AN ABSTRACT OF THE DISSERTATION OF

John Giovanini for the degree Doctor of Philosophy in Statistics presented on June, 6 2008.
Title: Generalized Linear Mixed Models with Censored Covariates

Abstract approved: ______________________________
Daniel W. Schafer

This dissertation is about statistical methods for data analysis using generalized linear mixed models (GLMMs) with censored covariates. Special attention in given to the particular problem of inference about age-specific reproductive success in wild animal populations using some animals with known ages and some animals with ages only known to exceed some lower bound. GLMMs allow for non-normal response distributions, such as a Poisson distribution for the number of offspring from a parent in one year, and they account for the correlation of repeated responses from the same observational unit, such as the correlation of the number of offspring from the same parent over multiple years. A computational algorithm for maximum likelihood estimation and two approximate estimation methods are proposed. The full solution uses the EM algorithm with Markov Chain Monte Carlo techniques for the E-step. The approximations are presented as techniques that may be nearly as good as the full maximum likelihood analysis but that are easier for wildlife biologists to use. One uses a Laplace approximation to the log-likelihood to capitalize on existing programs for GLMM estimation. The other is a regression calibration method in which the missing ages are simply replaced by predicted values. The full likelihood analysis is demonstrated on a study of age-specific reproductive success of Northern Spotted

Owls (*Strix occidentalis caurina*). A simulation study was used to evaluate the operating characteristics of the three methods and to highlight the potential gains of these methods over the common practice of ignoring animals with unknown ages. The conditions of the simulations were chosen to roughly match those in the spotted owl study. It appears that the use of the owls with censored ages reduces the widths of 95% confidence intervals for important regression coefficients by about 39% if full maximum likelihood analysis is used. The corresponding reduction for the regression calibration estimator is about 27%. A main conclusion of this thesis is that the regression calibration estimator can offer substantially higher efficiency than the commonly used GLMM estimator with animals of unknown ages excluded and, importantly, wildlife biologists can use it with computing modules that are already available in standard statistical computing packages.

Generalized Linear Mixed Models with Censored Covariates

by
John Giovanini

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented June 6, 2008
Commencement June 2009

Doctor of Philosophy dissertation of John Giovanini presented on June 6, 2008.

APPROVED:

Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

John Giovanini, Author

ACKNOWLEDGEMENTS

I would like to thank Dr. Dan Schafer for being my advisor during my dissertation research. He has been extremely helpful during the numerous times that I have needed a little push in the right direction. He always strived to make my dissertation a polished and professional work. I would also like to thank Dr. Alix Gitelman, Dr. Lisa Madsen, Dr. Cliff Pereira, and Dr. Bruce McGough for serving on my committee.

I am grateful for Pete Loschl and Eric Forsman for allowing us to work on the Northern Spotted Owl data set. Pete has always been willing to answer questions about his data set.

I would also like to thank the numerous teachers I have had during my twenty three years of formal education. I would like to mention three teachers who had a significant impact on my life. As an undergrad at the University of Montana, Dr. Johnny Lott was a mentor and friend. Also at the University of Montana, Dr. Jon Graham introduced me to the field of statistics. During my time as a research assistant at Oregon State University, I have learned a great deal from Dr. Cliff Pereira about being a consulting statistician. I would like to thank all the faculty, staff, and student s of the Oregon State University Statistics Department for everything that they have done, especially Dr. Dave Birkes for teaching most of the statistics PhD classes.

My parents, Bob and Lois Giovanini, deserve credit for instilling a love of math and learning. Finally, I would like to thank my wife Joy. Without her love and support, I could not have gotten this far. I am eternally grateful for all of the sacrifices that she has made during my time as a graduate student.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF APPENDIX FIGURES

LIST OF APPENDIX FIGURES (Continued)

LIST OF APPENDIX TABLES

Generalized Linear Mixed Models with Censored Covariates

1. INTRODUCTION

In a typical longitudinal study of age-specific reproductive success, researchers observe many animals over the course of many years, and observe one or more measures of reproductive success on each animal each year. The measure could be a binary outcome for a successful mating, a count of the number of young produced in a year, or a count of the number of young produced in a year that survive and contribute to the breeding population. The researchers may wish to estimate the age profile for mean reproductive success, to determine the ages at optimal reproductive success, to test for a decline in reproductive success at older ages, or to examine physiological or environmental variables that are associated with yearly reproductive success (after accounting for the effect of age). They may wish to do these examinations either separately for males and females or jointly. References on reproductive success include Clutton Brock (1988) and others.

