	93	-1.32	0.99	91.5
SLSE	-0.27	1.00	95	-0.77	1.00	92	-1.54	1.00	92

Table 2.5: Simulation results: Population generated under quadratic model (2.3), moderate cluster effect ($c = 0.01$). Results are based on 200 simulations. Relative bias to the true population mean(%RB), Square root of mean square error (RMSE) relative to RMSE of fully weighted estimator, and true coverage of the nominal 95% confidence or credible interval of population mean estimator (CVR). Population means are estimated via design based unweighted (UNWT), fully weighted (FWT), and weight trimmed estimators (TWT), and under the exchangeable (XSE), linear (LSE), and linear spline (SLSE) models.

$k_1 = k_2 = 0$	$Var = 0.1$			$Var = 10$			$Var = 1000$		
Estimator	%RB	RMSE	Cvr	%RB	RMSE	Cvr	%RB	RMSE	Cvr
UNWT	-15.05	6.04	0	-15.15	5.33	0	-13.79	1.43	69
FWT	0.14	1	96.5	0.10	1	90.5	1.00	1	97
TWT	-3.14	1.48	78	-3.04	1.37	74	-2.24	0.95	95
XSE	-0.87	1.03	94	-0.92	1.04	89	-6.77	1.14	91
LSE	-0.33	0.98	95.5	-0.40	0.99	88.5	-0.03	0.98	97.5
SLSE	-0.14	1.00	95	-0.33	1.00	89.5	-0.28	0.99	96.5
$k_1 = 1, k_2 = 2$	$Var = 0.1$			$Var = 10$			$Var = 1000$		
UNWT	-15.14	5.08	0	-15.20	4.38	0	-16.67	1.96	41.5
FWT	0.01	1	95.5	-0.08	1	92	0.08	1	94.5
TWT	-3.08	1.33	77	-3.05	1.24	80	-3.37	0.99	93.5
XSE	-0.99	1.04	92	-1.06	1.04	88	-5.57	1.13	88.5
LSE	-0.47	0.99	94.5	-0.57	0.99	91.5	-0.81	0.98	95
SLSE	-0.29	1.00	96	-0.43	1.00	92	-0.98	0.99	94

Table 2.6: Simulation results: Population generated under quadratic model (2.3), large cluster effect ($c = 0.1$). Results are based on 200 simulations. Relative bias to the true population mean(%RB), Square root of mean square error (RMSE) relative to RMSE of fully weighted estimator, and true coverage of the nominal 95% confidence or credible interval of population mean estimator (CVR). Population means are estimated via design based unweighted (UNWT), fully weighted (FWT), and weight trimmed estimators (TWT), and under the exchangeable (XSE), linear (LSE), and linear spline (SLSE) models.

Method	Intercept(95%CI)	SE(Relative to FWT)	RMSE(Relative to FWT)
FWT	2.85(2.76,2.94)	0.0445(1.00)	0.0458(1.00)
UWT	2.80(2.75,2.86)	0.0486(1.06)	0.0497(1.08)
TWT	2.86(2.78,2.94)	0.0406(0.89)	0.0406(0.89)
20 strata			
XSE	2.81(2.75,2.87)	0.0306(0.67)	0.0323(0.70)
LSE	2.88(2.79,2.96)	0.0427(0.93)	0.0427(0.93)
SLSE	2.86(2.78,2.95)	0.0418(0.91)	0.0418(0.91)
50 strata			
XSE	2.80(2.75,2.85)	0.0268(0.58)	0.0293(0.64)
LSE	2.88(2.79,2.96)	0.0421(0.92)	0.0421(0.92)
SLSE	2.86(2.77,2.94)	0.0418(0.91)	0.0418(0.91)
85 strata			
XSE	2.80(2.75,2.86)	0.0279(0.61)	0.0300(0.65)
LSE	2.84(2.77,2.90)	0.0328(0.72)	0.0328(0.72)
SLSE	2.83(2.77,2.90)	0.0326(0.71)	0.0326(0.71)
	Birth weight(1)(95%CI)	SE(Relative to FWT)	RMSE(Relative to FWT)
FWT	-0.12(-0.21,-0.02)	0.0534(1.00)	0.0534(1.00)
UWT	0.00(-0.07,0.08)	0.0398(0.74)	0.1171(2.19)
TWT	-0.11 (-0.20,-0.02)	0.0452(0.85)	0.0452(0.85)
20 strata			
XSE	-0.01(-0.14,0.11)	0.0615(1.15)	0.1085(2.03)
LSE	-0.10(-0.21, 0.01)	0.0549(1.03)	0.0549(1.03)
SLSE	-0.10(-0.22, 0.01)	0.0600(1.13)	0.0600(1.13)
50 strata			
XSE	-0.00(-0.09,0.09)	0.0459(0.86)	0.1137(2.13)
LSE	-0.09(-0.19,0.01)	0.0501(0.94)	0.0501(0.94)
SLSE	-0.08(-0.19,0.02)	0.0531(0.99)	0.0531(0.99)
85 strata			
XSE	0.00(-0.07, 0.08)	0.0398(0.75)	0.1171(2.19)
LSE	-0.03(-0.11,0.05)	0.0407(0.76)	0.0850(1.59)
SLSE	-0.03(-0.11,0.05)	0.0400(0.75)	0.0970(1.63)
	Birth weight(2)(95%CI)	SE(Relative to FWT)	RMSE(Relative to FWT)
FWT	-0.03(-0.10,0.04)	0.0467(1.00)	0.0467(1.00)
UWT	-0.04(-0.08,-0.00)	0.0192(0.41)	0.0192(0.41)
TWT	-0.05(-0.10.0.00)	0.0296(0.63)	0.0296(0.63)
20 strata			
XSE	-0.04(-0.08,0.00)	0.0195(0.42)	0.0195(0.42)
LSE	-0.03(-0.08,0.02)	0.0273(0.59)	0.0273(0.59)
SLSE	-0.01(-0.08,0.05)	0.0306(0.66)	0.0306(0.66)
50 strata			
XSE	-0.04(-0.08,0.00)	0.0198(0.42)	0.0198(0.42)
LSE	-0.03(-0.08,0.02)	0.0251(0.54)	0.0251(0.54)
SLSE	-0.02(-0.07,0.03)	0.0256(0.55)	0.0256(0.55)
85 strata			
XSE	-0.04(-0.08,0.00)	0.0191(0.41)	0.0191(0.41)
LSE	-0.05(-0.10,0.00)	0.0231(0.50)	0.0231(0.50)
SLSE	-0.05(-0.10,0.00)	0.0296(0.63)	0296(0.63)

Table 2.7: Associations of non-HDL cholesterol level (mg/dL) and birth weight (lb.): Coefficients of linear regression assessed for intercept, birth weight (Birth weight (1)), quadratic birth weight (Birth weight (2)) among US 4-12 year-olds, by unweighted (UWT), fully-weighted (FWT), trimmed weight (TWT), exchangeable strata effect model(XSE), linear strata effect model(LSE), spline linear strata effect model(SLSE). 95% confidence interval in parenthesis. Standard error (SE) is estimated by jacknife method. RMSE is root mean square error, measured in both absolute values and values relative to FWT estimator (in parenthesis). Data from National Health and Nutrition Examination Survey III.

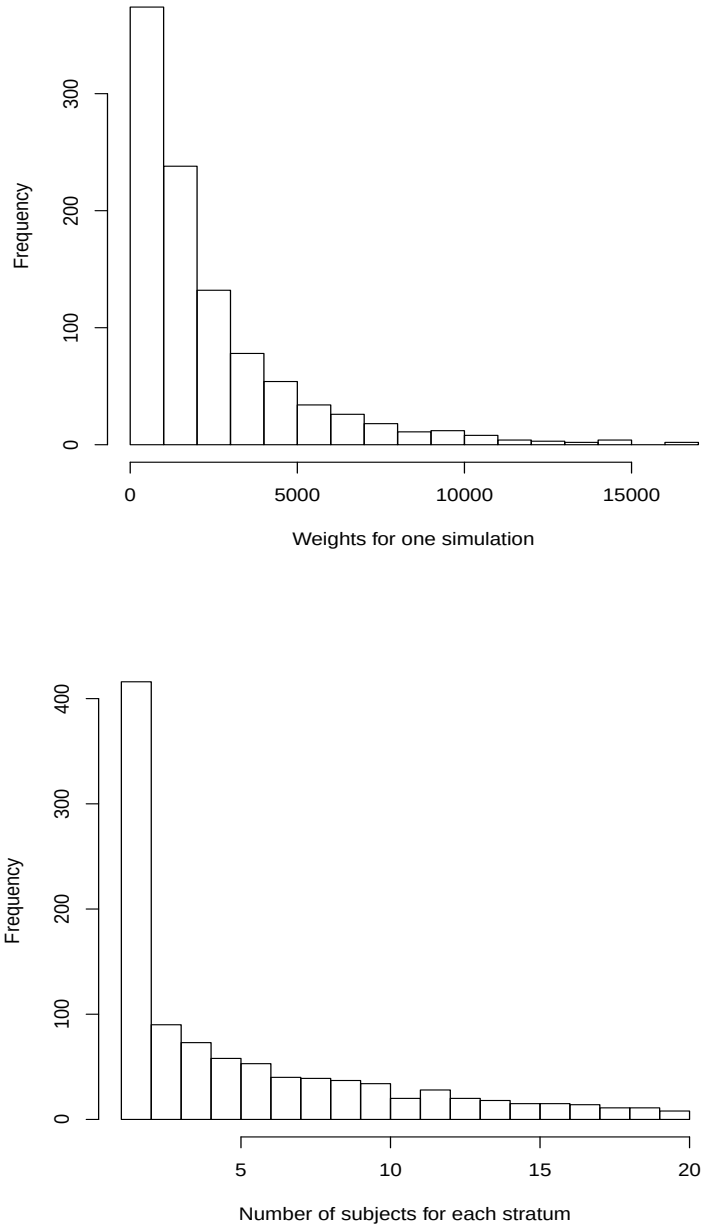

Figure 2.1: Characteristics of survey weights and weight strata of population with $Var = 10, k_1 = 0, k_2 = 0$: The upper panel is the histogram of weights; the bottom panel shows the number of subjects within each stratum.

CHAPTER III

Modeling Menstrual Cycle Length and Variability at the Approach of Menopause Using Bayesian Changepoint Models

3.1 Introduction

Menstrual cycles are the most easily observed markers of ovarian function throughout female reproductive life. Changes in menstrual bleeding patterns are important indicators of ovarian aging, endocrine disruption and endocrine risk factors for chronic disease (Harlow 1995). The menopausal transition is increasingly recognized to be a critical period in women's lives as physiologic changes and health practices adopted during this period frequently define women's long term chronic disease risk profile (Wildman et al. 2008, Sowers et al. 2006, Avis et al. 2004). Given this recent interest in the interface between reproductive and somatic aging, several proposals for staging reproductive aging have emerged. The Stages of Reproductive Aging Workshop (STRAW) recommendations (Soules et al. 2001), its modifications (Harlow et al. 2007) and several other proposals (Mitchell et al. 2000, Taffe and Dennerstein 2002, Mansfield et al. 2004) define criteria primarily by menstrual bleeding characteristics to determine onset of the transition, as well as the stages within the transition period.

Information on the patterns of menstrual bleeding across the reproductive lifespan

derives mainly from four seminal menstrual calendar studies, including three studies from Caucasian populations (Chiazze et al. 1968, Treloar et al. 1967, Vollman 1977) and one study from a Japanese population (Matsumoto et al. 1962, 1979). Treloar (1981) was the first to estimate age at entry into the menopausal transition by visual inspection of menstrual cycle lengths for the 12 year period prior to the final menstrual period (FMP). He observed that during the menopausal transition longer intervals become mixed with shorter than usual intervals, increasing the variability in cycle length. He defined onset of the menopausal transition as the age at which variability in cycle length visually increased, and estimated median age of entry into the transition at 45.5 years with a median duration of transition of 4.8 years. Brambilla et al. (1994) introduced the term "late perimenopause" and defined women as being in the late stage of the transition by self-report of 3-9 months of amenorrhea or menstrual irregularity. Subsequently, investigators from several longitudinal studies (Melbourne Women's Midlife Health Project [MWMHP] (Dennerstein et al. 1993), Seattle Midlife Women's Health Study [SWMHS] (Mitchell et al. 2000), TREMIN (Treloar et al. 1967)) proposed various bleeding criteria to define the transition period (Taffe and Dennerstein 2002, Mitchell et al. 2000, Mansfield et al. 2004, Lisabeth et al. 2004a).

STRAW defined stages principally by changes in menstrual bleeding characteristics and, to a lesser extent, by changes in serum follicle-stimulating hormone (FSH) levels (Soules et al. 2001). STRAW divided reproductive life prior to menopause into the reproductive years (3 stages) and the transition years (2 stages, early and late transition). Entry into the early transition is characterized by increased variability in menstrual cycle length while entry into the late transition is characterized by the occurrence of skipped cycles or amenorrhea. The STRAW recommendations

(Soules et al. 2001), although based on emerging results of the large cohort studies of midlife women, were not data-driven. The multi-study ReSTAGE Collaboration subsequently evaluated bleeding criteria that served as the basis of the STRAW recommendations and documented the extent to which the various proposed criteria identified a similar moment in women's reproductive life (Harlow et al. 2006, 2007, 2008). All of these proposals attempt to define bleeding criteria that identify a change-point in women's menstrual cycle histories. Notably, however, none of the papers attempted to model these changepoints longitudinally.

Harlow et al. (2000) longitudinally modeled change in mean cycle length, as well as in between-woman and within-woman variance across the reproductive lifespan and found that within-woman heterogeneity in cycle length was an important source of variation in menstrual patterns, especially after age 40. They fitted a bipartite cubic spline model that modeled the risk of both very short and very long segments using changepoints fixed at ages 34 and 40. Lisabeth et al. (2004b) used generalized estimating equations to model changes in mean cycle length and variance independent of the mean referenced to age at FMP and demonstrated that variance in menstrual cycle lengths increase on average 2 to 6 years before increases in the mean, depending on age at FMP.

Prior descriptive analyses also suggest that there is some heterogeneity in women's menstrual trajectories. Menstrual characteristics in young adult women are associated with fertility (Small et al. 2006) and the timing of menopause (Den Tonkelaar et al. 1998, Wallace et al. 1979, Lisabeth et al. 2004b). A prior analysis of the TREMIN data by (Wallace et al. 1979) reported that women with later menopause had longer mean cycle length and greater variability two years before menopause than women with earlier menopause. Lisabeth and colleagues (Lisabeth et al. 2004b) in

a longitudinal analysis of the same data also reported that longer cycles were associated with a later age of menopause. Another study (Den Tonkelaar et al. 1998) reported that women with a late age at menopause (55-59) had a longer mean cycle length in the nine years prior to menopause than women with an earlier menopause. Weinstein et al. (2003) found that low serial irregularity, a measure of the variability of the changes in cycle length, was associated with younger age at FMP, after adjusting for age at menarche, number of births, and hormone use.

Our goal is to model how menstrual cycle length and variability change when women approach menopause. We assume that there are underlying unknown mean and variance changepoints for each individual woman and build a Bayesian hierarchical change point model to estimate distributions of these changepoints. Furthermore, we impute cycles that are missing due to hormone use, gaps in the menstrual calendar, and gynecological surgery, allowing more subjects and information to be included. Most prior reports have censored women when they began using hormonal contraceptives or hormone therapy (HT)(Weinstein et al. 2003, Guo et al. 2006, Harlow et al. 2006, 2008).

Statistically, the objective is to model both the mean and variance of a set of curves. Several approaches have been proposed for correlated functional data of this type, including the bipartite spline model proposed by (Harlow et al. 2000) which modelled mean and between-subject variance by a linear random effect model and used a two-stage log-linear regression to study within-subject variance vs. age. Crainiceanu et al. (2007) proposed Bayesian penalized splines to model both mean and variance by using a set of fixed knots for the splines with structural covariance matrix and random effects to depict the heterogeneity of variance. Lisabeth et al. (2004b) modeled means and variances over time separately using independent gener-

alized estimating equations. Gunn and Dunson (2005) modeled hormone patterns in the menstrual cycles using a Bayesian hierarchical model and mapped the posterior draws to a constrained space which guarantees that each curve increases monotonically to an unknown changepoint and decreases afterwards. To model student test achievement, Thum and Bhattacharya (2001) proposed a hierarchical Bayesian regression model which included two-phase composite of $y_i \sim N(\beta_{01} + \beta_{11}x_i, \sigma_1^2), i = 1, 2, \ldots, k$ and $y_i \sim N(\beta_{02}+\beta_{12}x_i, \sigma_2^2), i = k+1, k+2, \ldots, n$ where k was the unknown change point. Hall et al. (2003) used unknown change points for the splines to capture individual cognitive function over time. These approaches estimated unknown changepoints for the mean but did not model the variance function over time. Davidian and Carroll (1987) proposed another approach for variance function estimation, which models the variance as proportional to a power of the mean response. This approach builds a separate function to model variance but did not include changepoints. Here we consider a Bayesian hierarchical model that estimates individual-level mean and variance profiles with unknown changepoints. These changepoints represent measures of menopausal transition, and, together with intercepts and pre- and post-changepoint slopes, provide detailed summaries of the menstrual cycle data that can be related to individual level covariates such as age at menarche, parity, and secular cohort membership.

Our article is organized as follows. In section 3.2 we describe the TREMIN study data. In section 3.3 we describe a Bayesian model to study the trajectories of women's menstrual cycle length that estimates unknown changepoints for both means and variances and allows these changepoints to be functions of subject-level covariates. In addition, we impute different forms of missingness in the data set and incorporate the imputation in the Markov Chain Monte Carlo sampling used to estimate model

parateters. In section 3.4 we give the results from fitted model to menstrual data, along with Bayesian posterior predictive model checks. In section 3.5 we discuss how our results compare to and extend previous menstrual cycle staging research, along with possible extensions of our model.

3.2 The TREMIN Dataset

Our models are designed for the TREMIN data, one of only two data sets available providing individual women's menstrual calendar data across their reproductive life span. The study, initiated by Dr. Alan Treloar (Treloar et al. 1967), recruited the first cohort of TREMIN: 2350 college-aged women attending the University of Minnesota between 1934 and 1939.

Definitions recommended by WHO (Belsey and Farley 1987) were used to summarize the calendar data. A bleeding segment, analogous to the term menstrual cycle, is a period of consecutive bleeding days and the subsequent bleeding-free days. Bleed-free intervals had to consist of at least 3 days; 1-2 bleed-free days between 2 bleeding days were considered part of the bleeding episode. Bleeding segment length is the dependent variable in our study. Age at menopause is determined by the date of final menstrual period, which is attributed retrospectively after 12 months of amenorrhea on the calendar cards (WHO 1996).

We used data from 617 women in the 1935-1939 cohort who were a) age 25 or less at enrollment, b) used hormones for less than four years continuously, c) had at least one observed segment before age 40, and d) were not censored before age 40 (Data tape TRUST998.FINAL, March 1993). We consider segment lengths beginning at age 35. After this left truncation, the data set has a total of 95,246 observed menstrual segment records. Each record consists of woman's age, bleed-

ing segment length, and status indicators for pregnancy, hormone use and surgery. Related subject-level information including age at menarche and parity are also available.

Pregnancy intervals as well as the first two segments after a birth and the first segment after a spontaneous abortion are coded as non-menstrual intervals. Many women used exogenous hormones at some point during their reproductive lives, mainly as hormonal replacement therapy. When hormones are used, the bleeding segment is coded as a treated interval, during which ovarian function is masked. Thus, the segment data are considered to be missing when women use hormones. A one-segment washout period after hormone use ended was also treated as missing. Many studies of menstrual characteristics censor women when they begin hormone use or ignore the time period during which women are using hormones. However, Wegienka and Baird (2003) suggested that these strategies may introduce bias since hormone users are not a random sample of menstruating women. Omitting these women or portions of their data will provide an incomplete description of experiences in the overall population. In our analysis, we consider these data as missing and impute their values for hormone use gaps of up to four years. Studies have not found that hormonal use influences menstrual segment length after stopping use and allowing for a washout period (Taylor et al. 1977, Treloar and Behn 1971).

The 617 women included in our analysis each contributed between 15 and 321 non-missing segments to the analysis. The observed segment lengths vary from 4 to 366 days with a median of 27 days. Final menstruation periods were observed for 313 subjects (50.7%). Only 105 (17.0%) have complete data.

Figure 3.1 displays log segment lengths for four typical women in the TREMIN data set. Subject A has complete data. She has a pregnancy gap which is not

included in the analysis, no gynecological surgery or periods of hormone use, and has an observed final menstruation period. Subject B was coded as using hormones from age 36.07 to age 37.24, and her information for this period is treated as missing. Her FMP is observed, however. Subject C has intermittent missingness at age 36.95. She had a hysterectomy at age 45.78, thus her menstrual history was truncated at this point and no FMP was observed. Subject D has intermittent missing at age 39.59 and from age 41.56 to 43.64. She began hormone therapy after age 50.21, with no untreated bleeds recorded afterwards; thus no FMP was observed.

3.3 Modeling Menstrual Cycle Data

We construct a Bayesian change point model for the mean and variance of the segment length.

3.3.1 Change Point Model for Mean and Variance

Let y_{it} denote the t^{th} menstrual segment length of subject i. Let a_{it} denote the age at the beginning of the t^{th} menstrual segment of subject i, where $i = 1, \ldots, N$, $t = 1, \ldots, T_i$, $N = 617$.

We consider a log-normal model with a linear change point for both the mean and variance for each subject:

$$log(y_{it})|\mu_{it}, \sigma^2_{it} \sim N(\mu_{it}, \sigma^2_{it})$$

$$\mu_{it} = \alpha^{\mu}_i + \beta^{\mu}_i(a_{it} - 35) + \gamma^{\mu}_i(a_{it} - \theta^{\mu}_i)_+$$

$$log(\sigma^2_{it}) = \alpha^{\sigma}_i + \beta^{\sigma}_i(a_{it} - 35) + \gamma^{\sigma}_i(a_{it} - \theta^{\sigma}_i)_+$$

The function $(x)_+ = x$ if $x \geq 0$, $(x)_+ = 0$ if $x < 0$; θ^{μ}_i and θ^{σ}_i are the unknown change points of mean and variance for subject i. The change points create a linear spline

for each mean and variance model. We denote these eight subject-level parameters for each woman as $\Phi_i = (\alpha_i^\mu, \beta_i^\mu, \gamma_i^\mu, \theta_i^\mu, \alpha_i^\sigma, \beta_i^\sigma, \gamma_i^\sigma, \theta_i^\sigma)'$.

To link the subject-level models, we postulate a multivariate normal prior for the subject-level parameters:

$$\Phi_i \overset{ind}{\sim} N(x_i'\Lambda, \Omega)$$

where x_i are covariates associated with subject i. Thus Λ and Ω can also be considered as population level parameters, with Λ as the regression coefficients and $\Omega \otimes I_N$ as the covariance matrix for the regression of Φ_i on x_i.

We complete the model specification by postulating an Inverse-Wishart hyperprior for Λ and Ω:

$$p(\Lambda, \Omega) = \text{Inv-Wishart}(\Omega; 1, I)$$

which is completely flat for Λ and weakly informative for Ω.

3.3.2 Posterior Inference

Let $z_{it} = log(y_{it})$. The goal of our analysis is to obtain inference on the joint posterior distribution of Φ, Λ, and Ω conditional on the observed data z^{obs}. The posterior based on the complete data z is given by

$$
\begin{aligned}
& p(\boldsymbol{\Phi}, \Lambda, \Omega | \boldsymbol{z}) \\
& \propto \prod_{i=1}^{N} \left[\prod_{t=1}^{T_i} p(z_{it}|\Phi_i) p(\Phi_i|\Lambda, \Omega) \right] p(\Lambda, \Omega) \\
& \propto \left[\prod_{i=1}^{N} \left[\prod_{t=1}^{T_i} \frac{1}{\sigma_{it}} \exp(-\frac{(z_{it}-\mu_{it})^2}{2\sigma_{it}^2}) \right] |\Omega|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\Phi_i - x_i'\Lambda)'\Omega^{-1}(\Phi_i - x_i'\Lambda)) \right] \\
& \qquad \times |\Omega|^{-\frac{k+2}{2}} \exp(-\frac{1}{2} tr(\Omega^{-1})) \\
& = \left[\prod_{i=1}^{N} \prod_{t=1}^{T_i} \sigma_{it}^{-1} \right] |\Omega|^{-\frac{N+k+2}{2}} \exp\left\{ \sum_{i=1}^{N} \left[\sum_{t=1}^{T_i} \frac{(z_{it}-\mu_{it})^2}{\sigma_{it}^2} + (\Phi_i - x_i'\Lambda)'\Omega^{-1}(\Phi_i - x_i'\Lambda) \right] \right. \\
& \qquad \left. + tr(\Omega^{-1}) \right\}
\end{aligned}
$$

We sample the parameters via a MCMC algorithm that uses Metropolis-within-Gibbs sampling. Details of the procedure are in Appendix A.

Missing data are imputed under a missing at random (MAR) assumption (Little and Rubin 2002) using a standard selection model. Imputation is embedded within the MCMC algorithm. Details are provided in the next section.

3.3.3 Imputation of Missing Data

The majority (512 of the 617 women) have some form of missing data. For 313 women, their segment lengths are censored due to dropout while still menstruating, surgical termination of menstruation due to hysterectomy or bilateral oophorectomy, or hormone use that began before FMP and continued past FMP. For the remaining 207 women, missingness was only intermittent. Intermittent missingness occurred due to sporadic non-reporting (women failing to report an individual segment or series of segments), or to periodic hormone use that stopped before one of the censoring events.

There is concern that missingness, particularly missingness due to hormone use, is

not missing completely at random. In order to deal with the different types of missingness, we impute the missing data under a missing at random (MAR) assumption. To ensure that the imputation is proper (i.e., fully conditions on the observed data), we need to ensure that the imputed segment lengths sum to the length of the gap between observed segments. In addition, when censoring is present, we need to estimate the age of the FMP in order to terminate the imputation process.

When missingness is intermittent, we ensure that the imputed missing segment lengths sum to the length of the gap using an importance sampling algorithm. For notational simplicity, we assume that we have a single missing gap of length L_i for subject i, starting after segment y_{ik}. Conditional on Φ_i, the unobserved segment lengths $(y_{i,k+1}, ..., y_{i,k+S})' = \tilde{y}_i$ in the gap are independent, subject to the constraint that $\sum_{s=1}^{S} y_{i,k+s} = T$. We obtain a draw $\log(y_{i,k+1}^{rep}) \sim N(\mu_{i,k+1}, \sigma_{i,k+1}^2)$ where $\mu_{i,k+1} = \alpha_i^{\mu} + \beta_i^{\mu}(a_{i,k+1} - 35)_+ + \gamma_i^{\mu}(a_{i,k+1} - \theta_i^{\mu})_+$ and $\sigma_{i,k+1}^2 = \exp(\alpha_i^{\sigma} + \beta_i^{\sigma}(a_{i,k+1} - 35)_+ + \gamma_i^{\sigma}(a_{i,k+1} - \theta_i^{\sigma})_+)$ and $a_{i,k+1} = a_{ik} + y_{ik}$ is the age of the start of segment $y_{i,k+1}^{rep}$. A draw of $y_{i,k+2}^{rep}$ is then obtained as for $y_{i,k+1}^{rep}$, where now $a_{i,k+2} = a_{i,k+1} + y_{i,k+1}^{rep}$. This process is repeated until we obtain $y_{i,k+S}^{rep}$ such that $\sum_{s=1}^{S} y_{i,k+s}^{rep} > L_i$. We then replace $y_{i,k+S}^{rep}$ with $\tilde{y}_{i,k+S}^{rep} = L_i - \sum_{s=1}^{S-1} y_{i,k+s}^{rep}$. Let $(y_{i,k+1}^{(t)}, ..., y_{i,k+S-1}^{(t)}, \tilde{y}_{i,k+S}^{(t)}) = \tilde{y}_i^{(t)}$ be the t^{th} vector of imputations, $t = 1, ..., 50$. Finally, we draw one of the 50 sets with probability $p^t = \frac{f(\tilde{y}_i^{(t)}|\Phi_i)}{\sum_t f(\tilde{y}_i^{(t)}|\Phi_i)}$, where $f(\tilde{y}_i^{(t)}|\Phi_i) = \prod_{s=1}^{S-1} \phi\left(\frac{\log(y_{i,k+s}) - \mu_{i,k+s}}{\sigma_{i,k+s}}\right) \times \phi\left(\frac{\log(\tilde{y}_{i,k+S}) - \mu_{i,k+S}}{\sigma_{i,k+S}}\right)$, where $\phi(\cdot)$ is the pdf of the standard normal distribution. On rare occasions where $y_{i,k+s}^{rep} < 4$, the imputed values were truncated to be 4; similarly $y_{i,k+s}^{rep} > 365$ was truncated to 365.

When subjects' segment lengths are censored, we need to impute an FMP since it is unobserved. We model the age at FMP Q_i as a piecewise exponential distribution with hazard $h_i(t) = \eta_k$ for $A_{k-1} \leq t < A_k$ for knots $k = 1, ..., K$. Knots are set at age

40, 42, 43, 44, 45, 46, 46.5, 47, 47.5, 48, 48.5, 49, 49.5, 50, 50.5, 51, 51.5, 52, 52.5, 53, 53.5, 54, 55, 56, 57, and 60. Assuming a prior of the form $\eta_k \sim \text{GAMMA}(a, b)$, obtain a draw from $p(\eta_k \mid Q) \sim \text{GAMMA}(\sum_i I(A_{k-1} \leq Q_i \leq A_k) + a, \sum_i I(Q_i \geq A_{k-1}) + b)$ for $k = 1, \ldots, K$, where Q includes both the observed FMP and those imputed at the previous iteration of Gibbs sampler (see Appendix B). As in the intermittent missing setting, we then obtain a draw $\log(y^{rep}_{i,T^{rep}_i+1}) \sim N(\mu_{i,T^{rep}_i}, \sigma^2_{i,T^{rep}_i})$ where $\mu_{i,T^{rep}_i} = \alpha^\mu_i + \beta^\mu_i(a_{i,T^{rep}_i} - 35)_+ + \gamma^\mu_i(a_{i,T^{rep}_i} - \theta^\mu_i)_+$ and $\sigma^2_{i,T^{rep}_i} = exp(\alpha^\sigma_i + \beta^\sigma_i(a_{i,T^{rep}_i} - 35)_+ + \gamma^\sigma_i(a_{i,T^{rep}_i} - \theta^\sigma_i)_+)$, $a_{i,T^{rep}_i} = a_{iT^{rep}_i - 1} + y_{iT^{rep}_i}$ is the age of the start of segment y^{rep}_{i,T^{rep}_i}, and T^{rep}_i is the number of observed segments plus the number of imputed segments in any intermittent missing gaps. Let $W_{i,1}$ be an indicator for whether this first imputed cycle is FMP, we then obtain a draw $W_{i,1}$ from a Bernoulli distribution with probability (see Appendix B)

$$
\begin{aligned}
&P(a_{i,T^{rep}_i} \leq Q_i \leq a_{i,T^{rep}_i+1} \mid Q_i > \max(a_{i,T^{rep}_i}, \theta^\mu_i, \theta^\sigma_i)) \\
&= \begin{cases}
\left[1 - e^{-\eta_k(a_{i,T^{rep}_i+1} - a_{i,T^{rep}_i})}\right] I\left[a_{i,T^{rep}_i} > \max(\theta^\mu_i, \theta^\sigma_i)\right] \\
\qquad \times I\left[A_{k-1} \leq a_{i,T^{rep}_i} < a_{i,T^{rep}_i+1} \leq A_k\right] \\
\left[1 - e^{-\{\eta_k(a_{i,T^{rep}_i+1} - A_{k-1}) - \eta_{k-1}(a_{i,T^{rep}_i} - A_{k-1})\}}\right] I\left[a_{i,T^{rep}_i} > \max(\theta^\mu_i, \theta^\sigma_i)\right] \\
\qquad \times I\left[A_{k-2} \leq a_{i,T^{rep}_i} \leq A_{k-1} \leq a_{i,T^{rep}_i+1} \leq A_k\right] \\
\left[1 - e^{-\{\eta_{k+1}(a_{i,T^{rep}_i+1} - A_k) + \eta_k(A_k - A_{k-1}) - \eta_{k-1}(a_{i,T^{rep}_i} - A_{k-1})\}}\right] \\
\qquad \times I\left[a_{i,T^{rep}_i} > \max(\theta^\mu_i, \theta^\sigma_i)\right] \\
\qquad \times I\left[A_{k-2} \leq a_{i,T^{rep}_i} \leq A_{k-1} < A_k \leq a_{i,T^{rep}_i+1} \leq A_{k+1}\right]
\end{cases}
\end{aligned}
$$

Note that the FMP must occur after both the last observed segment and the latent mean and variance changepoints; also, since none of our knots are less than six months apart, a segment can cover a maximum of 3 intervals. If $W_{i,1} = 1$,

$y^{rep}_{i,T_i^{rep}+1}$ is the length of the final FMP. If $W_{i,1} = 0$, we draw $\log(y^{rep}_{i,T_i^{rep}+2}) \sim N(\mu_{i,T_i^{rep}+1}, \sigma^2_{i,T_i^{rep}+1})$ and repeat the process s times until one of the following occurs: $W_{i,s} = 1$, $y^{rep}_{i,T_i^{rep}+1} > 365$ or $a^{rep}_{i,T_i^{rep}+s} \geq 60$. For the vast majority of subjects, the FMP variable trigged the end of the imputation.

3.4 Results

We use the methodology described in section 3 to analyze TREMIN data using MATLAB software. We ran two MCMC chains for 10,000 iterations each after discarding the first 10,000 draws as "burn-in". We assessed convergence using the Gelman and Rubin statistic (Gelman et al. 2004), with a thinning interval of 5 segments. All of the population and 98% of the individual-level parameters had a value of less than 1.2, indicating reasonable convergence.

3.4.1 Individual Level Parameters

To visually assess model fit at the individual level, Figure 3.2 plots the observed segment lengths and predicted means and variances for the same four sampled women described in Figure 3.1. The model appears to capture the trajectories well, with approximately 5% of cycle lengths excluded from the 95% predictive intervals. The uncertainty in the position of the variance changepoint is highlighted in (b) and (c).

Figure 3.3 plots the posterior means and 90% credible intervals of the mean and variance changepoints for 50 randomly selected women. As noted by Treloar (1981) and Lisabeth et al. (2004b), variability generally begin to increase before mean length. Subjects with earlier changepoints averaged 4-5 years between mean and variance changepoints, whereas subjects with later changepoints averaged only 1-2 years between mean and variance changepoints, consistent with the findings of Harlow et al. (2008). Uncertainty in the variance changepoints is generally greater

than in the mean changepoints.

Figure 3.4 plots the posterior medians of the mean and variance changepoints against the final menstural periods for the 315 women with observed FMPs. FMPs occured on average 3.6 years after the mean changepoint, with a standard deviation of 1.4 years. FMPs occured on average 6.5 years after the variance changepoint, with a standard deviation of 2.5 years. The mean time to FMP after the mean changepoint was fairly constant with respect to age at mean changepoint; mean time to FMP after the variance changepoint was considerably shorter in women with later variance changepoints than in younger women.

3.4.2 Population Level Parameters

Table 3.1 summarizes the posterior means and associated 95% credible intervals for the population level segment length mean and variance regression parameters. The population mean age at the changepoint for segment length means is 46.20 years (95% CI 45.87-46.54 years), older than the population mean age at changepoints for segment length variability, which is 42.21 years (95% CI 41.84-42.58 years); thus variability in segment length is predicted to begin increasing 3.24 years earlier (95% CI 2.97-3.51 years) in the population than the mean segment length itself. Mean segment length declined about 1% per year before the changepoint and increased about 15% per year afterwards. Variability of log-segment length was stable before the changepoint and increased by 79% per year after the changepoint.

Table 3.2 presents the posterior mean and associated 95% posterior predictive interval for the correlation matrix corresponding to the covariance matrix Ω. The 95% credible intervals of correlations that exclude zeros are denoted in bold.

- Later changepoints for variance are highly associated with later changepoints

for mean.

- Later changepoints for both mean and variance are also correlated with longer and more variable segment lengths, and more rapid increases in mean and variance after the changepoint; consequently mean and variance slopes after changepoints are positively correlated.
- Greater mean length is associated with greater declines in variability before the variance changepoint and greater increases in variability after.
- Larger segment variability is associated with longer mean segment length.
- Larger segment variability is highly associated with more rapid declines in variability before but larger increases in variability after the variance changepoint: thus change in variability before and after the variance changepoint is negatively correlated.

We conducted a principal components analysis of Ω to determine if the relationships among the eight parameters governing perimenopause segment lengths could be summarized in a smaller number of dimensions. Table 3.3 shows that four components explained 82% of the variance of the individual level parameters governing menstrual segment length. The first component loads heavily on the inverse relationship between the slope of the variances before and after the variance changepoint, and on late mean and variance changepoints. The second component also loads on the inverse relationship between the slope of the variances before and after the variance changepoint, but picks up a relationship between early changepoints in means and variances and smaller increases in means after the mean changepoint. The third and fourth components load on the relationship between the mean intercepts and slopes: the third component relates longer mean segment lengths at age 35 with

more rapid declines in mean length before the mean changepoint and less rapid increases thereafter, while the fourth component relates shorter mean segment lengths at age 35 with more rapid declines in mean length before the mean changepoint and more rapid increases thereafter.