If the measures of reproductive success are binary, sums of Bernoulli trials, or small integer counts, it is appropriate for researchers to consider binomial or Poisson regression models, i.e. generalized linear models (GLMs). In longitudinal studies, a response is measured repeatedly on each of several animals and different responses from a given animal should not be considered independent. A common modeling approach for handling this lack of independence involves the inclusion of random effects for the different animals. Generalized linear mixed models (GLMMs; see McCulloch and Searle, 2001) permit the inclusion of such random effects along with fixed effects in generalized linear models. They also allow for departures from the binomial and Poisson distributions with an additional parameter for extra-binomial or extra-Poisson variation. GLMMs have been used for studying age-specific reproductive success in barn owls (Altwegg et al. 2007), snow petrels (Angelier et al. 2007), brown thornbills (Green 2001), goshawks (Kruger 2005), brown bears (Zedrosser, et al. 2007), mountain goats (Côté, Festa-Bianchet 2001), and sparrowhawks (Newton, Rothery 2002).

GLMMs offer a useful approach to these biological investigations. They permit the inclusion of time-dependent and time-independent explanatory variables, detailed modeling of age effects and variance terms that lead to conclusions about between-male and between-female variability. Furthermore, easy-to-use computer routines are available, such as GLIMMIX in SAS (Schabenberger 2005 and Schabenberger 2007) and lmer in R (Bates 2005 and Bates 2007).

In practice, researchers observe reproductive success measures on some of the animals with known ages (because the animals have been observed their entire lives) and some animals of unknown age (because they were adults at the time of their first observation). A common practice for investigation of age-specific reproductive success in this case is to use only those animals with known ages. This is true in studies of snow petrels (Angelier et al. 2007), barn owls (Altwegg et al. 2007), goshawks (Kruger 2005), and sparrowhawks (Newton, Rothery 2002). In a northern spotted owl (Strix occidentalis caurina) investigation, which motivated this work, yearly reproductive responses were observed on 404 male owls of known ages, 463 male owls of unknown ages, 463 female owls of known age, and 579 female owls of unknown ages.

There is some information available, though, from the owls of unknown ages, and there may be substantially more power and precision available for answering the scientific questions of interest by incorporating them in the analysis. In fact, the ages are known to exceed some number, so they are "right censored." At the time that an adult spotted owl is first observed, for example, the biologists can conclude from its adult characteristics that it is at least 3 years old. Furthermore, if such an owl is observed in the reproductive success study for 10 years, then it is known to be at least 13 years old in the 10th year of observation. It seems very likely that this partial information could be useful in the estimation of age-specific reproductive success models.

This dissertation, therefore, is about GLMM estimation with right censored explanatory variables, with specific attention to the problems inherent in studies of age-specific reproductive success. The goals are to describe the increased efficiency that is available by incorporating animals whose ages are only known to exceed some number of years, to describe potential biases that result by replacing unknown ages by some fixed number, and to describe approaches for GLMM estimation that include animals with censored ages. Particular emphasis will be on the study of approaches that are easy to use with currently available software.

1.1 Example

In a recent study on Northern Spotted Owls (Loschl 2008) researchers are interested in determining the age of peak reproductive output and other features of age-specific reproductive success. The study consists of three areas in Oregon and one area in Washington, collected from 1984 to the 2005. One response of interest is the number of young fledged per year, which ranges from 0 to 3 (with 3 being very rare).