We also fit a two-covariate model, including parity and age at menarche. As covariates showed no significant relationships with the eight parameters describing the menopausal transition, we do not show the results here.

3.4.3 Posterior Predictive Model Check

We used posterior predictive distribution checks (Gelman et al. 1996) to assess model fit. We calculated the χ^2 discrepancy statistic for observed segment lengths of each individual woman given by $\sum_t \frac{(y_{it}-\mu_{it}^{rep})^2}{(\sigma_{it}^{rep})^2}$, which will have a $\chi^2_{T_i^{obs}}$ distribution if the model is correct, where T_i^{obs} is the total number of observed segments for the i^{th} woman. We assessed corresponding predictive p-values for these χ^2 test statistics based on 250 replications. Figure 3.5 shows the predictive p-values for all subjects. No subjects had a posterior predictive p-values greater than 0.95 and only one subject has a posterior predictive p-value smaller than 0.05. Review of subjects with low posterior predictive p-values show that they contain one or two sporadic very short or very long segments well before the onset of the increase in variability, suggesting that these subjects contain outlying segment lengths rather than indicating more general model failure. Subjects with high posterior predictive p-values generally had relatively few observations with little variability – the variance estimates were smoothed back toward larger values, yielding small χ^2 discrepancy statistics.

To consider the appropriateness of the final menstrual period modeling, we plot the observed and predicted FMPs together with the censoring ages for 100 randomly selected women in Figure 3.6. The method for estimating FMP when not observed

appears to have worked well, with the distribution for the predicted FMPs corresponding closely to the observed FMPs when the censoring age is relatively early and little information is usually available to predict FMP.

3.5 Discussion

In this article we have provided a Bayesian changepoint model for describing the patterns of means and variances of women's menstrual segment lengths as they approach menopause. Our model detects individual changepoints of mean and variance of segment lengths for each individual woman. The model is applied to the TREMIN data. Multiple imputations integrated with an MCMC chain are carried out to impute the different kinds of missingness in the data set. Instead of setting splines at a certain fixed point for all women and using traditional random effect models to study menstrual patterns (Harlow et al. 2000), our model allows the changepoints to be unknown parameters that vary for different subjects. This setting provides a flexible way of capturing both the mean and variability of each individual's segment length trajectory.

Our work develops a data-driven definition of early and late transition defined by subject-level variance and mean changepoints respectively. We observed a 3.2 year difference in age between mean and variance changepoints at the population level, somewhat shorter than that of Lisabeth et al. (2004b), who reported a 3.9 year difference between cycle lengths with standard deviations of 6 days and the first cycle of 60 days or more. In addition, our results were consistent with those of Wallace et al. (1979), Den Tonkelaar et al. (1998), and Lisabeth et al. (2004b), who found that longer mean segment lengths were associated with later FMPs. Our results were also consistent with those of Weinstein et al. (2003), who found that lower

variability was associated with early FMPs. We further found relationships between rates of change in length and variability before and after changepoints themselves, in particular that greater baseline variability was associated with more rapid declines in variability before variance changepoints and greater increases thereafter; and later mean changepoints were associated with greater increases in mean length and more mean variability after mean changepoints. These data contribute to efforts to define a staging system for reproductive aging as they further our understanding of the timing and duration of the menopausal transition and describe the nature of heterogeneity in women's experience.

Our next step is to add the second TREMIN cohort data to assess changes of women's menstrual pattern in different generations by adding secular cohort (1935-1970 vs. 1960-1995) as a population-level covariate to the model. Also, while model checking showed that the model provides an adequate fit to the data, the model might still be improved. Distributions of individual level variance parameters (not shown) are somewhat skewed or heavy-tailed, suggesting a mixture distribution might be more suitable than one normal distribution for all subjects. Thus, a latent class model with subjects belonging to one of several underlying categories might fit the data even better. Estimation in the presence of left censoring is also of interest as many recent and ongoing studies enrolled prevalent cohorts including women who had already begun the menopausal transition.

Carroll (2003), in a paper entitled "Variances are not Always Nuisance Parameters," called for increased focus on developing methods for "variance structures" in order to better understand how "systematic dependence of variability on known factors" could yield both better prediction and improved inference. We agree with Carroll that incorporating information from subject-level variability in longitudinal

data settings is underutilized in clinical and epidemiological research settings, at least in part because of the lack of methods for such analysis. In our application, it would be of interest to identify sub-groups of women who experience distinct patterns of variability during the menopausal transition and evaluate whether these subgroups also differ in their risk for developing chronic disease. We believe the analysis provided here begins to fill in some of the gaps in this area.

Parameter	Λ (95%CI)
Mean intercept	3.313(3.306,3.321)
Mean slope before changepoint	-0.007(-0.010,-0.003)
Mean slope after changepoint	0.139(0.124,0.155)
Segment length mean changepoint	45.95(45.66,46.24)
Log-variance intercept	-4.814(-4.927,-4.704)
Log-variance slope before changepoint	0.016(-0.015,0.047)
Log-variance slope after changepoint	0.583(0.528,0.636)
Segment length variance changepoint	42.71(42.38,43.03)

Table 3.1: Posterior mean of population level regression coefficients (Λ) estimates and associated 95% posterior predictive intervals.

	Mean intercept	Mean slope before changepoint	Mean slope after changepoint	Mean changepoint	Log-variance intercept	Log-variance slope before changepoint	Log-variance slope after changepoint	Variance changepoint
Mean intercept	1.00	**-0.11**	-0.02	**0.28**	**0.16**	**-0.12**	**0.15**	**0.27**
	–	**(-0.19,-0.03)**	(-0.14,0.09)	**(0.19,0.36)**	**(0.07,0.25)**	**(-0.22,-0.03)**	(0.06,0.24)	**(0.19,0.35)**
Mean slope		1.00	-0.02	-0.01	-0.04	0.02	-0.01	-0.01
before changepoint		–	(-0.10,0.07)	(-0.10,0.07)	(-0.12,0.04)	(-0.06,0.10)	(-0.10,0.07)	(-0.10,0.07)
Mean slope			1.00	**0.28**	0.07	-0.08	**0.31**	**0.26**
after changepoint			–	**(0.16,0.40)**	(-0.05,0.25)	(-0.20,0.05)	**(0.20,0.42)**	**(0.13,0.38)**
Mean changepoint				1.00	**0.16**	**-0.25**	**0.46**	**0.81**
				–	**(0.07,0.25)**	**(-0.37,-0.15)**	**(0.37,0.55)**	**(0.77,0.85)**
Log-variance					1.00	**-0.69**	**0.45**	0.04
intercept					–	**(-0.74,-0.65)**	**(0.38,0.53)**	(-.04,0.13)
Log-variance slope						1.00	**-0.75**	0.06
before changepoint						–	**(-0.79,-0.70)**	(-0.05,0.16)
Log-variance							1.00	**0.37**
slope after changepoint							–	**(0.29,0.45)**
Variance								1.00
changepoint								–

Table 3.2: Posterior means and 95% credible intervals of correlations of population level regression coefficients (corresponding to covariance matrix Ω). Credible intervals that exclude zero are boldfaced.

	Loadings			
	PC1	PC2	PC3	PC4
Mean intercept	0.22	-0.14	0.63	-0.32
Mean slope before changepoint	-0.03	0.00	-0.61	-0.75
Mean slope after changepoint	0.24	-0.19	-0.44	0.55
Mean changepoint	0.45	-0.40	-0.01	-0.10
Log-variance intercept	0.36	0.45	0.09	-0.06
Log-variance slope before changepoint	-0.42	-0.50	0.02	0.03
Log-variance slope after changepoint	0.51	0.16	-0.15	0.06
Log-variance changepoint	0.36	-0.55	0.00	-0.10
Cumulative percent variance explained	35.3	56.1	69.9	81.9

Table 3.3: Principal components analysis of menstrual segment length parameters.

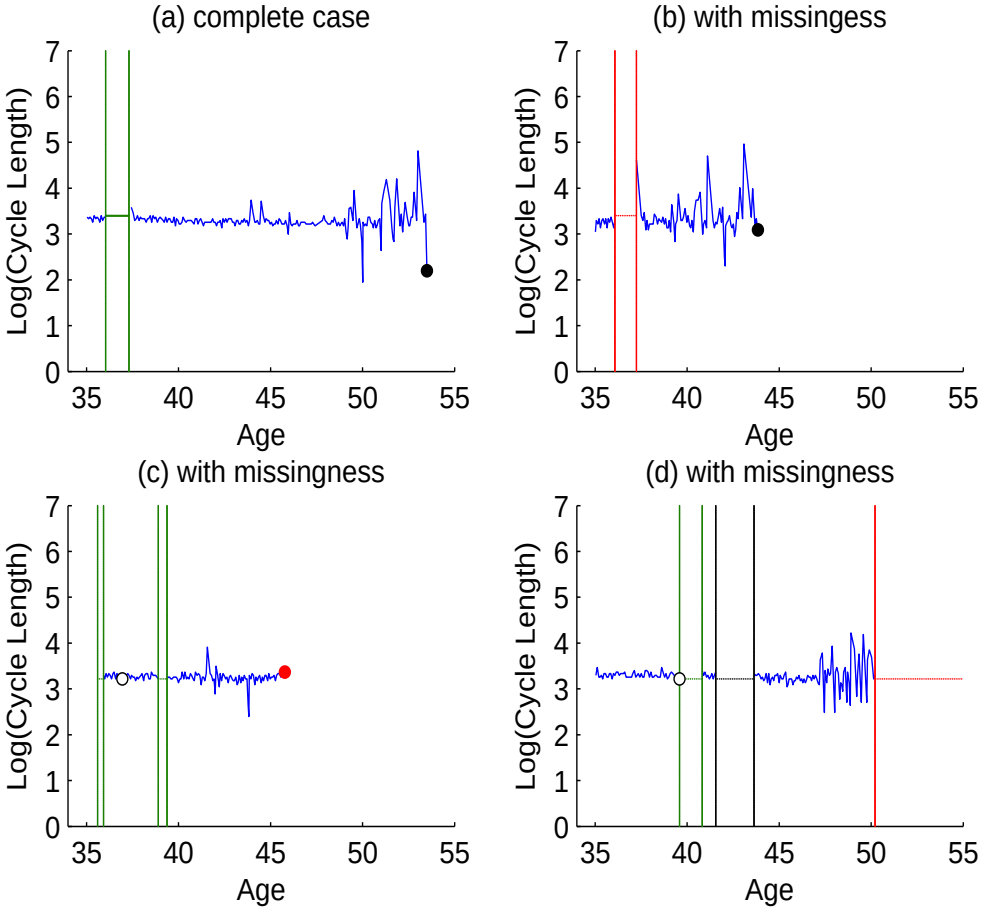

Figure 3.1: Four sampled women's log-segment-length trajectory after age 35: subject (a) has no missing data, the green gap is due to pregnancy and no imputation is needed. The black dot at the end means that FMP was observed for this subject. Subject (b) used hormones for a period of time, the red gap is due to hormone use. FMP is observed for this subject. Subject (c) has two pregnancy gaps (green gaps) and intermittent missingness at around age 36 (black circle). The red dot at the end represent that the subject's menstruation was truncated by surgery. Subject (d) has a pregnancy gap (green gap), an intermittent missingness (black circle) and a loss of contact gap (black gap). Her menstruation was censored due to hormone use (red line) and missing afterwards.

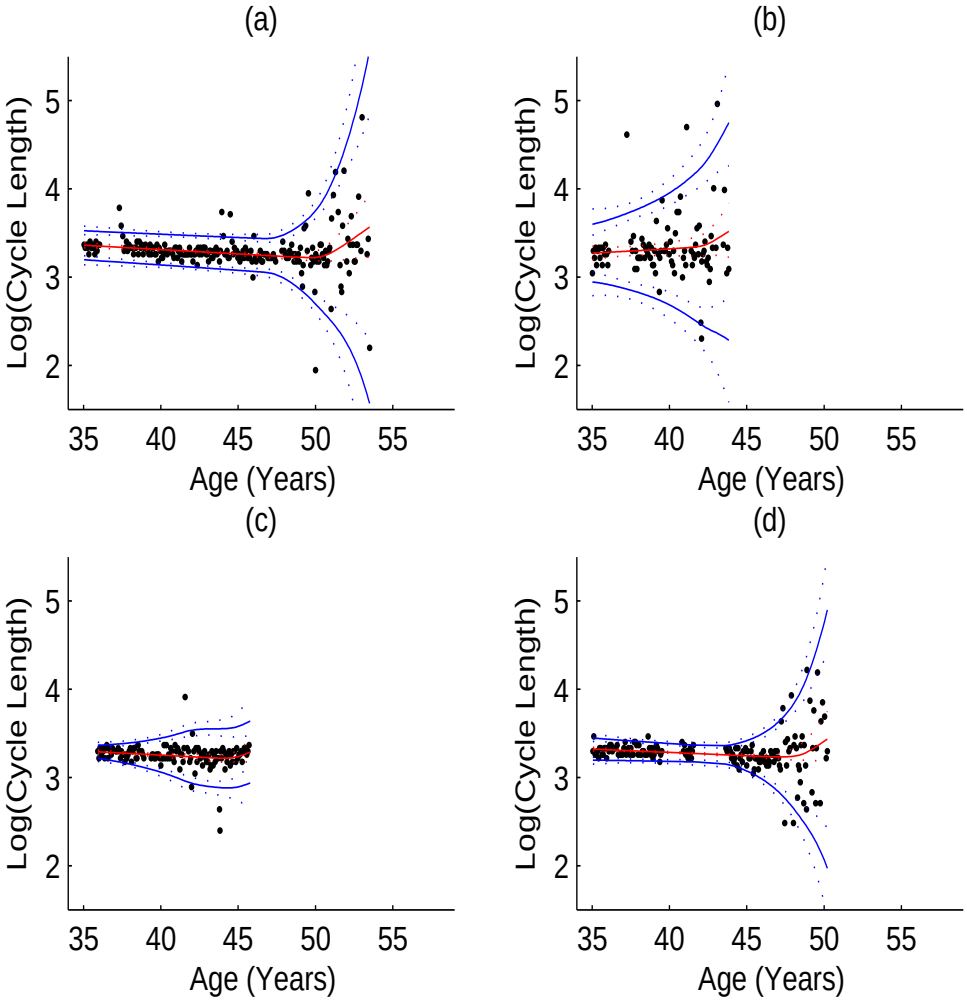

Figure 3.2: Changepoint model applied to the data for the four women in Figure 3.1: Red lines represent posterior mean of the mean segment length and associated 95% credible intervals; blue lines represent posterior mean for the upper and lower 2.5 percentiles for the segment distribution and their associated 95% credible intervals. Black dots represent log of observed segment lengths.

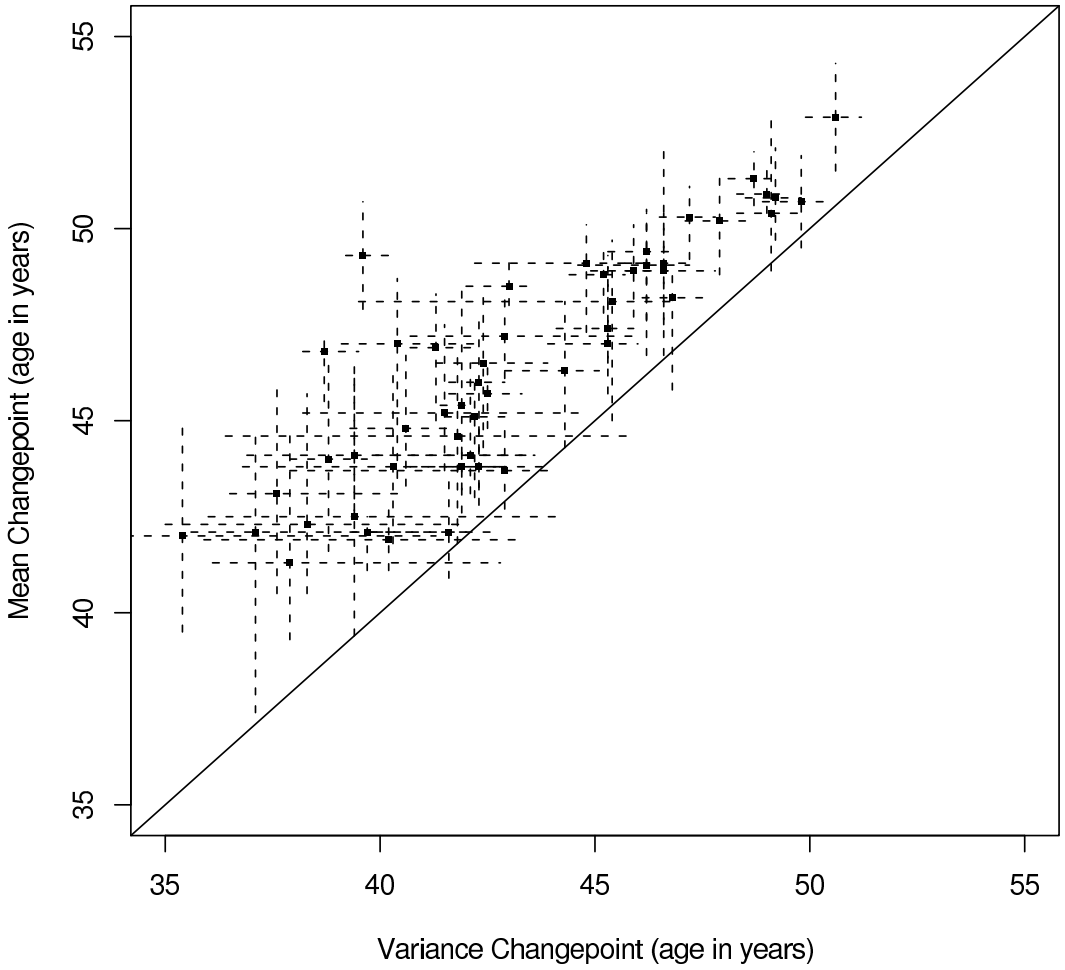

Figure 3.3: Posterior means and 90% posterior predictive intervals for mean changepoints and variance changepoints (100 randomly selected women).

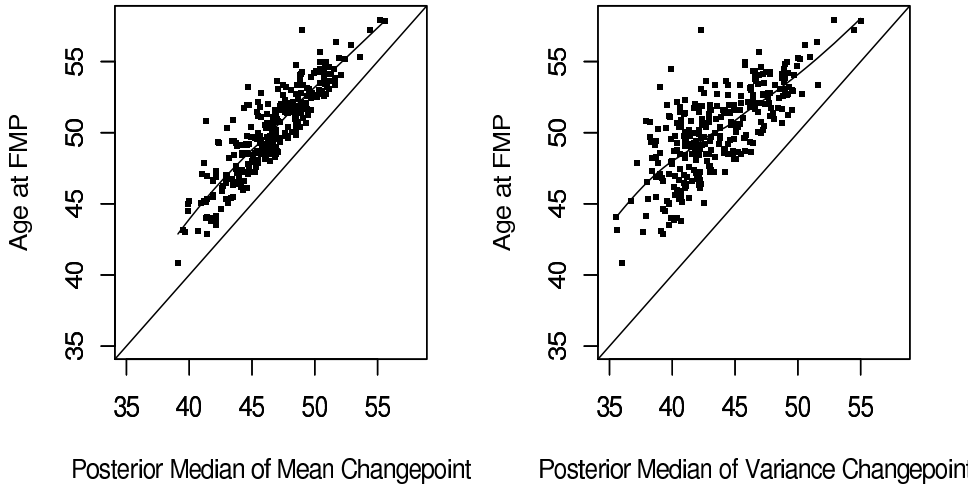

Figure 3.4: Posterior means of mean and variance changepoints versus final menstrual period for 313 women with observed final menstrual periods. Lines show loess fit.

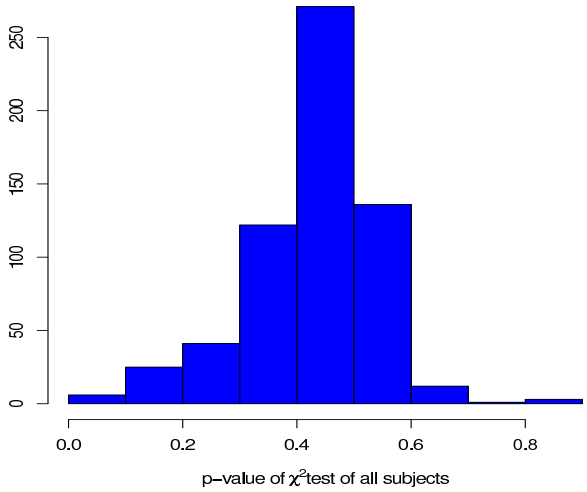

Figure 3.5: Histogram of p-values of subject level posterior predictive χ^2 tests.

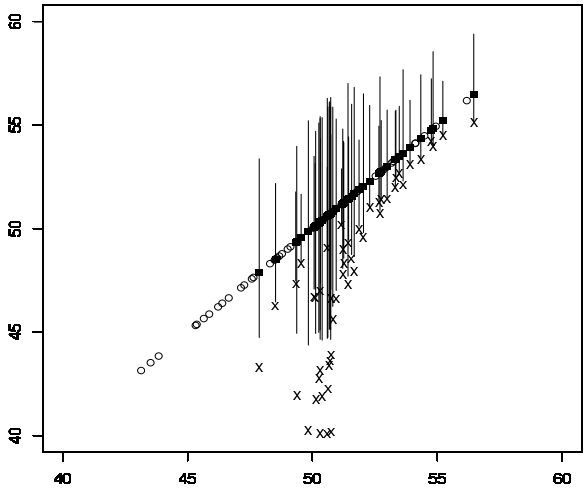

Figure 3.6: Observed FMP (circle) and posterior medians (squares) and 95% predictive interval for unobserved FMPs. X indicates age at censoring. (100 randomly selected women.)

CHAPTER IV

Patterns of Menstrual Bleeding and Their Relations to the Onset of Menopause, Markers of Stages and Health Outcomes

4.1 Introduction

Menstrual cycles can provide rich information about women's health status. They serve as markers for ovarian aging, endocrine disruption and endocrine risk factors for chronic diseases (Harlow 1995). Changes in women's menstrual patterns also predict the onset of menopause. Weinstein et al. (2003) found that a decrease of serial irregularity of menstrual cycles in late reproductive life is a strong predictor for the onset of menopause. Small et al. (2006) found menstrual cycle characteristics to be associated with fertility and spontaneous abortion.

Researchers have employed a range of methods to study menstrual patterns across women's reproductive lifespan (Treloar et al. 1967, Harlow et al. 2000). Most of these analyses describe the usual or average pattern of change in menstrual function over time. In Chapter III we quantified women's menstrual patterns across women's reproductive life using a more formal statistical modeling strategy than has typically been employed. We developed a Bayesian change point model with eight parameters for each woman: mean cycle length at age 35, along with rate of change in mean cycle length before and after a latent changepoint age; and equivalent parameters for

cycle variability. Few researchers have attempted to understand how these menstrual patterns may be classified into subgroups. Gorrindo et al. (2007) identified five types of menstrual patterns based on several key features including:

- variability of two-year running medians;
- mean interquartile range (IQR) of medians of each 2-year window;
- IQR consistency;
- slope of 5 year medians;
- stability of medians, based on count of stable 5-year intervals.

The five types they defined are (type I) very stable and consistent histories, flat (difference between 95th and 5th percentiles of less than 2 days) and with no evidence of a definable menopausla transition (fourth largest group); (type II) stable but more variable in cycle lengths, also flat, suggesting a muted menopausal transition experience (third largest group); (type III) oscillating or erratic cycles (difference between 95th and 5th percentiles of 2 or more but less than 5 days), with decreasing mean cycle lengths toward age 50, but no apparent increasing trend in cycle lengths at the approach of menopause (second largest group); (type IV) oscillating or erratic, with medium to high median IQRs and increasing cycle lengths as menopause approaches, which is the classic pattern with transition across age (largest group); (type V) highly erratic (difference between 95th and 5th percentiles of 5 days of more) and variable, typically driven by high early variability between age 15 and age 30 (smallest group with very few subjects). Women with stable menstrual patterns (type I and II) had later age of menarche and more births than women with more variable patterns (type III, IV, V). While Gorrindo et al. (2007) have a detailed classification system, their system was based on a visual inspection of cycle lengths over time, although specific quantitative measures of cycle length features were eventually used to classify

subjects. Moreover, the timing of transition was not considered.

Our goal is to quantitatively define subgroups of women based on menstrual patterns based on the eight characteristics defined in Chapter III and to study how pattern characteristics can predict the onset of menopause. Section 4.2 reviews the TREMIN data set used in this paper and the Bayesian changepoint model of Chapter III, which generated the individual level summaries of menstrual cycle characteristics. Section 4.3 discusses the relationship between individual summary measures and standard transition markers. Section 4.4 introduces the K-medoids algorithm used to cluster women to menstrual pattern subgroups based on the summary measures and relates these subgroups to age of menopause, age at menarche, number of births, as well as standard transition markers. Section 4.5 concludes with a discussion of the relationship between the clusters determined by our K-Medoids algorithm and those determined by Gorrindo et al. (2007), as well as plans for future work.

4.2 TREMIN Study and Bayesian Changepoint Model

We base our analysis on the TREMIN study, initiated by Dr. Alan Treloar at the University of Minnesota (Treloar et al. 1967). The study recruited the first cohort of 2350 college-age women attending the University of Minnesota between 1934 and 1939. Chapter III created a Bayesian changepoint model to describe women's menstrual patterns during their late reproductive life, from 35 years old until onset of menopause. Segments were assumed to follow a log-normal distribution with a linear change point for both the mean and variance for each subject:

$$\log(y_{it})|\mu_{it}, \sigma^2_{it} \sim N(\mu_{it}, \sigma^2_{it})$$

$$\mu_{it} = \alpha^{\mu}_i + \beta^{\mu}_i(a_{it} - 35) + \gamma^{\mu}_i(a_{it} - \theta^{\mu}_i)_+$$

$$\log(\sigma^2_{it}) = \alpha^{\sigma}_i + \beta^{\sigma}_i(a_{it} - 35) + \gamma^{\sigma}_i(a_{it} - \theta^{\sigma}_i)_+$$

The function $(x)_+ = x$ if $x \geq 0$, $(x)_+ = 0$ if $x < 0$. The model used eight parameters for each women to describe their individual menstrual pattern, including mean segment length at age 35 (α^{μ}_i), mean slope of segment length before changepoint (β^{μ}_i), mean slope of segment length after changepoint (γ^{μ}_i), segment length mean changepoint age (θ^{μ}_i), log-variance of segment length at age 35 (α^{σ}_i), slope of log-variance before variability changepoint (β^{σ}_i), slope of log-variance after variability changepoint (γ^{σ}_i), and segment length variability changepoint age (θ^{σ}_i). These eight subject-level parameters for each woman were denoted as $\Phi_i = (\alpha^{\mu}_i, \beta^{\mu}_i, \gamma^{\mu}_i, \theta^{\mu}_i, \alpha^{\sigma}_i, \beta^{\sigma}_i, \gamma^{\sigma}_i, \theta^{\sigma}_i)'$. Further, Chapter III used a multivariate prior $\Phi_i \overset{ind}{\sim} N(x_i'\Lambda, \Omega)$ to link the individual models, as well as a non-informative hyper prior $p(\Lambda, \Omega) = \text{Inv-Wishart}(\Omega; 1, I)$. Since no covariates were used, the prior parameter Λ and Ω can also be considered as the population mean and covariance matrix. The analysis included 95,246 observed menstrual segment records from 617 women in this cohort who were younger than 25 years old at enrollment, used hormones for less than four years continuously, had at least one observed segment before age 40 and were not censored before age 40. Missing segment lengths and missing final menstruation periods were imputed under a missing at random (MAR) assumption. Posterior means from the individual level parameter values are displayed in Figure 4.1.

4.3 Changepoint Characteristics in Late Reproductive Life

4.3.1 Changepoint Characteristics and Age of Final Menstruation Periods

To study how characteristics of menstrual patterns are related to the onset of menopause, we apply a linear accelerated failure time (AFT) model and semiparametric AFT model (Jin et al. 2006) to regress the eight identified characteristics of menstrual patterns on women's age at final menstruation periods (FMPs) separately and jointly. To accommodate subjects with censored FMPs, accelerated failure time models are used, which are parametric regression models assuming that the effect of a covariate is to multiply the predicted event time by some constant (Klein and Moeschberger 2003). In our application, we assume a linear relationship between failure times and covariates, rather than log-transformed failure time. Linear AFT model assumes that failure times are normally distributed. Semiparametric AFT model assumes failure times are linearly related to the covariates while leaving the error distribution unspecified. Jin et al. (2006) developed an iterative resampling technique to get inferences from the semiparametric AFT model.

All individual level posterior means of the eight characteristic parameters are standardized before analysis to facilitate comparisons of their influences of age at FMP.

Table 4.1 shows the associations between each characteristic and age at FMPs. In the linear AFT model, mean age at changepoints have the largest influence on FMP with one standard deviation increase resulting in FMPs occuring 2.52 (95% CI: 2.35 - 2.70) years later on average. Mean age at variance changepoints have the second largest influence on FMP, with one standard deviation increase associated with a 1.95 (95% CI: 1.70 - 2.19) year increase in age at FMP. Mean segment lengths at age 35 have the third largest influence on FMP, with one standard deviation increase

associated with a 1.12 (95% CI: 0.83 - 1.41) year increase in FMP. More rapid increase in mean slope after changepoint is associated to an earlier age of FMP. More rapid decrease in log-variance before changepoint is associated with an older age of FMP, while larger variance at age of 35 is associated with an older age of FMP. The rate of mean change before changepoint and rate of variance increase after changepoint have no significant influence on age at FMP. The semiparametric AFT model results are very similar to the linear AFT model.

As we model menstrual characteristics jointly, the story is different. Table 4.2 shows how all characteristics jointly influence age at FMPs. The influence of mean segment lengths at age 35 and changes of mean segment lengths before changepoint are no longer significant. Mean and variance changepoints' influence on FMPs decreases somewhat to third and fourth largest, while the influence of variance slope before and after changepoints substantially increases to the top two largest.

We next consider whether the changes in segment lengths and variability have multiplicative associations with age at FMP. Table 4.3 considers interactions between the before and after changepoints slopes on age at FMP after applying backward selection to eliminate non-significant influences. There is evidence of such a multiplicative effect for mean cycle lengths, with more rapid increases in segment lengths after changepoints associated with more rapid decreases before the changepoints and vice versa. No significant multiplicative effect was seen for variance slopes.

Finally, we consider the degree to which "early" measures of menopausal transition predict age at FMP by considering only mean and variability of segment lengths at age 35 together with the variance parameters. Table 4.4 suggests all of these five characteristics have significant influence on age at FMPs. Changes in variability before and after changepoint are still the top two most influential factors. Variance

changepoint is the third largest influence on age of FMP, followed by the influence of mean and variance of cycle lengths at age of 35.

From the three models jointly evaluating associations of women's menstrual pattern characteristics and their age of FMP (Table 4.2 - Table 4.4), we find that in general, women with older age of FMP tend to have less rapid decrease in variance of cycle lengths before changepoints, less rapid increase in mean and variance of cycle lengths after changepoints, later mean and variance changepoints, smaller variance ate age of 35, as well as less multiplicative effect of variance slopes before and after changepoints.

We then compare the predictive ability of FMP for the last three models: accelerated failure time model using all eight characteristics jointly (Model A), accelerated failure time model using all eight characteristics and interactions after backward selection (Model B), accelerated failure time model using with "early" information (Model C). Leave one out cross-validation for subjects with observed FMP are carried out to check the predictive ability. Table 4.5 calculates the mean square error of predicted FMP as well as adjusted R^2 measures of linear regression of predicted FMP versus observed FMP, also referring to Figure 4.2 for a visual view of predicted FMP versus observed FMP. All three models perform well. The model with "early" information has the lowest predictive ability, while the model with all characteristics and interactions after backward selection has the highest predictive ability. However, the difference is small across models.

4.3.2 Changepoints and Bleeding Markers of the Menopausal Transition

In 2001, the Stages of Reproductive Aging Workshop (STRAW) proposed a staging system for defining the end of women's reproductive life which includes early and late menopausal transition stages (Soules et al. 2001, Harlow et al. 2007). In-

creasing levels follicle stimulating hormone (FSH) is a signal of both early and late menopausal transition. In menstrual pattern changes, entry into early transition is characterized by increasing variability in menstrual cycle lengths (without skipping cycles), while skipped cycles or amenorrhea are typical characteristics of entry into the late transition. Different markers have been defined for early and late transition stages. We are interested in five markers based on menstrual bleeding data. Three of these markers have been proposed for defining the onset of late menopausal transition, including:

(a) D90: the first segment of at least 90 days (Brambilla et al. 1994);

(b) D60: the first segment of at least 60 days (Lisabeth et al. 2004a);

(c) RR10: the first instance of a running range of more than 42 days. The running range is computed as the difference between the maximum and minimum length of ten consecutive segments (Taffe and Dennerstein 2002).

Two of these markers have been proposed for defining the onset of early menopausal transition:

(d) Irregularity: the occurrence of more than two menstrual cycles outside the 21 to 35 day range over 10 cycles (Taffe and Dennerstein 2002).

(e) Diff6p: the first segment length more than 6 days different from the previous segment, and this magnitude of difference is observed again within 10 segments (persistent $>$ 6-day difference) (Mitchell et al. 2000).

We use correlations and loess regression to compare the mean and variance change-points and the four markers (D60, D90, RR10, Irregularity, Diff6p). Only subjects with observed markers are included in our analysis.

The characteristics of four markers are summarized in Table 4.6. We are interested in comparing age at these markers with age at with mean and variance

changepoints determined by the Bayesian changepoint model of Chapter III. Figure 4.3 through Figure 4.7 plot the markers against the changepoints, and Table 4.7 summarizes the correlations between markers and changepoints. All markers are more strongly correlated with mean changepoints than variance changepoints, even though mean changepoints are conceptually consistent with late transition and variance changepoints are conceptually consistent with early transition. Marker D90 has the highest correlations with both mean and variance changepoints, followed by RR10, D60 and Irregularity while marker Diff6p has lowest correlations with both changepoints, suggesting late markers are more correlated with changepoints.

We consider the relationship between the markers and changepoints in more detail after we discuss the clustering of model parameters with the K-medoids method.

4.4 Clustering of Change Menstrual Patterns in Late Reproductive Life

We use the K-medoids method to cluster women into different subgroups according to their eight characteristic summary measures. We also compare three important aspects of women's reproductive lives by classification types: age at menopause, age at menarche and number of births.

4.4.1 K-medoids Algorithm

The K-medoids method (Kaufman and Rousseeuw 1990) uses medoids to divide a p-dimensional data set into different clusters using medoids. A medoid is considered as the center of a cluster, defined as the elements with the average distance between the medoid and other objects in this cluster is minimized. The algorithm attempts to search medoids and to partition other objects to their nearest medoid. In order to accelerate the computation, we apply random sampling and E-M algorithm to identify medoids and clusters. The algorithm is as follows:

Step 1. Calculate Mahalanobis distance matrix for 617 women:

$$\{D(i,j) = \sqrt{(\Phi_i - \Phi_j)^T \Omega^{-1} (\Phi_i - \Phi_j)} : i = 1, \ldots, 617, j = 1, \ldots, 617\}$$

where $\hat{\Phi}_i = E(\Phi_i|y)$ is the posterior mean of the eight summary characteristic measures of subject i. $\hat{\Omega} = E(\Omega|y)$ is the population level covariance matrix for these eight parameters.

Step 2. Sample K subjects randomly from 617 women as temporary medoids: $\{m_1, \ldots, m_K\}$

Step 3. E-step: assign subjects to closest clusters by minimizing total distance to cluster centers (medoids)

$$C(i) = \arg \min_{1 \le k \le K} D(i, m_k)$$

Step 4. M-step: for a given assignment C, find the center in the cluster minimizing total distance to other subjects in that cluster. This center is the new medoid

$$i_k^* = \arg \min_{\{i:C(i)=k\}} \sum_{C(i')=k} D(i, i')$$

Step 5. Iterate step 3 and 4 until the assignments do not change.

Step 6. Iterate from step 2 to step 5 until the combination of medoids do not change.

We use *silhouettes* (Kaufman and Rousseeuw 1990) to choose optimal number of clusters. For each subject in a given classification, we can calculate its *silhouette*. For any subject i assigned to cluster A, $a(i)$ is the average distance of subject i to all other subjects in A. For any cluster C other than A, we define $d(i, C)$ to be the average distance of subject i to all subjects in C. We then calculate $d(i, C)$ for all clusters $C \neq A$ and select the smallest of those: $b(i) = \min_{C \neq A} d(i, C)$, which is the average distance of subject i to all subjects of the closest neighbor cluster. For a given

number of clusters k, the average silhouette for all subjects is $\overline{s}(k) = \sum_{i=1}^{n} s(i)/n$, where $s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}}$. We then calculate the average silhouette for a variety of numbers of clusters k. The one with the highest average silhouette is an appropriate number of clusters for our data.