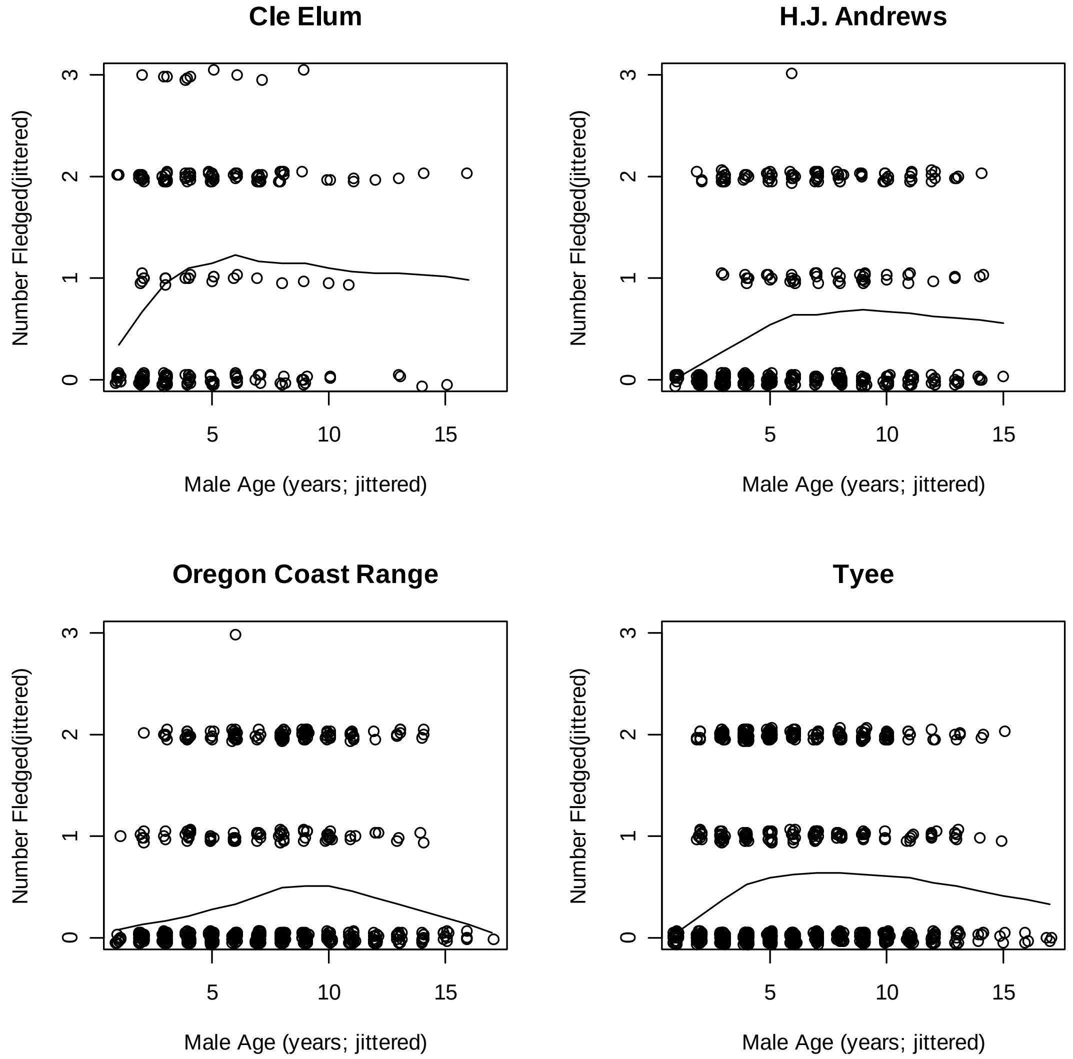


Figure 1.1 Number of fledglings versus male age the four study areas. The solid lines are lowess (locally-weighed polynomial regression smoother) curves.

Figure 1.1 shows the number fledged for the known-age males vs. the male age for the four study areas. There appears to be an increase in the number fledged at early ages, with some leveling off and possibly a decrease in reproductive success as the male age increases.

If the birds were first observed and banded as fledglings or sub-adults, then the exact age is known. These birds were recorded as 0, 1, or 2 year olds depending on physical characteristics. However, if the birds were first observed as an adult, then it is only known that their age at first observation is at least 3 years. At one of the study sites, only 40% of the owl/years observed were of known age owls. Restricting attention to known age owls is an easy solution but the inclusion of the additional owls with censored ages—using statistical techniques for handling missing data—may result in important gains in efficiency and precision in answering the scientific questions of interest.

1.2 Contributions of the Dissertation

In this dissertation, we propose three methods for fitting GLMMs with censored explanatory variables. The first is an MCEM algorithm for maximum likelihood estimation that treats both the random effects and known ages as "missing data." The second is an approximate MCEM algorithm that only treats the known ages as "missing data" and uses a Laplace approximation to integrate out the random effects. The third is a regression calibration method that replaces the censored ages with their expected value given the lower bound and possible covariates.

Our goals are to clarify the potential efficiency gains and to provide techniques for incorporating animals with censored ages into the GLMM analysis for age-specific reproductive success. We are particularly interested in finding easy-to-use solutions, if possible. Our statistical interest is in the Spotted Owl study in particular, but the same data characteristics are found in other studies of age-specific reproductive success. We can imagine that there are other applications of GLMMs in which a covariate is censored and our results would pertain to those. Austin and Hoch (2004) report a regression problem, for example, in which the explanatory variable “household income” is obtained from a survey questionnaire with highest category “greater than $80,000.” The main effort here, though, is directed towards the particular problems encountered in the study of age-specific reproductive success from wild animal populations.

Chapter 2 presents a likelihood analysis based on the Markov Chain Expectation Maximization (MCEM) Algorithm. The E-step of the algorithm uses Monte Carlo simulation to approximate the expected value of the “complete data” log likelihood and the simulation requires the use of the Metropolis-Hastings algorithm (as in McCulloch and Searle, 2001). While the M-step of the algorithm is fairly straightforward, the overall algorithm is intricate, slow, and probably difficult for non-statisticians to understand.

One approximation in Chapter 3 uses a Laplace approximation to the complete data log likelihood, which permits the use of existing GLMM modules within a broader EM Algorithm. The E-step, however, still requires the Metropolis-Hastings algorithm and, therefore, does not go very far in simplifying the more full likelihood approach.

The regression calibration approach, which is based on a commonly used technique for regression estimation in the presence of explanatory variable measurement errors, is more intuitive and easy to use. In a first stage of the analysis, the user estimates the unknown parameters in a probability distribution for the covariate that is censored on some subjects. In a second stage, a GLMM module is used but with censored ages replaced by their estimated expectations given that they exceed the specified lower bound (and with unknown parameters replaced by their first-stage estimates).

A simulation study in Chapter 3 clarifies the sampling distributions of the three estimators and compares their characteristics to those of the naïve estimator that excludes animals with unknown ages and to the naïve estimator that replaces unknown ages by their lower bounds. There is convincing evidence of substantial gains in precision due to including the subjects with censored covariates.

The simulation suggests that the regression calibration estimator performs quite well. Given it's simplicity, it strikes us as the right approach for biologists to use. Furthermore, it allows for more complex modeling of random effects. In particular, the user can formulate a model in which there are random effects due to the male parent and to the female parent in a single model (provided there are enough partner changes to make the model identifiable), and to include random effects due to different years. While it would be possible to include these multiple random effects in the MCEM approaches, the complexity and slowness would make them practically unusable.

1.3 Organization of the Dissertation

The rest of the dissertation proceeds as follows. In Chapter 2, we propose a Monte Carlo EM algorithm for generalized linear mixed models with censored covariates that treats both the random effects and the censored explanatory variables as "missing data." This algorithm is then demonstrated on a subset of the Northern Spotted Owl data set. Chapter 3 details the two easier-to-compute approximations and a simulation study, which shows the increased precision due to including owls with censored ages and clarifies the operating characteristics of the three estimators that do include these owls. Chapter 4 contains a discussion of the conclusions and possible directions of future research. A major conclusion is that the regression calibration procedure offers an easy-to-apply approach that can substantially improve power and efficiency by including those units with censored covariates.

2. Likelihood Analysis for Generalized Linear Mixed Models with Censored Covariates

John N. Giovanini[*] and Daniel W. Schafer[**]
Department of Statistics,
Oregon State University,
Corvallis, OR 97331
[*]email: giovanin@science.oregonstate.edu
[**]email: schafer@science.oregonstate.edu

2.1 Abstract

This paper is about likelihood analysis of generalized linear mixed models (GLMMs) when some observational units have censored values of an explanatory variable. Special attention in given to age-specific reproductive success studies from wild animal populations when some animals have known ages and some have ages that are only known to exceed a lower bound. GLMMs permit a small integer count response—such as the number of offspring produced in a season—and address the non-independence of repeated observations on the same animal in different seasons with random animal effects. A Monte Carlo Expectation Maximization (MCEM) algorithm is proposed for maximum likelihood analysis. The random effects and the censored covariates are both treated as "missing data." The algorithm is demonstrated on a recent Northern Spotted Owl dataset.