4.4.2 Classification of Women's Menstrual Patterns

Using the K-medoids algorithm described above, we classify women into six pattern categories based on the silhouette optimality criteria (see Table 4.8). The parameters of medoids of each category, *i.e.*, for subjects at the center of each cluster, are listed in Table 4.9. Figure 4.8 shows the observed segment lengths as well as posterior means and the upper and lower 2.5 percentiles of segment lengths distributions of the medoid subjects if FMP is observed, or subjects with an observed FMP that has the smallest distance to the corresponding medoid subject if FMP is censored for the medoid subjects. Figure 4.9 presents the category means (means of posterior measures of all subjects in a certain category) for posterior means and the upper and lower 2.5 percentiles of segment lengths distributions. Figure 4.10 describes how menstrual characteristics are distributed for each category, corresponding to the colors listed in Table 4.9. The number of subjects in each category as well as their proportions can also be found in Table 4.9.

Category 1 has 74 (12.0%) subjects and category 2 has 80 (13.0%) subjects. They represent subjects with early changepoints. The medoid variance changepoint is 40.5 years old and the medoid mean changepoint is at age 42.7. The biggest difference is that category 1 has a generally stable pattern before changepoints and a rapid increase in both mean and variability of segment lengths thereafter, while category 2 has generally stable variability before and after the changepoints. Examining Figure 4.8 and Figure 4.9, we can conclude that category 2 represents subjects with generally

larger variability overall but less obvious changepoints compared to other categories, thus less apparent transition and higher uncertainty of changepoints.

Category 3 has 187 (30.3%) subjects and category 4 has 76 (12.3%) subjects. They represent subjects with medium changepoint ages. For category 3, the medoid mean changepoint is at age 44.83 and the medoid variance changepoint is at age 41.45. For category 4, the medoid mean changepoint is at age 45.69 and the medoid variance changepoint is at age 39.84. Their biggest difference is that category 4 has larger variability at age 35 and decreases before variance changepoints, while category 3 has smaller variability at age 35 as well as a more stable pattern before variance changepoints. Category 4 also has an exceptionally large difference between late (mean) and early (variance) changepoints, which is nearly 6 years.

Category 5 has 87 (14.1%) subjects and category 6 has 113 (18.4%) subjects. They represent subjects with older changepoint ages. Their medoids mean changepoints are around age 48.5 while their medoids variance changepoints are around age 46.5. The biggest difference here is that category 5 have smaller variability at age 35 and a more stable pattern before variance changepoints, while category 6 has larger variability at age 35 and decreases before variance changepoints.

We can also observe from these results that subjects with early transitions tend to have early FMPs and subjects with late transitions tend to have late FMPs.

4.4.3 Menstrual Pattern Categories in Women's Late Reproductive Life and Women's Reproductive Characteristics

We compare three important aspects of women's reproductive lives by these six menstrual pattern categories: age at menopause, age at menarche and number of births. We conduct analysis of variance and pairwise comparisons to study differences between categories or groups of categories.

Table 4.10 suggests that age at FMP differs between categories. From the comparisons between single categories and comparisons between women with early (category 1 and 2), medium (category 3 and 4) and late changepoints (category 5 and 6), it's very apparent that women with older changepoints tend to have older age at FMP than women with younger changepoints. Women with similar changepoints generally do not differentiate in age at FMP, except that women with early changepoints but generally stable variability (category 2) have an FMP of 3.12 years later on average than women with apparent early transition (category 1).

Table 4.11 suggests that age at menarche also differs between categories. However, the difference is not significantly associated with changepoints. Pairwise comparisons show that women in category 4 and category 6 tend to have older age at menarche. These two categories share the same characteristics of having larger variability at age of 35. We further compare age of menarche between category 4 and 6 with other categories altogether and found that women with larger variability at age of 35 have menarche at 1.58 years older on average.

Both analysis of variance and pairwise comparisons suggest that there are no significant difference of number of births between categories.

4.4.4 Menstrual Categories and Transition Markers

Figure 4.3 through Figure 4.7 show how changepoints are associated with markers, as well as how categories of menstrual patterns are aligned with each markers. These figures suggest that for most subjects there are strong linear associations between changepoints and markers, while some subjects' changepoints have no clear relationship with markers. For early transition markers, most of these subjects are in categories 4 (cyan) and 6 (blue), as well as a few in category 2 (red). Category 4 and 6 share a similarity that they both have early larger variability. Thus their

transition markers may occur early while their actual changepoints may occur much later. Subjects in category 2 tend to have less apparent changepoints, *(*i.e.), less apparent transitions, thus it is reasonable that their transition markers may not be well aligned with their changepoints. For early markers, most subjects without clear linear relationship with changepoints are in categories categories 4 (cyan) 5 (magenta) and 6 (blue), indicating early markers are not quite linearly associated with late changepoints. Similar with late transition markers, early transition markers may occur much earlier than their actual changepoints for subjects with early large variability. These findings suggests that changepoints identified by Chapter III tend to capture menstrual patterns over longer life span, while transition markers concentrates more on local changes.

4.5 Discussion

In this paper, we use eight summary measures of individual woman's menstrual patterns defined by Chapter III to study associations between menstrual pattern and other features of women's late reproductive life. We find that women with later changepoints, smaller variance of segment lengths at age of 35, less rapid decrease in variance of segment lengths before changepoints, less rapid increase in mean and variance of segment lengths after changepoints as well as less abrupt changes of variance slopes before and after changepoints tend to have a later age at menopause.

We classify women's menstrual patterns into six subgroups by K-medoids algorithm. While previous classification by Gorrindo et al. (2007) was based on visual examination and simple statistics of 2-year sliding windows, our approach is more deeply rooted in statistical modeling and features of menstrual patterns during women's late reproductive life. The two classifications also differ in the time

frame of study. While classification by Gorrindo et al. (2007) was based on women's lifetime data, our classification only use data after 35 years old. Thus our classification is based purely on women's late reproductive life. The five types given by Gorrindo et al. (2007) depended mostly on variability measures and on the trend of cycle lengths changes overtime, which were based on their belief of key features before conducting quantitative analysis. Our findings suggest that time of transitions (changepoints) is the most important features with respect to classification, while features such as early mean and variance of segment lengths, rate of changes before and after changepoints also play a role. We further compare age of menopause (FMP), age at menarche and number of births among different groups. While Gorrindo et al. (2007) found no significant associations between their categories and age of menopause, we find age of menopause is strongly associated with menstrual transition time (changepoints) and women with older transition age tend to have older age of menopause. While Gorrindo et al. (2007) suggested that women with stable menstrual patterns tend to have later menarche than women with erratic patterns, our findings suggest that the key factor here is early variability (variance of segment lengths at age of 35), with women with larger early variability tending to have later age at menarche. Gorrindo et al. (2007) found women with stable menstrual patterns had fewer births than women with erratic pattern, but our findings suggest no significant association between number of births and menstrual patterns.

To assess whether changepoints can serve as transition markers, we study their associations with previously defined transition markers D60, D90, RR10, Irregularity and Diff6p. Results suggest most mean and variance changepoints are well aligned with transition markers, while some are less clearly associated. The lack of association is due to time frame considered. Transition markers are calculated by

sliding windows, which are local time frames. From the classification of subjects by menstrual cycle patterns, we can see that women with highly variable cycles and/or weakly defined transitions can "trigger" transition markers long before the true transition begin. Thus mean and variance changepoints, which are identified using data throughout women's late reproductive life, provide more comprehensive information about menopausal transition than previously defined transition markers.

	Estimated Effects (95%CI)	
	Model 1	Model 2
Mean intercept	**1.12(0.83,1.41)**	**1.12(0.83,1.41)**
Mean slope before changepoint	-0.34 (-0.67,0.00)	-0.33 (-0.79,0.13)
Mean slope after changepoint	**-0.99 (-1.33,-0.65)**	**-0.99 (-1.32,-0.66)**
Log-variance intercept	**0.34 (0.01,0.66)**	**0.33 (0.01,0.66)**
Log-variance slope before changepoint	**-0.61 (-0.96,-0.26)**	**-0.61 (-0.98,-0.23)**
Log-variance slope after changepoint	0.25 (-0.13,0.63)	0.25 (-0.15,0.65)
Segment length mean changepoint	**2.52 (2.35,2.70)**	**2.47 (2.26,2.68)**
Segment length variance changepoint	**1.95 (1.70,2.19)**	**1.93 (1.68,2.18)**

Table 4.1: Influence of each menstrual pattern characteristics on age of final menstruation periods: influences are assessed one character per model(not jointly); Model 1 is a Gaussian AFT model with identity link which included censored FMPs; Model 2 is semiparametric AFT model which included censored FMPs. Bold numbers represent significant associations.

	Estimated Effects (95%CI)	
	Model 1	Model 2
Mean intercept	0.08(-0.09,0.25)	0.07(-0.10,0.24)
Mean slope before changepoint	0.01(-0.15,0.18)	0.03(-0.20,0.25)
Mean slope after changepoint	**-0.60(-0.79,-0.42)**	**-0.64(-0.85,-0.43)**
Log-variance intercept	**-0.53(-0.84,-0.22)**	**-0.54(-0.88,-0.20)**
Log-variance slope before changepoint	**-2.01(-2.75,-1.28)**	**-2.08(-2.99,-1.18)**
Log-variance slope after changepoint	**-2.06(-2.63,-1.50)**	**-2.14(-2.80,-1.48)**
Segment length mean changepoint	**1.80(1.24,2.35)**	**1.75(1.11,2.38)**
Segment length variance changepoint	**1.32(0.67,1.98)**	**1.37(0.65,2.10)**

Table 4.2: Influence of each menstrual pattern characteristics on age of final menstruation periods: influences are assessed jointly for all characteristics; Model 1 is a Gaussian AFT model with identity link which included censored FMPs; Model 2 is semiparametric AFT model which included censored FMPs. Bold numbers represent significant associations.

	Estimated Effects (95%CI)	
	Model 1	Model 2
Mean intercept	-	-
Mean slope before changepoint	0.05(-0.10,0.19)	0.06(-0.13,0.24)
Mean slope after changepoint	**-0.52(-0.70,-0.35)**	**-0.53(-0.71,-0.35)**
Log-variance intercept	**-0.44(-0.74,-0.14)**	**-0.44(-0.76,-0.12)**
Log-variance slope before changepoint	**-1.94(-2.64,-1.24)**	**-1.94(-2.77,-1.11)**
Log-variance slope after changepoint	**-2.08(-2.61,-1.54)**	**-2.09(-2.69,-1.48)**
Segment length mean changepoint	**1.95(1.42,2.48)**	**1.94(1.34,2.55)**
Segment length variance changepoint	**1.19(0.56,1.81)**	**1.18(0.50,1.87)**
Mean Slope (before*after) Changepoint	**-0.36(-0.49,-0.22)**	**-0.37(-0.58,-0.17)**
Log-var Slope (before*after) Changepoints	-	-

Table 4.3: Influence of each menstrual pattern characteristics on age of final menstruation periods: influences are assessed jointly after backward selection for all characteristics as well as before and after changepoints interactions; Model 1 is a Gaussian AFT model with identity link which included censored FMPs; Model 2 is semiparametric AFT model which included censored FMPs. All covariates are standardized. Bold numbers represent significant associations.

	Estimated Effects (95%CI)	
	Model 1	Model 2
Mean intercept	**0.22(0.06,0.38)**	**0.21(0.06,0.36)**
Log-variance intercept	**-1.36(-1.59,-1.13)**	**-1.37(-1.60,-1.13)**
Log-variance slope before changepoint	**-4.44(-4.81,-4.07)**	**-4.50(-4.93,-4.07)**
Log-variance slope after changepoint	**-3.90(-4.25,-3.55)**	**-3.99(-4.39,-3.59)**
Segment length variance changepoint	**3.36(3.16,3.57)**	**3.37(3.17,3.58)**

Table 4.4: Influence of each menstrual pattern characteristics on age of final menstruation periods: influences are assessed jointly for early characteristics; Model 1 is a Gaussian AFT model with identity link which included censored FMPs; Model 2 is semiparametric AFT model which included censored FMPs. All covariates are standardized. Bold numbers represent significant associations.

	Model A	Model B	Model C
Adjusted R^2	0.825	0.828	0.812
MSE	1.72	1.68	1.90

Table 4.5: Predictive ability of menstrual pattern characteristics for FMP: model A is the accelerated failure time model adjusted for all eight characteristics; model B is the accelerated failure time model adjusted for all eight characteristics and interactions after backward selection; model C is the accelerated model adjusted for "early" characterstics, which includes cycle length at age of 35 and all variability characteristics.

Makers	Stages	Mean (SE)	Number of Observations
D60	Late	46.96 (4.25)	443
D90	Late	48.80 (3.45)	359
RR10	Late	47.18 (3.96)	440
Irregularity	Early	44.76 (4.51)	502
Diff6p	Early	40.98 (4.24)	572

Table 4.6: Characteristics of transition markers.

	D60	D90	RR10	Irregularity	Diff6p
Mean changepoints	0.61	0.86	0.69	0.48	0.28
Variance changepoints	0.47	0.68	0.55	0.39	0.17

Table 4.7: Correlations of mean and variance changepoints and menstrual transition markers.

Number of Clusters (k)	Silhouette
2	0.1431
3	0.1404
4	0.1265
5	0.1304
6	0.1434
7	0.1363
8	0.1351
9	0.1182
10	0.1287

Table 4.8: Average silhouette $\overline{s}(k)$ for clustering with different number of clusters.

	Medoids					
Parameters	1	2	3	4	5	6
Mean Intercept	3.24	3.30	3.29	3.31	3.30	3.43
Mean slope before changepoint	-0.005	0.008	-0.006	-0.007	-0.004	-0.011
Mean slope after changepoint	0.21	0.09	0.12	0.16	0.15	0.17
Segment length mean changepoint	42.76	42.75	44.83	45.69	48.25	48.94
Log-variance intercept	-4.93	-5.34	-5.37	-3.26	-5.64	-3.79
Log-variance slope before changepoint	0.12	0.39	0.07	-0.38	0.15	-0.08
Log-variance slope after changepoint	0.75	-0.25	0.51	0.83	0.62	0.89
Segment length variance changepoint	40.89	40.40	41.45	39.84	46.45	46.73
Number of subjects in cluster(Proportion)	74(12.0%)	80(13.0%)	187(30.3%)	76(12.3%)	87(14.1%)	113(18.3%)
Cluster means of age at menarche (SE)	11.87 (1.15)	11.70 (1.25)	11.78 (1.25)	12.29 (1.34)	11.91 (1.29)	12.15 (1.28)
Cluster means of number of births (SE)	2.24 (1.80)	2.54 (1.80)	2.64 (1.86)	2.45 (1.76)	2.41 (1.92)	2.60 (1.81)
Cluster means of age at FMP* (SE)	47.05 (2.04)	50.17 (2.28)	50.44 (1.72)	50.62 (2.40)	52.63 (1.84)	52.69 (2.08)
Color in figure.	Green	Red	Black	Cyan	Magenta	Blue

*Including observed and imputed age of FMP.

Table 4.9: Parameters of cluster medoids. Also shown are the number of subjects (and proportion to total number of subjects), age at menarche, number of births and age at FMP by cluster, together with the corresponding color in figures.

Pattern category	1	2	3	4	5	6	1 and 2	3 and 4
1	-							
2	**3.12**	-						
3	**3.38**	0.26	-					
4	**3.57**	0.45	0.19	-				
5	**5.57**	**2.46**	**2.19**	**2.01**	-			
6	**5.64**	**2.51**	**2.26**	**2.07**	0.06	-		
3 and 4							**1.92**	-
5 and 6							**4.05**	**2.13**

Overall analysis of variance: $F_{5,611} = 89.05, P <= 0.001$

Table 4.10: Pairwise differences in age at FMP by menstrual pattern category: entry in row i and column j is (mean age at FMP in group i)-(mean age at FMP in group j); bold numbers represent significant differences. Category 1 and 2 are women with early changepoints; category 3 and 4 are women with medium changepoints; category 5 and 6 are women with late changepoints.

Pattern category	1	2	3	4	5	6	1 and 2	3 and 4	1,2,3,5
1	-								
2	-0.17	-							
3	-0.10	0.08	-						
4	**0.41**	**0.59**	**0.51**	-					
5	0.03	0.21	0.13	**-0.38**	-				
6	0.27	**0.45**	**0.37**	-0.14	0.24	-			
3 and 4							0.24	-	
5 and 6							0.24	-0.00	
4 and 6									**1.58**

Overall analysis of variance: $F_{5,611} = 2.87, P <= 0.014$

Table 4.11: Pairwise differences in age at menarche by menstrual pattern category: entry in row i and column j is (mean age at menarche in group i)-(mean age of menarche in group j); bold numbers represent significant differences. Category 1 and 2 are women with early changepoints; category 3 and 4 are women with medium changepoints; category 5 and 6 are women with late changepoints; category 4 and 6 are women with larger variability at age of 35; category 1,2,3,and 5 are women with smaller variability at age of 35.

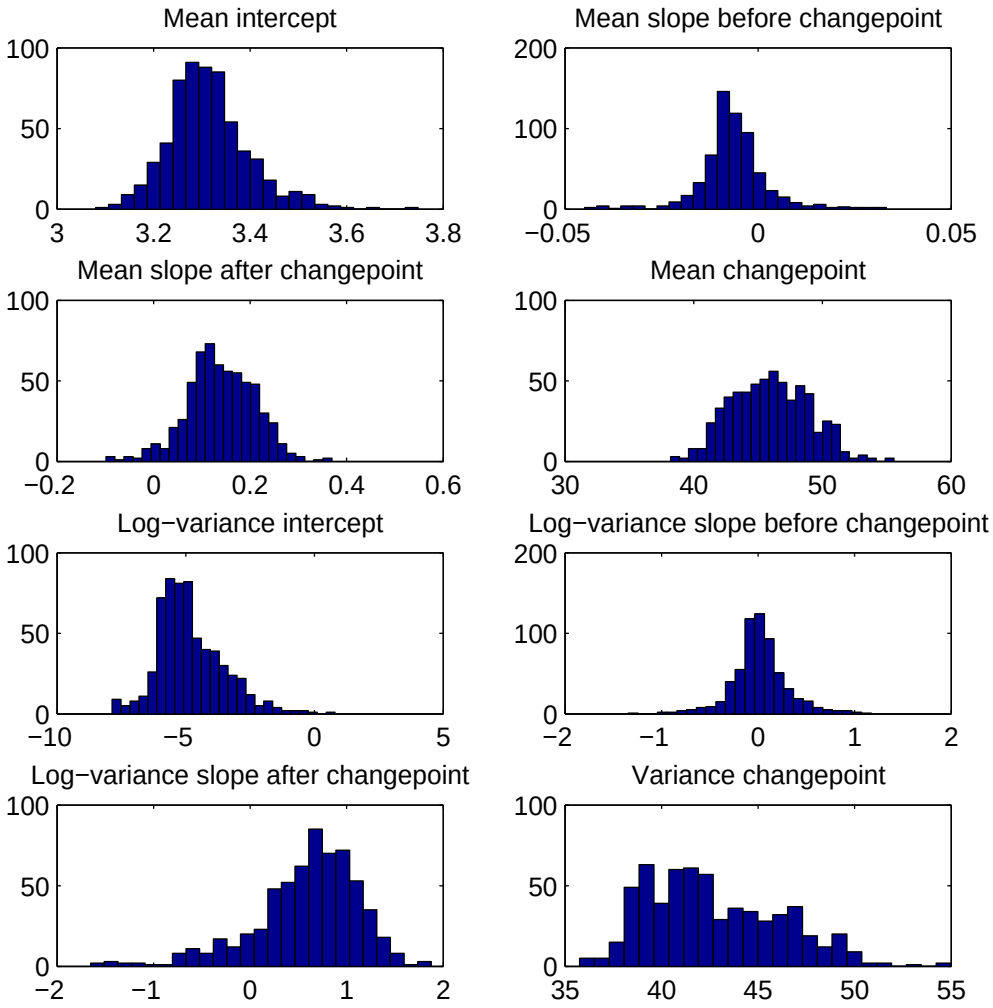

Figure 4.1: Histogram of parameters for all subjects.