2.2 Introduction

This work was motivated by an investigation of factors affecting individual reproductive success in a wild animal population. In particular, Figure 2.1 shows the number of fledged spotted owls in a year versus male parent age in that year, for multi-year observations on 108 male northern spotted owls. The solid points are the sample means for each age and the vertical lines are approximate 95% confidence intervals (data from Pete Loschl, personal communication; see Loschl, 2008). The smooth curve is a nonparametric lowess fit, which indicates an apparent increase in mean number fledged up to a maximum of 0.7 fledglings per year at about age 9, with a subsequent decrease.

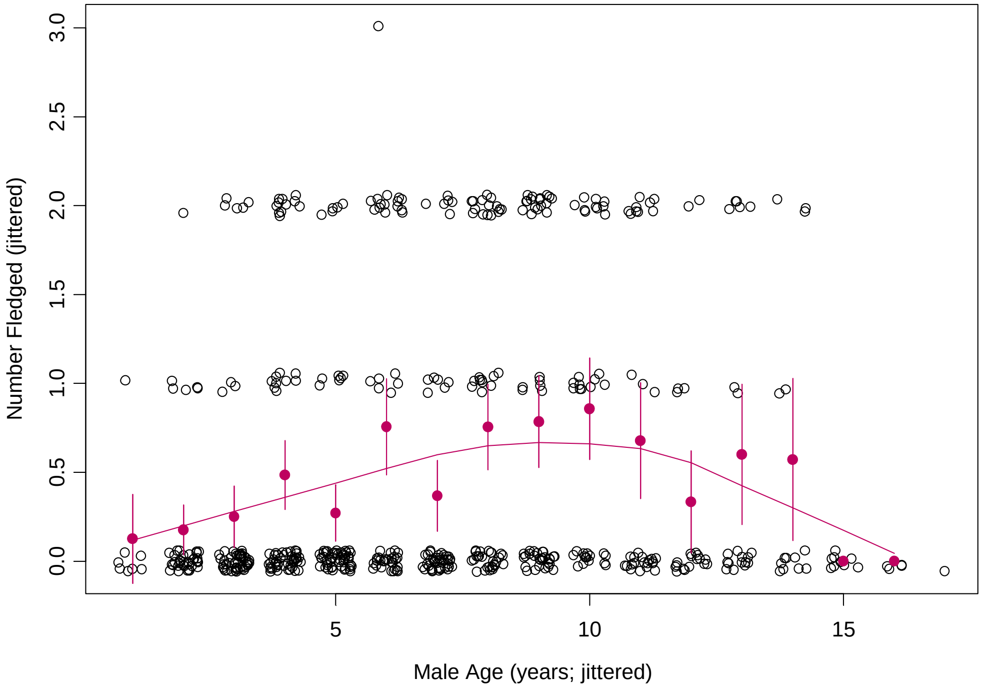


Figure 2.1 Number of fledglings versus male parent age from multiple-year observations on each of 108 Northern Spotted Owl males in the Oregon Coast Range; with sample means for each age, naïve 95% confidence intervals; and a lowess (locally-weighed polynomial regression soother) curve.

More formal investigation into patterns of age-specific reproductive success should account for dependence of different observations from the same male. Figure 2.2 reproduces Figure 2.1 but includes the fit to a generalized linear mixed model (GLMM) that accounts for the dependence with random intercepts for the 108 different males. The solid curve is the GLMM estimate of a log-linear model with linear and quadratic effects of age, averaged over all males. The solid curve is the approximate maximum likelihood estimate of this GLMM. The dashed lines show a 95% confidence band for the mean fixed effect of male age. (This confidence band includes between-owl variability in intercepts.)