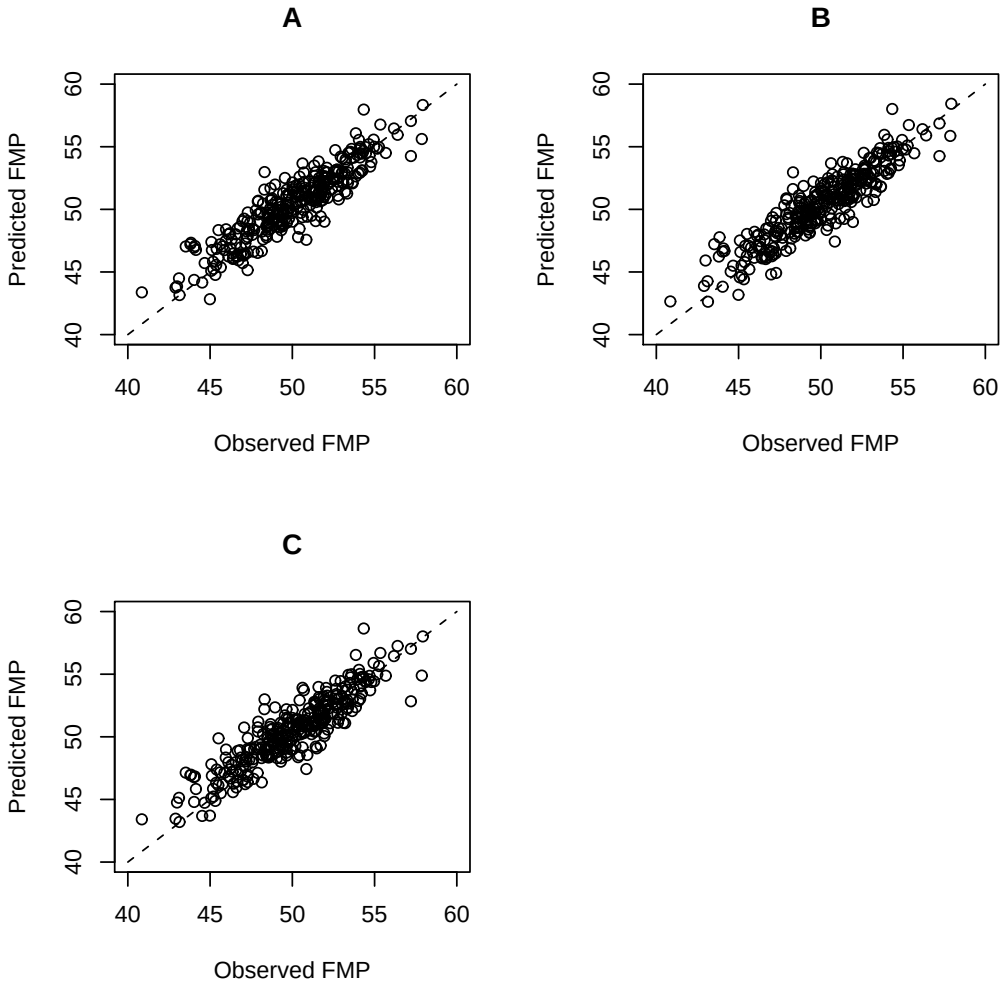

Figure 4.2: Predicted FMP and observed FMP: model A is the accelerated failure time model adjusted for all eight characteristics; model B is the accelerated failure time model adjusted for all eight characteristics and interactions after backward selection; model C is the accelerated model adjusted for "early" characterstics, which includes cycle length at age of 35 and all variability characteristics. Predicted FMP are calculated using leave-one-out cross-validation.

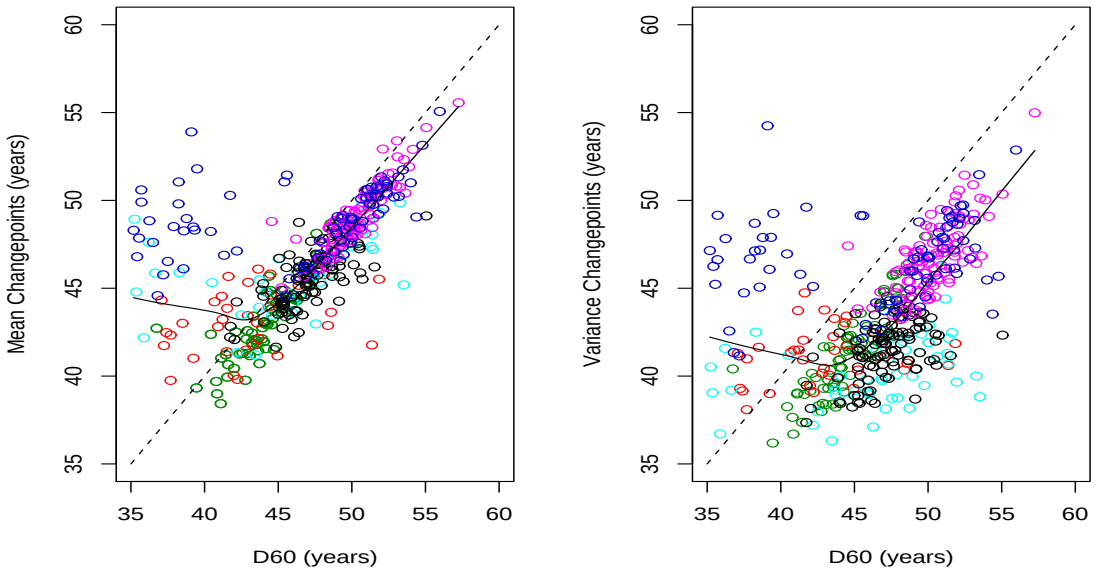

Figure 4.3: D60 and changepoint: solid line is the loess regression line and dotted line is the diagonal line. Corresponding medoids can be found in Table 4.9 by color names.

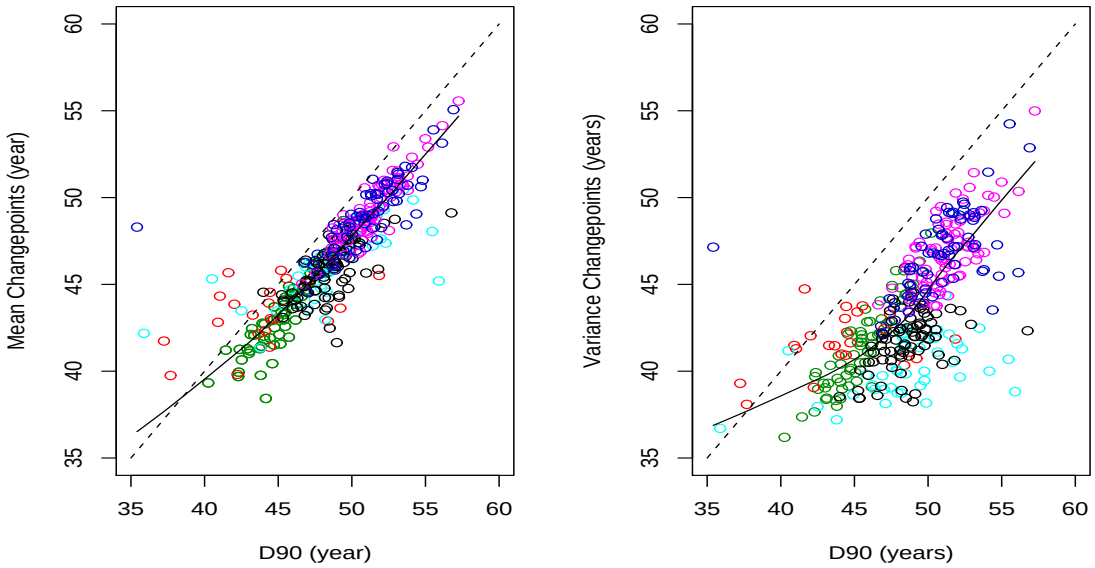

Figure 4.4: D90 and changepoint: solid line is the loess regression line and dotted line is the diagonal line. Corresponding medoids can be found in Table 4.9 by color names.

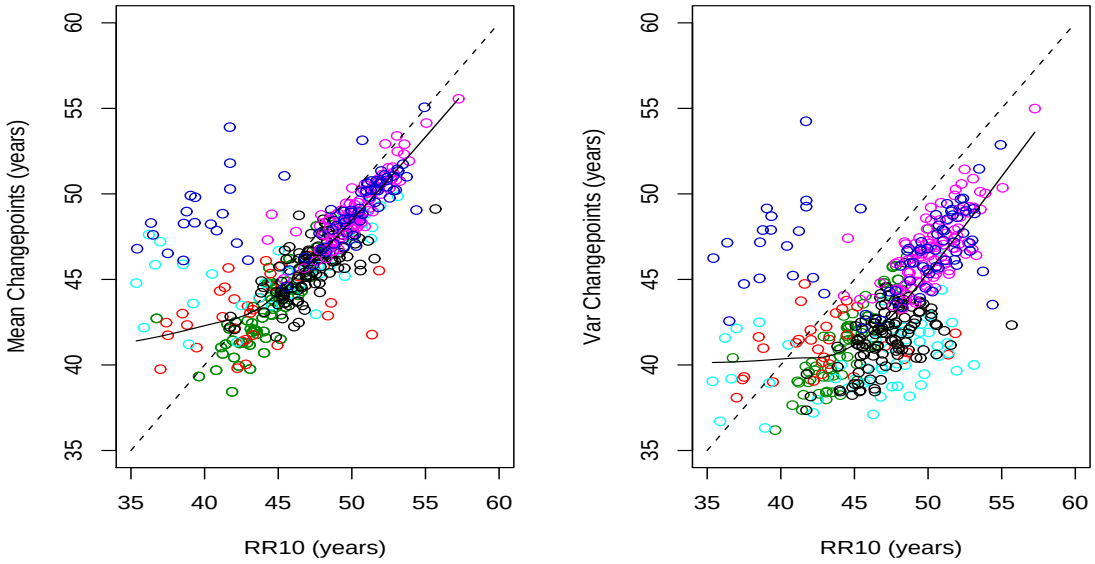

Figure 4.5: RR10 and changepoint: solid line is the loess regression line and dotted line is the diagonal line. Corresponding medoids can be found in Table 4.9 by color names.

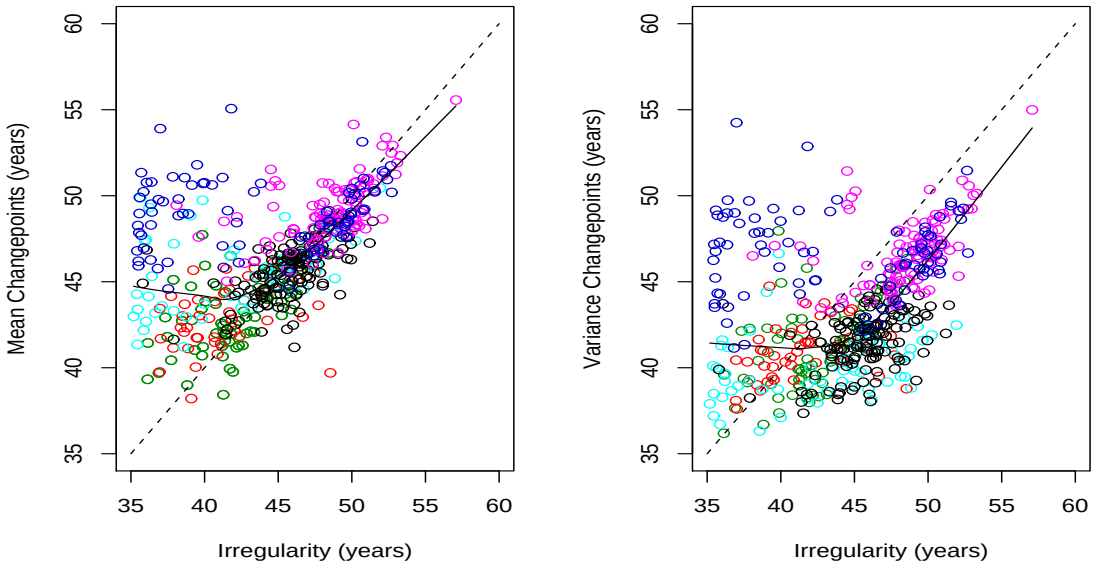

Figure 4.6: Irregularity and changepoint: solid line is the loess regression line and dotted line is the diagonal line. Corresponding medoids can be found in Table 4.9 by color names.

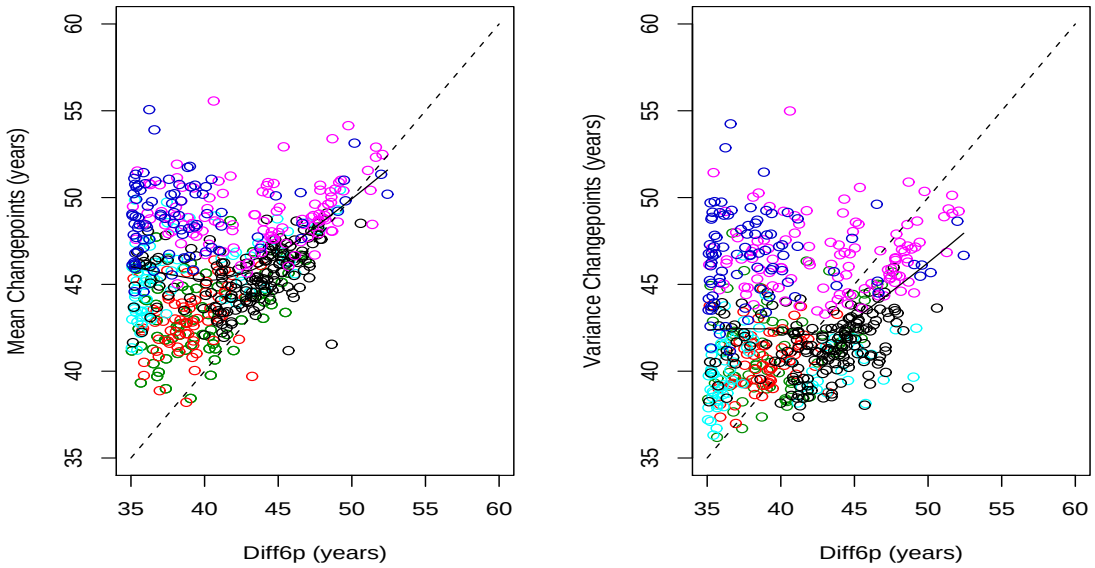

Figure 4.7: Diff6p and changepoint: solid line is the loess regression line and dotted line is the diagonal line. Corresponding medoids can be found in Table 4.9 by color names.

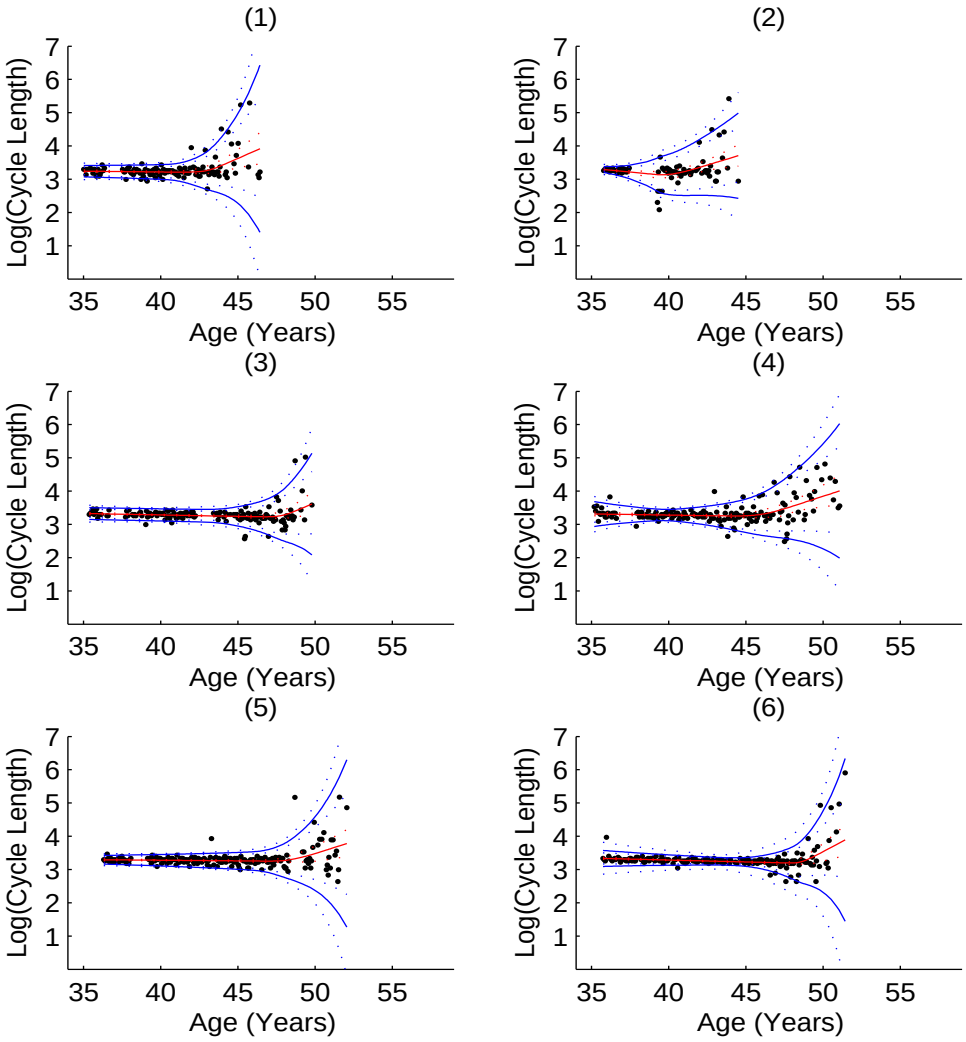

Figure 4.8: Profiles of medoids. Red lines represent of posterior mean of the mean segment length and associated 95% credible intervals; blue lines represent posterior mean for the upper and lower 2.5 percentiles for the segment distribution and their associated 95% credible intervals. Panel titles represent category IDs.