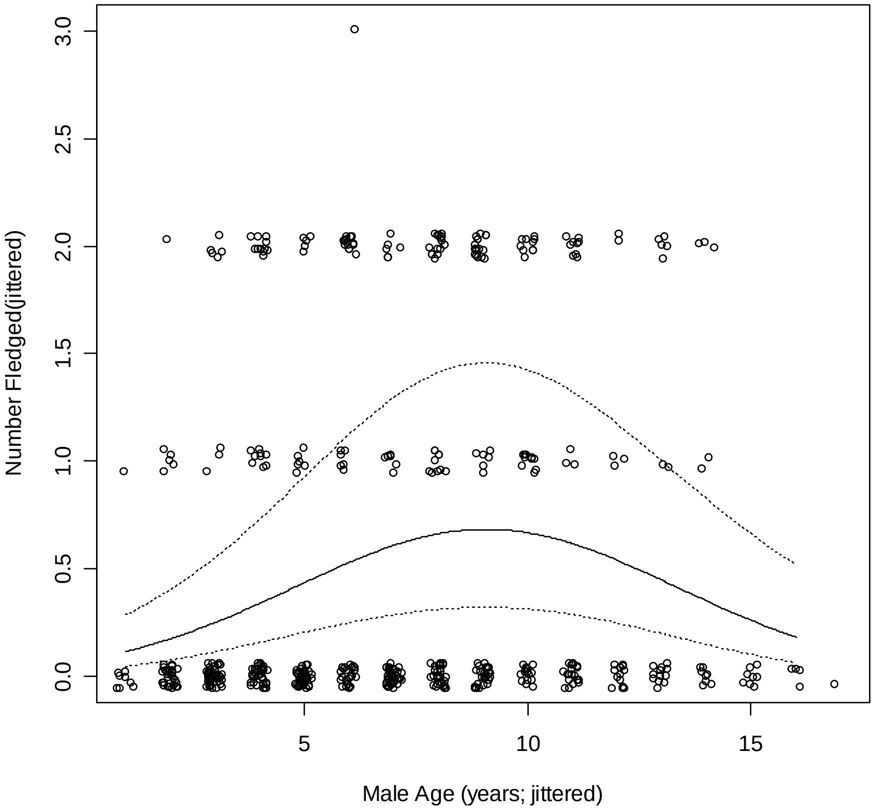


Figure 2.2 Number of fledglings versus male age with GLMM fit for a typical year and an approximate 95% confidence band (Oregon Coast Range study area)

Such a GLMM is useful for investigating several scientific questions about age-specific reproductive success: 1.What evidence is there that male reproductive success decreases in older ages? 2. What is the age at which maximal mean reproductive success is achieved? 3. What proportion of variability in reproductive success can be explained by between-male differences after accounting for effects of age? 4. What evidence is there that various landscape and climate variables affect mean number fledged, after accounting for the effects of parent age?

When number of fledglings or some other small integer count is used as a measure of reproductive success, these types of questions (and similar questions for females) can be addressed with GLMM analysis (McCulloch & Searle 2001 and Jiming, 2007). The measure of reproductive success is taken to have a Poisson (or possibly binomial) distribution with a mean that depends on parent age and other explanatory variables, but with the inclusion of random effects to account for variable reproductive successes between males (or females). This has been used, for example in studies of barn owls (Altwegg et al. 2007), snow petrels (Angelier et al. 2007), brown thornbills (Green 2001), goshawks (Kruger 2005), brown bears (Zedrosser, et al. 2007), mountain goats (Côté, Festa-Bianchet 2001), and sparrowhawks (Newton, Rothery 2002).

Our interest is in the use of GLMMS for this purpose when a substantial number of the animals in the data set have ages that are only known to exceed some lower bound. The plots and fitted models in the figure above, for example, are based on 542 observations from 108 male owls whose ages are known exactly. Also available are 839 additional observations from 165 male owls whose exact ages on their first season of observation are only known to exceed 3 years. When the researchers first band an owl, they might conclude with certainty—from adult characteristics—that the owl is at least 3 years old. In the following year, therefore, they can be sure it is at least 4 years old. After 10 years of observation, the researchers can be sure that the owl is at least 13 years old.