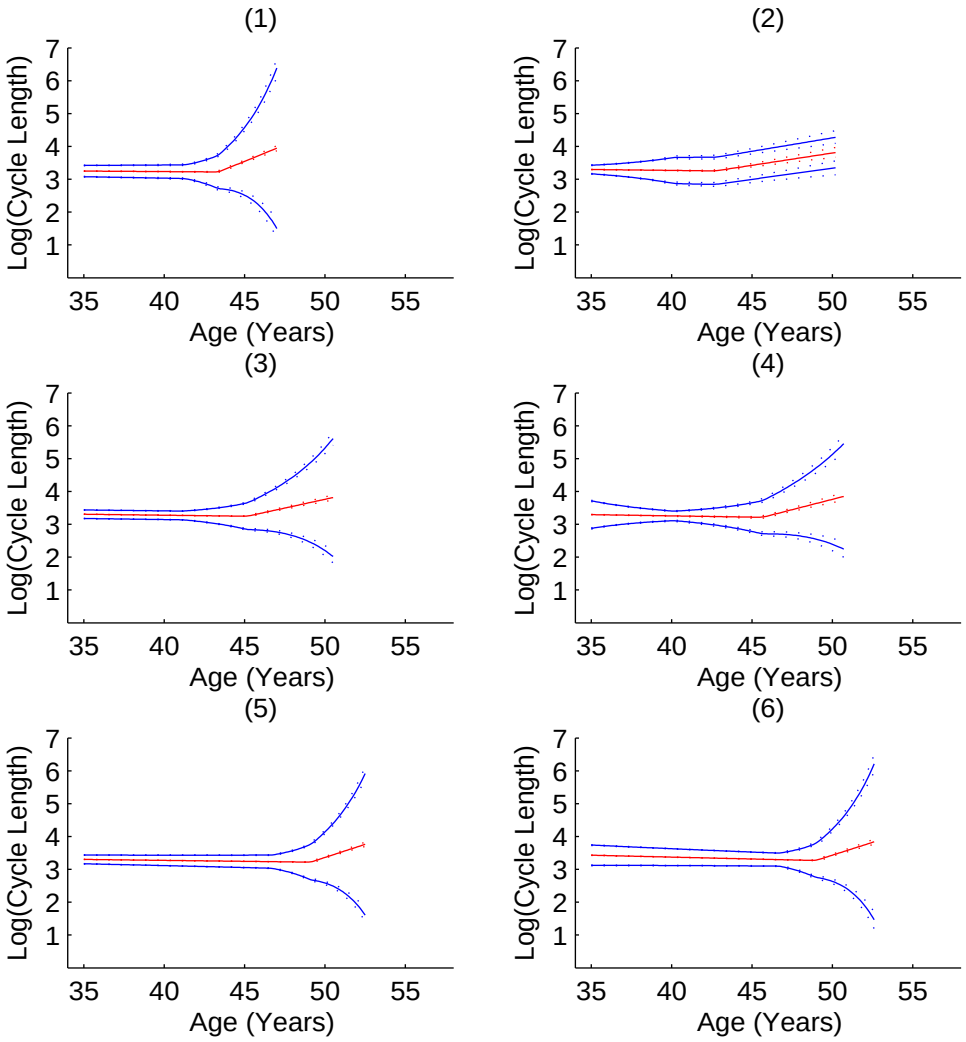

Figure 4.9: Cluster average pattern profiles. Red lines represent cluster mean of posterior mean of the mean segment length and associated 95% credible intervals; blue lines represent cluster mean of posterior mean for the upper and lower 2.5 percentiles for the segment distribution and their associated 95% credible intervals. Panel titles represent category IDs.

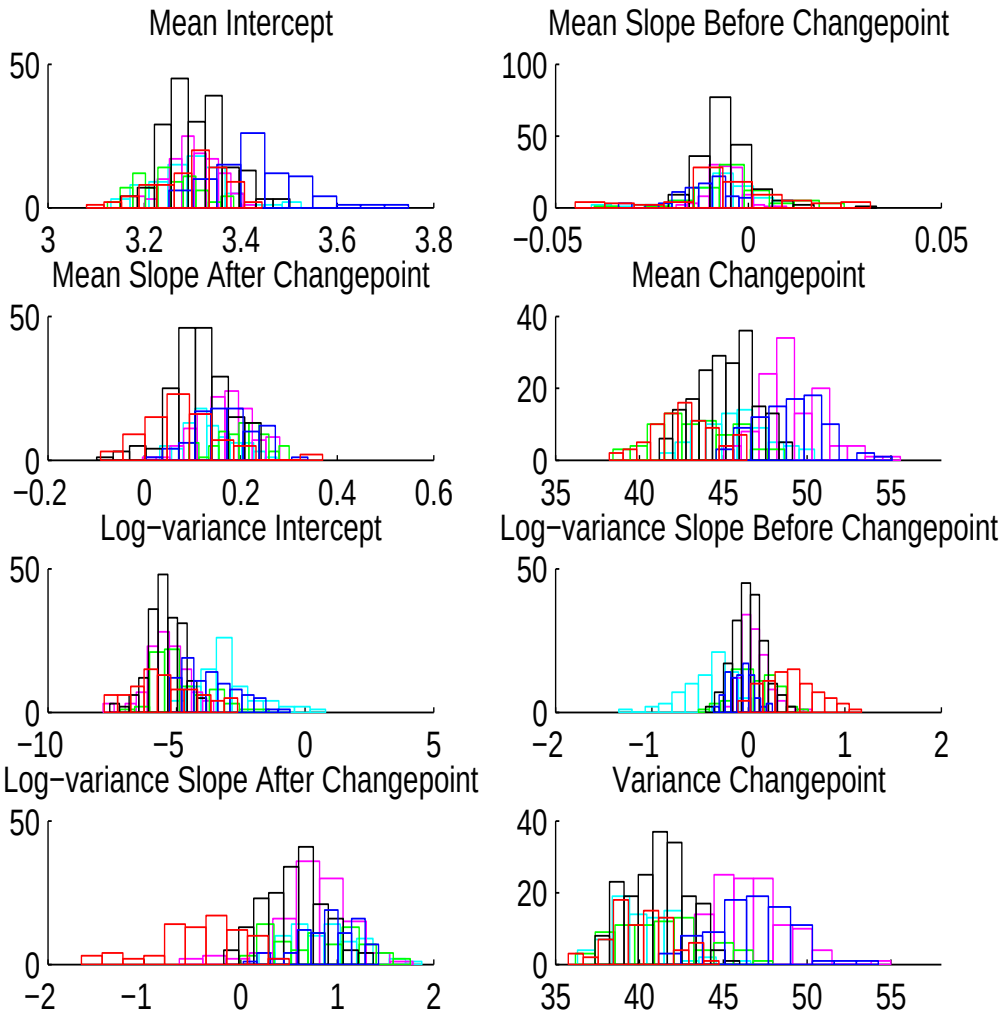

Figure 4.10: Histogram of parameters by clusters. Corresponding medoids can be found in Table 4.9 by color names.

CHAPTER V

Conclusion and Discussion

In this dissertation, I proposed statistical modeling in two different topics: weight smoothing model for complex designed surveys and modeling women's menstruation history data.

5.1 Weight smoothing model for complex designed surveys

In sample surveys, highly disproportional sample designs may have large weights or large variability of weights, which will lead to large variability in statistical estimates of population quantities. *Weight trimming*, the most commonly used design based approach, reducing variability of estimates by introducing some bias, however has limitation that it does not use data to optimize bias-variance tradeoffs. Elliott and Little (2000) and Elliott (2007) introduced model based method *weight smoothing* to reduce variance of statistical estimates by allowing outlying data elements of large weights to borrow strength from the rest of data. In Chapter II, I proposed weight smoothing models for more complex sample design that include single or multi-stage cluster samples or strata that "cross" the weight strata. Simulation suggests that model based approaches have better bias-variance property than designed based approaches in general. XSE estimators have best property when there is no correlation between outcomes and weights, especially when cluster effect is small.

When outcome and weights are associated, XSE estimators' performance is not as good as fully-weighted estimators. LSE and SLSE estimators are more robust and have slightly reduction in RMSE compared to fully weighted estimators generally.

Similar techniques can be used in other situations wherever highly variable weights present. For example, in open-label extension study of randomized controlled clinical trials, only a subset of patients may choose to participate in the extension phase (Wainwright 2002, Taylor and Wainwright 2005). Such choices are not made at random, thus a complete analysis would suffer from selection bias, which could be adjusted by inclusion probability (weights). Our model can be used to reduce MSE when highly variable weights present.

Elliott (2008, 2009) further developed *model averaging* methods to average estimators from a set of weight pooling models with all possible weight trimming levels. This approach involves variable selection procedures and has high robustness. Future exploration on this topic could extend *model averaging* method to surveys with cluster design. The challenge here is that the number of weight strata in complex designed surveys are usually large, which will rapidly increase the number of models to be averaged. Thus, future research should develop a variable selection algorithm to limit the number of models under consideration to the subset large enough to contain most appropriate models but small enough to allow averaging in a reasonable time frame.

5.2 Women's menstruation history study

Women's menstruation history provide rich information for women's reproductive health. TREMIN, the ongoing 70-year longitudinal study initiated by Dr. Alan Trelore in 1967 provide a valuable data resource to study the nature of women's

menstruation and reproductive health. However, the biggest challenge for previous researches is lack of precision. The statistical question here is how to comprehensively model women's menstruation pattern. The main contribution of Chapter III is to provide a more precise method to capture women's menstruation patterns. Chapter IV uses results of Chapter III to quantitatively study women's menstruation more thoroughly. In Chapter III, we develop a Bayesian changepoint model that uses separate linear splines with unknown changepoints for individual women's mean cycle lengths and variability of cycle lengths, which are then tied together using a hierarchical model to smooth parameter estimates for women with limited cycle length data, and to provide population-level estimates. We also developed an algorithm to impute various types of missingness, including imputation of final menstruation periods. Results suggest the model captured well each individual woman's menstruation pattern, especially the mean and variability changepoints of cycle lengths that can serve as late and early menopausal transition markers. The model also provides a good summary of population level menstruation characteristics and their associations.

The Bayesian changepoint model uses eight characteristics to describe individual woman's menstruation pattern. Quantitative estimates of these characteristics make it possible to study the association of women's menstruation patterns and their age at menopause; to define subgroups of women with similar pattern characteristics and their associations with age at menopause, age at menarche and parity; and to compare mean and variability changepoints with previously defined transition markers. We address these issues in Chapter IV. Gaussian accelerated failure time models suggest women with late menopause tend to have later changepoints, smaller variance of cycle lengths at age of 35, less rapid decrease in variance of cycle lengths

before changepoints, less rapid increase in mean and variance of cycle lengths after changepoints as well as less multiplicative effect of changes of variance slopes before and after changepoints. We next used a K-medoids algorithm to classify women's menstruation patterns to six categories. Category 1 has early changepoints and a rapid increase in both mean and variability of cycle lengths after changepoints; category 2 also has early changepoints, but the changes are not quite apparent compared to other categories. Category 3 and category 4 have medium changepoints, but category 4 has larger variability at age 35 with more rapid decrease in variability than category 3. Category 5 and category 6 have late changepoints, but category 6 has larger variability at age 35 with more rapid decrease in variability than category 5. We also compared several health outcomes by different categories. Results suggest that age at menopause is significantly positively associated with changepoints; age at menarche is primarily influenced by early variability of menstruation cycle lengths, with subjects with larger early variability have later age at menarche; and number of births does not differ between different categories. The comparison of changepoints and previously defined transition markers suggests changepoints are better associated with late transition markers. Subjects with larger early variability tend to have early transition markers but not necessarily early changepoints, which means changepoints are more comprehensive and better measures to capture menopausal transition time.

The next step of our research is to include TREMIN corhort II data and compare menstruation patterns between different generation of women, as well as adjusting for more covariates such as BMI and smoking.

Methodologically, since we need to impute final menstruation periods, a possible future extension to this work is to combine the menstruation pattern information

with the age of FMP in a joint modeling approach. This approach will use pattern characteristics to impute final menstruation periods. However, it could be more complex and computationally intensive.

APPENDICES

APPENDIX A

Gibbs Sampling Algorithm for Chapter III

Gibbs sampling is used to draw from the posterior distribution $p(\mathbf{\Phi}, \Lambda, \Omega|\boldsymbol{z})$, where $\Phi_i = (\alpha_i^\mu, \beta_i^\mu, \gamma_i^\mu, \theta_i^\mu, \alpha_i^\sigma, \beta_i^\sigma, \gamma_i^\sigma, \theta_i^\sigma)'$. The algorithm outline is as follows:

1. Initialize $\mathbf{\Phi}, \Lambda, \Omega$. Perform an initial imputation of missing data.

2. For $i = 1, ..., n$ and z_i consisting of both observed and imputed data:

2a.

$$(\alpha_i^\mu, \beta_i^\mu, \gamma_i^\mu|rest) \sim N(U, V)$$

$$U = (A_i^{\mu'} W_i^{-1} A_i^\mu + \Omega_\mu^{-1})^{-1}(A_i^{\mu'} W_i^{-1} z_i + \Omega_\mu^{-1} x_i' \Lambda_\mu), V = (A_i^{\mu'} W_i^{-1} A_i^\mu + \Omega_\mu^{-1})^{-1}$$

where $W_i = Diag(\sigma_{it}^2)$, $A_i^\mu = \begin{pmatrix} 1 & (a_{i1} - 35) & (a_{i1} - \theta_i^\mu)_+ \\ \vdots & \vdots & \vdots \\ 1 & (a_{iT_i} - 35) & (a_{iT_i} - \theta_i^\mu)_+ \end{pmatrix}$, and Λ_μ and Ω_μ are the corresponding part of prior multivariate normal mean Λ and covariance matrix Ω conditional on other parameters.

2b.

$$p(\alpha_i^\sigma|rest) \propto exp(-\frac{1}{2}(\sum_{t=1}^{T_i} \frac{z_{it} - (\alpha_i^\mu + \beta_i^\mu(a_{it} - 35) + \gamma_i^\mu(a_{it} - \theta_i^\mu)_+)^2}{exp(\alpha_i^\sigma + \beta_i^\sigma(a_{it} - 35) + \gamma_i^\sigma(a_{it} - \theta_i^\sigma)_+)} + \frac{(\alpha_i^\sigma - \mu_{\alpha^\sigma})^2}{\Omega_{\alpha^\sigma}}))$$
$$\times exp(-\frac{\alpha_i^\sigma T_i}{2})$$

where $\mu_{\alpha^\sigma} = x_i'\Lambda_{\alpha^\sigma}$ and Ω_{α^σ} are the corresponding part of prior multivariate normal mean and variance conditional on other parameters. The inverse CDF method is used to obtain the conditional draws.

2c.

$$p(\beta_i^\sigma|rest) \propto exp(-\frac{1}{2}(\sum_{t=1}^{T_i} \frac{z_{it} - (\alpha_i^\mu + \beta_i^\mu(a_{it} - 35) + \gamma_i^\mu(a_{it} - \theta_i^\mu)_+)^2}{exp(\alpha_i^\sigma + \beta_i^\sigma(a_{it} - 35) + \gamma_i^\sigma(a_{it} - \theta_i^\sigma)_+)} + \frac{(\beta_i^\sigma - \mu_{\beta^\sigma})^2}{\Omega_{\beta^\sigma}}))$$
$$\times exp(-\frac{1}{2}\beta_i^\sigma \sum_{t=1}^{T_i}(a_{it} - 35))$$

where $\mu_{\beta^\sigma} = x_i'\Lambda_{\beta^\sigma}$ and Ω_{β^σ} are the corresponding part of prior multivariate normal mean and variance conditional on other parameters. The inverse CDF method is used to obtain the conditional draws.

2d.

$$p(\gamma_i^\sigma|rest) \propto exp(-\frac{1}{2}(\sum_{t=1}^{T_i} \frac{z_{it} - (\alpha_i^\mu + \beta_i^\mu(a_{it} - 35) + \gamma_i^\mu(a_{it} - \theta_i^\mu)_+)^2}{exp(\alpha_i^\sigma + \beta_i^\sigma(a_{it} - 35) + \gamma_i^\sigma(a_{it} - \theta_i^\sigma)_+)} + \frac{(\gamma_i^\sigma - \mu_{\gamma^\sigma})^2}{\Omega_{\gamma^\sigma}}))$$
$$\times exp(-\frac{1}{2}\gamma_i^\sigma \sum_{t=1}^{T_i}(a_{it} - \theta_i^\sigma)_+)$$

where $\mu_{\gamma^\sigma} = x_i'\Lambda_{\gamma^\sigma}$ and Ω_{γ^σ} are the corresponding part of prior multivariate normal mean and variance conditional on other parameters. The inverse CDF method is used to obtain the conditional draws.

2e.

$$p(\theta_i^\mu|rest) \propto exp(-\frac{1}{2}(\sum_{t=1}^{T_i} \frac{z_{it} - (\alpha_i^\mu + \beta_i^\mu(a_{it} - 35) + \gamma_i^\mu(a_{it} - \theta_i^\mu)_+)^2}{exp(\alpha_i^\sigma + \beta_i^\sigma(a_{it} - 35) + \gamma_i^\sigma(a_{it} - \theta_i^\sigma)_+)} + \frac{(\theta_i^\mu - \mu_{\theta^\mu})^2}{\Omega_{\theta^\mu}}))$$

where $\mu_{\theta^\mu} = x_i'\Lambda_{\theta^\mu}$ and Ω_{θ^μ} are the corresponding part of prior multivariate normal mean and variance conditional on other parameters. The inverse CDF method is used to obtain the conditional draws.

2f.

$$p(\theta_i^\sigma|rest) \propto exp(-\frac{1}{2}(\sum_{t=1}^{T_i} \frac{z_{it} - (\alpha_i^\mu + \beta_i^\mu(a_{it} - 35) + \gamma_i^\mu(a_{it} - \theta_i^\mu)_+)^2}{exp(\alpha_i^\sigma + \beta_i^\sigma(a_{it} - 35) + \gamma_i^\sigma(a_{it} - \theta_i^\sigma)_+)} + \frac{(\theta_i^\sigma - \mu_{\theta^\sigma})^2}{\Omega_{\theta^\sigma}}))$$
$$\times exp(-\frac{1}{2}\gamma_i^\sigma \sum_{t=1}^{T_i}(a_{it} - \theta_i^\sigma)_+)$$

where $\mu_{\theta^\sigma} = x_i'\Lambda_{\theta^\sigma}$ and Ω_{θ^σ} are the corresponding part of prior multivariate normal mean and variance conditional on other parameters. The inverse CDF method is used to obtain the conditional draws.