It is common for researchers to exclude the animals with unknown ages from the statistical analysis of age-specific reproductive success. This is true in the studies of spotted owls (Loschl, 2008), snow petrels (Angelier et al. 2007), barn owls (Altwegg et al. 2007), goshawk (Kruger 2005), and sparrowhawks (Newton, Rothery 2002). While the common practice of excluding animals of unknown ages isn't likely to induce any bias into the scientific conclusions, the incorporation of information from the owls with censored ages may provide important gains in efficiency and power. One does need to consider the possible that the age-specific reproductive success curves are different for the known age vs. the censored animals. This is especially true if the animals that are excluded from the analysis are the older animals that were first observed as adults at the beginning of a study.

Notice, for example, that there is some visual indication from Figure 2.2 that the mean number of fledglings decreases with older ages. The fairly wide confidence band at that end of the graph, though, suggests that the evidence for the decrease is not convincing. Including the additional owls with censored ages may result in more precise model estimation and therefore more resolution to this and other scientific questions of interest.

Our goals are to clarify the potential efficiency gains and to provide techniques for incorporating animals with censored ages into the GLMM analysis for age-specific reproductive success. We are particularly interested in finding easy-to-use solutions, if possible.

Our focus is on the EM (expectation-maximization) algorithm for computing maximum likelihood estimates in the presence of missing data (Dempster, Laird, & Rubin 1977 and McLachlan & Thriyambakam 1997). For full likelihood analysis of models for studying age-specific reproductive success, the unavailable true ages and the random effects are all treated as “missing data.” The E-step is accomplished via Markov Chain Monte Carlo techniques. This algorithm parallels one suggested by Wu and Wu (2007) for GLMMs with missing data. Approximations that lead to easier calculations are discussed later.

2.3 Notation and Model Specification.

The model of interest specifies repeated measures on each of *m* subjects (or clusters) with responses that follow a generalized linear model with random intercepts for each subject, with time-dependent and time-independent explanatory variables, and with a time-independent explanatory variable that is censored. Note that the first observed nesting age is a time independent, while the actual age is not. Let y_i represent the response observed for observational unit *i*, for $i = 1,\ldots,n$. In the spotted owl example, the response is the number of young fledged and the “observational unit” is an “owl year” of observation.

Let $\underline{z}_i = (z_{i1},...,z_{im})$ where z_{ij} = 1 if observational unit *i* is associated with subject or cluster *j* and 0 if not; for *i* = 1,…,*n* and *j* = 1,…,*m.* In the example, this variable indicates the particular male associated with observational unit *i.*

Let a_j represent the explanatory variable that is censored on some of the subjects or clusters, for subject or cluster *j*, for *j* = 1,…, *m*, and let $\underline{a} = (a_1,...,a_m)$. In the example, a_j is the age of male owl *j* at the time it was first observed.

Let c_j be a censoring indicator, that takes on the value 0 if a_j is observed and takes on the value 1 if it is only known that a_j is greater than or equal to some known value. Let a_j^* be the true age, a_j, for those owls with known ages and the lower bound for age at first observation otherwise. Let $\underline{a}^* = (a_1^*,...,a_m^*)$. Let $\underline{x}_i$ be a vector of explanatory variables associated with observational unit *i*, which may be time-variant or time- invariant. Let *X* be the matrix whose ith row is $\underline{x}_i^T$. Let $\underline{u} = (u_1,...,u_m)$ represent "random effects" associated with the *M* clusters or subjects.

We suppose that the y_i's are conditionally independent, given $\underline{u}$, with density

$$f(y_i \mid \underline{a},\underline{u},\underline{x}_i) = f(y_i \mid \underline{z}_i^T\underline{a},\underline{z}_i^T\underline{u},\underline{x}_i) \text{ with mean } \mu_i\text{, where}$$

$$g(\mu_i) = \underline{x}_i^T\beta + h(\underline{z}_i^T\underline{a};\alpha) + \underline{z}_i^T\underline{u}\ ,$$

where α and β are *p*- and *q*-vectors of unknown parameters and *g*( ) is a known "link function." In the owl example, the density is taken to be Poisson and the link function is the logarithm. Since we will be using maximum likelihood analysis, we will not be using a dispersion parameter. Such a parameter would complicate the analysis.