3.

$$\Lambda|rest \sim N((X'(\Omega \otimes I_N)^{-1}X)^{-1}X'(\Omega \otimes I_N)^{-1}\mathbf{\Phi}, (X'(\Omega \otimes I_N)^{-1}X)^{-1})$$

where X is the covariate matrix of all subjects, which consists of stacked rows of x_i', and $\mathbf{\Phi}$ consists of the stacked rows of Φ_i'.

4.

$$\Omega|rest \sim Inv - Wishart(\Omega|(\sum_{i=1}^{N}(\Phi_i - x_i'\Lambda)(\Phi_i - x_i'\Lambda)' + I))$$

5. Use the updated parameters to create a new imputation data set. Then go to step 2.

APPENDIX B

Piecewise Exponential Distribution for Chapter III

Assume that Q_i, the age at FMP, follows a piecewise exponential distribution. The baseline hazard is constant within each interval, so that

$$\lambda_0(t) = \eta_k, t \in [A_{k-1}, A_k]$$
$$f(Q_i = t : t \in [A_{k-1}, A_k]) = \eta_k e^{-\eta_k t}$$

Here, $A_0, ...A_K$ are a set of age knots, which are set at age 40, 42, 43, 44, 45, 46, 46.5, 47, 47.5, 48, 48.5, 49, 49.5, 50, 51.5, 52, 52.5, 53, 53.5, 54, 55, 56, and 57; we define $A_{-1} = 0$ and $A_{K+1} = \infty$ and assume $\eta_0 = 0$ (no risk of FMP before age 40).

We postulate a very weakly informative prior for $\eta_k : \eta_k \sim \text{GAMMA}(0.001, 0.001)$. The posterior distribution for η_k is:

$$p(\eta_k|\tilde{q}) \propto p(\tilde{q}|\eta_k)p(\eta_k) \propto \text{GAMMA}(m_k + 0.001, r_k + 0.001)$$

where $m_k = \sum_i I(A_{k-1} \leq Q_i \leq A_k)$ is the number of women with FMPs that occur between time A_{k-1} and A_k and $r_k = \sum_i I(Q_i \geq A_{k-1})$ is the number of women without an FMP at time A_{k-1}.

The hazard function for each interval is

$$\lambda(t) = \eta_k I(A_{k-1} \leq t \leq A_k)$$

The cumulative hazard and survival functions are then given by

$$\Lambda(t) = \int_0^t \lambda(t)dt = \sum_{j=1}^{k-1} \eta_j(A_j - A_{j-1}) + \eta_k(t - A_{k-1}), t \in [A_{k-1}, A_k]$$

$$S(t) = exp(-\Lambda(t))$$

The probability that the event occurs in the interval $[t_1, t_2]$ given that the event has not occured by t_1 is

$$P(Q_i \in [t_1, t_2] | Q_i > t_1) = \frac{S(t_1) - S(t_2)}{S(t_1)} = 1 - \frac{S(t_2)}{S(t_1)} =$$

$$\begin{cases} 1 - e^{-\eta_k(t_2 - t_1)} \text{ if } A_{k-1} \leq t_1 < t_2 \leq A_k \\ 1 - e^{-[\eta_k(t_2 - A_{k-1}) - \eta_{k-1}(t_1 - A_{k-1})]} \text{ if } A_{k-2} \leq t_1 < A_{k-1} \leq t_2 \leq A_k \\ 1 - e^{-[\eta_{k+1}(t_2 - A_k) + \eta_k(A_k - A_{k-1}) - \eta_{k-1}(t_1 - A_{k-1})]} \text{ if } A_{k-2} \leq t_1 < A_{k-1} < A_k \leq t_2 \end{cases}$$

BIBLIOGRAPHY

BIBLIOGRAPHY

Avis, N., Assmann, S., Kravitz, H., Ganz, P., and Ory, M. (2004), "Quality of life in diverse groups of midlife women: assessing the influence of menopause, health status and psychosocial and demographic factors," *Journal Quality of Life Research*, 13, 933–946.

Barker, D. J., Gluckman, P. D., Godfrey, K. M., Harding, J. E., Owens, J. A., and Robinson, J. S. (1993), "Fetal nutrition and cardiovascular disease in adult life," *Lancet*, 341, 938–941.

Basu, D. (1971), "An Essay on the Logical Foundations of Survey Sampling," in *Foundations of Statistical Inference*, eds. V.P.Godambe and Sproot, D., Toronto: Holt, Rinehard, and Winston.

Belsey, E. and Farley, T. (1987), "The analysis of menstrual bleeding patterns: a review." *Applied Stochastic Models and Data Analysis*, 3, 125–150.

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

Brambilla, D., McKinlay, S., and Johannes, C. (1994), "Defining the perimenopause for applications in epidemiologic investigations," *American Journal of Epidemiology*, 140, 1091–1095.

Carlin, B. and Louis, T. (1996), *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall.

Carroll, R. (2003), "Variance are not always nuisance parameters," *Biometrics*, 59, 211–220.

Chiazze, L., Brayer, F., Macisco, J., Parker, M., and Duffy, B. (1968), "The length and variability of the human menstrual cycle," *Journal of American Medical Association*, 203, 377–380.

Cochran, W. G. (1977), *Sampling Techniques, 3^{rd} Ed.*, New York: John Wiley & Sons, Inc.

Crainiceanu, C., Ruppert, D., Carroll, R., Jpshi, A., and Goodner, B. (2007), "Spatially adaptive Bayesian penalized splines with heteroscedastic errors," *Journal of Computational and Graphical Statistics*, 16, 265–288.

Davidian, M. and Carroll, R. (1987), "Variance Function Estimation," *Journal of American Statistical Association*, 82, 1079–1091.

Den Tonkelaar, I., te Velde, E., and Looman, C. (1998), "Menstrual Cycle Length Preceding Menopause in Relation to Age at Menopause," *Maturitas*, 29, 115–123.

Dennerstein, L., Smith, A., Morse, C., Burger, H., Green, A., Hopper, J., and M., R. (1993), "Menopausal symptomatology in Australian women," *Medical Journal of Australia*, 159, 232–236.

Deville, J. C. and Sarndal, C. E. (1992), "Calibration estimators in survey sampling." *Journal of the American Statistical Association*, 87, 376–382.

Elliott, M. R. (2007), "Bayesian weight trimming for generalized linear regression models," *Survey Methodology*, 33, 22–34.

— (2008), "Model averaging methods for weight trimming," *Journal of Official Statistics*, 24, 517–540.

— (2009), "Model averaging methods for weight trimming in generalized linear regression," *Journal of Official Statistics*, 25, 1–20.

Elliott, M. R. and Little, R. J. A. (2000), "Model based approaches for weight trimming," *Journal of Official Statistics*, 16, 191–210.

Ericson, W. A. (1969), "Subjective bayesian modeling in sampling finite populations." *Journal of the Royal Statistical Society*, Series B, 31, 195–234.

Forrester, T., Wilks, R., Bennett, F., Simeon, D., Osmond, C., Allen, M., Chuang, A., and Scott, P. (1996), "Fetal growth and cardiovascular risk factors in Jamaican schoolchildren." *BMJ*, 312, 156–160.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004), *Bayesian Data Analysis*, Chapman&Hall/CRC.

Gelman, A. and Carlin, J. B. (2002), "Poststratification and weighting adjustments." *Survey Nonresponse*, 289–302.

Ghosh, M. and Lahiri, P. (1988), "Bayes and Empirical Bayes Analysis in Multistage Sampling," *Statistical Decision Theory and Related Topics*, 1, 195–212.

Ghosh, M. and Meeden, G. (1986), "Empirical Bayes estimation of means from stratified samples." *Journal of the American Statistical Assiciation*, 81, 1058–1062.

Gorrindo, T., Lu, Y., Pincus, S., Riley, A., Simon, J. A., Singer, B. H., and Weinstein, M. (2007), "Lifelong Menstrual Histories are Typically Erratic and Trending a Taxonomy," *Menopause*, 14, 74–88.

Gunn, L. and Dunson, D. (2005), "A transformation approach for incorporating monotone or unimodal constraints," *Biostatistics*, 6, 434–449.

Guo, Y., Manatunga, A., Chen, S., and Marcus, M. (2006), "Modeling menstrual cycle length using a mixture distribution," *Biostatistics*, 7, 100–114.

Hall, C., Ying, J., Kuo, L., and Lipton, R. (2003), "Bayesian and profile likelihood change point methods for modeling cognitive function over time," *Computational Statistics & Data Analysis*, 42, 91–109.

Hansen, M. H. and Hurwitz, W. N. (1943), "On the Theory of Sampling from Finite Populations," *The Annals of Mathematical Statistics*, 14, 333–362.

Harlow, S., Cain, K., Crawford, S., Dennerstein, L., Little, R., Mitchell, E., Nan, B., Randolph, J., Jr. Taffe, J., and Yosef, M. (2006), "Evaluation of four proposed bleeding criteria for the onset of late menopausal transition," *Journal of Clinical Endocrinology & Metabolism*, 91, 3432–3438.

Harlow, S. and Campbell, O. (2004), "Epidemiology of Menstrual Disorders in Developing Countries: A Review." *BJOG: An International Journal of Obstetrics and Gynaecology*, 111, 6–16.

Harlow, S., Crawford, S., Dennerstein, L., Burger, H., Mitchell, E., and Sowers, M. (2007), "Recommendations from a multi-study evaluation of proposed criteria for staging reproductive aging," *Climacteric*, 10, 112–119.

Harlow, S., Lin, X., and Ho, M. (2000), "Analysis of menstrual diary data across the reproductive life span: Applicability of the bipartite model approach and the importance of within-woman variance." *Journal of Clinical Epidemiology*, 53, 722–733.

Harlow, S., Mitchell, E., Crawford, S., Nan, B., Little, R., and Taffe, J. (2008), "The ReSTAGE collaboration: defining optimal bleeding criteria for onset of early menopausal transition," *Fertil Steril*, 89, 129–140.

Harlow, S.D.and Ephross, S. (1995), "Epidemiology of menstruation and its relevance to women's health." *Epidemiology Review*, 17, 265–286.

Holt, D. and Smith, T. M. F. (1979), "Poststratification." *Journal of the Royal Statistical Society*, Series A, 142, 33–46.

Horvitz, D. G. and Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe." *Journal of the American Statistical Association*, 47, 663–685.

Huxley, R.R. Shiell, A. and Law, C. (2000), "The role of size at birth and postnatal catch-up growth in determining systolic blood pressure: a systematic review of the literature." *Journal of Hypertension*, 18, 815–831.

Huxley, R. N. and A. Collins, R. (2002), "Unravelling the fetal origins hypothesis: is there really an inverse association between birth weight and subsequent blood pressure?" *Lancet*, 360, 659–656.

Isaki, C. T. and Fuller, W. A. (1982), "Survey design under a regression superpopulation model." *Journal of the American Statistical Association*, 77, 89–96.

Jin, Z., Lin, D., and Ying, Z. (2006), "On least-squares regression with censored data," *Biometrika*, 93, 147–161.

Kalton, G. (2002), "Models in the Practice of Survey Sampling (Revisited)," *Journal of Official Statistics*, 18, 129–154.

Kaufman, L. and Rousseeuw, P. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, New York, NY: Wiley-Interscience.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons, Inc.

Kish, L. and Frankel, M. (1974), "Inference from Complex Samples," *Journals of the Royal Statistical Society*, 36, 1–37.

Klein, J. and Moeschberger, M. L. (2003), *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, 2nd ed.

Korn, E. and Graubard, B. (1999), *Analysis of Health Surveys*, New York: John Wiley & Sons, Inc.

Laird, N. M. and Ware, J. H. (1982), "Random effects models for longitudinal data," *Biometrics*, 38, 963–974.

Lazzeroni, L. C. and Little, R. J. A. (1998), *Random-effects models for smoothing post-stratification weights.*, New York: John Wiley & Sons, Inc.

Lindstrom, M. and Bates, D. (1988), "Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of the American Statistical Association*, 83, 1014–1022.

Lisabeth, L., Harlow, S., Gillespie, B., Lin, X., and Sowers, M. (2004b), "Staging reproductive aging: a comparison of proposed bleeding criteria for the menopausal transition." *Menopause*, 11, 186–197.

Lisabeth, L., Harlow, S., and Qaqish, B. (2004a), "Marginal models for mean and variance: a new approach to modelling menstrual variation across the menopausal transition." *Journal of Clinical Epidemiology*, 57, 484–496.

Little, R., Lewitzky, S., Heeringa, S., Lepkowski, J., and Kessler, R. (1997), "Assessment of Weighting Methodology for the National Comorbidity Survey," *American Journal of Epidemiology*, 146, 439–449.

Little, R. and Rubin, D. (2002), *Statistical Analysis with Missing Data*, John Wiley.

Little, R. J. A. (1983), "Estimating a finite population mean for unequal probability samples," *Journal of the American Statistical Association*, 78, 596–604.

— (1991), "Inference with survey weights," *Journal of official statistics*, 7, 405–424.

— (1993), "Poststratification: A modeler's perspective." *Journal of the American Statistical Assiciation*, 88, 1001–1012.

— (2004), "To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling," *Journal of the American Statistical Assiciation*, 99, 546–556.

Mansfield, P., Carey, M., Anderson, A., Barsom, S., and Koch, P. (2004), "Staging the menopausal transition: Data from the TREMIN research program on women's health," *Women's Health Issues*, 24, 220–226.

Mansfield, P. and Voda, A. (1997), "Women-centered information on menopause for health care providers:Findings from the Midlife Women's Health Survey." *Health Care for Women International*, 18, 55–72.

Matsumoto, S., Nogami, Y., and Ohkuri, S. (1962), "Statistical studies on menstruation: A criticism on the definition of normal menstruation." *Gunma Journal of Medical Sciences.*, 11, 294–318.

Matsumoto, S., Tamada, T., and Konuma, S. (1979), "Endocrinological analysis of environmental menstrual disorders." *International Journal of Fertility*, 24, 233–239.

Matthes, J., Lewis, P., Davies, D., and Bethel, J. (1994), "Relation between birth weight at term and systolic blood pressure in adolescence." *BMJ*, 308, 1074–1077.

Mitchell, E., Woods, N., and A., M. (2000), "Three stages of the menopausal transition from the Seattle Midlife Women's Health Study: toward a more precise definition." *Menopause*, 7, 334–349.

Oh, H. L. and Scheuren, F. J. (1983), "Weighting Adjustment for Unit Nonresponse." *Incomplete Data in Sample Surveys*, 2, 143–184.

Owen, C., Whincup, P., Odoki, K., Gilg, J., and Cook, D. (2003), "Birth Weight and Blood Cholesterol Level: A Study in Adolescents and Systematic Review." *Pediatrics*, 111, 1081–1089.

Potter, F. (1990), "A study of procedures to identify and trim extreme sample weights." in *Proceedings of Survey Research Methods Section*, American Statistical Association, pp. 225–230.

Rasbash, J. and Goldstein, H. (1994), "Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model," *Journal of Educational and Behavioral Statistics*, 19, 337–350.

Rubin, D. B. (1983), "Comments: Probabilities of Selection and Their Role for Bayesian Modeling in Sample Surveys," *Journal of the American Statistical Assiciation*, 78, 803–805.

— (1987), *Multiple imputation for non-response in surveys.*, New York: John Wiley & Sons, Inc.

Scott, A. J. (1977), "Large Sample Posterior Distributions in Finite Populations," *The Annals of Mathematical Statistics*, 42, 1113–1117.

Skinner, D. J. Holt, D. and Smith, T. M. F. (1979), "Analysis of complex surveys." *Journal of the Royal Statistical Society*, Series A, 142, 33–46.

Small, C., Manatunga, A., Klein, M., Feigelson, H., Dominguez, C., McChesney, R., and Marcus, M. (2006), "Menstrual cycle characteristics: associations with fertility and spontaneous abortion." *Epidemiology*, 17, 52–60.

Soules, M., Sherman, S., Parrott, E., Rebar, R., Santoro, N., Utian, W., and N., W. (2001), "Stages of Reproductive Aging Workshop (STRAW)." *Journal of Womens Health and Gender-Based Medicine*, 10, 843–848.

Sowers, M., Jannausch, M., McConnell, D., Little, R., Greendale, G., Finkelstein, J., Neer, R., Johnston, J., and Ettinger, B. (2006), "Hormone predictors of bone mineral density changes during the menopausal transition," *Journal of Clinical Endocrinology & Metabolism*, 91, 1261–1267.

Taffe, J. and Dennerstein, L. (2002), "Menstrual patterns leading to the final menstrual period," *Menopause*, 9, 32–40.

Taylor, G. J. and Wainwright, P. (2005), "Open label extension studies: research or marketing?" *BMJ*, 331, 572–574.

Taylor, R. J., Berger, G., and Treloar, A. (1977), "Changes in menstrual cycle length and regularity after use of oral contraceptives." *International Journal of Gynaecology and Obstetrics*, 15, 55–59.

Thum, Y. and Bhattacharya, S. (2001), "Detecting a change in school performance: a Bayesian analysis for a multilevel joint point problem," *Journal of Educational and Behavioral Statistics*, 26, 443–468.

Treloar, A. (1981), "Menstrual cyclicity and the pre-menopause," *Maturitas*, 3, 249–264.

Treloar, A. and Behn, B. (1971), "Effects of oral contraceptives on the human menstrual cycle." Proceedings of the VII World Congress, Tokyo and Kioto, Japan, Fertility and Sterility, pp. 886–890.

Treloar, A., Boynton, R., Behn, B., and Brown, B. (1967), "Variation of the human menstrual cycle through reproductive life." *International Journal of Fertility*, 12, 77–113.

Vollman, R. (1977), *The menstrual cycle.*, Philadelphia, PA: W.B. Saunders and Company.

Wainwright, P. (2002), "Consent to open label extension sstudies: some ethical issues," *Journal of Medical Ethics*, 28, 373–376.

Wallace, R., Sherman, B., Bean, J., Treloar, A., and Schlabaugh, L. (1979), "Probability of Menopause with Increasing Duration of Amenorrhea in Middle-Aged Women," *American Journal of Obstetrics and Gynecology*, 135, 1021–1024.

Wegienka, G. and Baird, D. (2003), "Potential bias due to excluding oral contraceptive users when estimating menstrual cycle characteristics," *American Journal of Epidemiology*, 158, 947–950.

Weinstein, M., Gorrindo, T., Riley, A., Mormino, J., Niedfeldt, J., Singer, B., and Rodriguez, G. (2003), "Timing of Menopause and Patterns of Menstrual Bleeding," *American Journal of Epidemiology*, 158, 782–291.

WHO (1996), "Research on Menopause in the 1990's," WHO Technical Report Series, World Health Organization, Geneva, Switzerland.

Wildman, R., Colvin, A., Powell, L., Matthews, K., Everson-Rose, S., Hollenberg, S., Johnston, J., and Sutton-Tyrrell, K. (2008), "Associations of endogenous sex hormones with the vasculature in menopausal women: the Study of Women's Health Across the Nation (SWAN)," *Menopause*, 15, 414–421.