The term $h(\underline{z}_i^T \underline{a};\alpha)$ is of unspecified form to permit the incorporation of nonlinear effects of the censored explanatory variable, such as a quadratic effect of age for the model displayed in Figure 2.2. Note that if s_i represents the number of years since the male associated with observational unit *i* was first observed, then $\underline{z}_i^T \underline{a} + s_i$ is its age associated with observational unit *i*. One possible model, for example, is

$$h(\underline{z}_i^T \underline{a} + s_i;\alpha) = \alpha_1(\underline{z}_i^T \underline{a} + s_i) + \alpha_2(\underline{z}_i^T \underline{a} + s_i)^2 .$$

Suppose also that the random effects are independent and identically distributed, and independent of the explanatory variables:

$$u_j \sim f(u_i \mid \underline{a}, X;\tau) = f(u_i;\tau) .$$

In GLMMs, it is convenient to take this distribution to be normal with mean 0 and variance τ .

It is also necessary to assume some distributional model for the marginal distribution of *a*. Let

$$a_j \sim f(a_j \mid \underline{w}_j;\gamma) ,$$

where $\underline{w}_j$ is a vector of explanatory variables that would be useful for predicting a_j. In the owl example, the total number of years that the owl was observed would be such an explanatory variable. We assume that a_j is independent of u_j and of $a_{j'}$ for $j \neq j'$.

2.4 EM Algorithm Treating Random Effects and Unknown Ages as "Missing Data"

The EM Algorithm is often used for estimation of parameters in GLMMs without the additional problem of censored covariates (McCulloch 1997 and Booth & Hobert 1999), by treating the random effects, $\underline{u}$, as missing data. The approach here is to use the same techniques, extended to also treat the unknown ages at first observation, $\underline{a}$, as missing. Let $\theta = (\alpha, \beta, \gamma, \tau)$ denote the vector of unknown parameters. The "observed data" are $\underline{y}$ and $\underline{a}^*$. The "complete data" are $\underline{y}$, $\underline{u}$ and $\underline{a}$. The complete data log likelihood is:

$$l_c(\theta; \underline{y}, \underline{a}) = \log[f(\underline{y}, \underline{u}, \underline{a} \mid X; \theta)]$$

$$= \log[f(\underline{y} \mid \underline{u}, \underline{a}, X; \theta)] + \log[f(\underline{u} \mid \underline{a}, X; \theta)] + \log[f(\underline{a} \mid X; \theta)]$$

$$= \sum_{i=1}^{n} \log[f(y_i \mid \underline{z}_i^T \underline{u}, \underline{z}_i^T \underline{a}, x_i; \alpha, \beta, \tau)] + \sum_{j=1}^{m} \log[f(u_j; \tau)] + \sum_{j=1}^{m} \log[f(a_j \mid c_j, \underline{w}_j; \gamma)] .$$

The E-step (expectation) requires the expectation of the complete data log likelihood given the observed data and with unknown parameters in the expectation replaced by their estimates after *t* iterations. Let $\theta^{(t)}$ denote the estimate of θ after *t* iterations of the EM algorithm. Then the expectation is:

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^{n} E\left\{ \log[f(y_i \mid z_i^T \underline{a}, x_i; \alpha, \beta, \tau)] \mid \underline{y}, X, \underline{a}^*; \theta^{(t)} \right\} +$$

$$\sum_{j=1}^{m} E\left\{ \log[f(u_j; \tau)] \mid \underline{y}, X, \underline{a}^*; \theta^{(t)} \right\} +$$

$$\sum_{j=1}^{m} E\left\{\log[f(a_j \mid \underline{w}_j;\gamma)] \mid \underline{y}, X, \underline{a}^*;\theta^{(t)}\right\}.$$

The expectations are with respect to the distribution of $(\underline{u},\underline{a})$ given $\underline{y}$, $\underline{a}^*$, and X. In general, the expectations are intractable, but they can be approximated by Monte Carlo methods in a way that parallels the approach for maximum likelihood with random effects alone (McCulloch and Searle, 2001, Sect. 10.3):

$$Q(\theta \mid \theta^{(t)}) \approx \sum_{i=1}^{n}\sum_{r=1}^{R}\frac{1}{R}\log[f(y_i \mid \underline{z}_i^T\underline{u}^{(r)}, \underline{z}_i^T\underline{\tilde{a}}^{(r)}, \underline{x}_i;\alpha,\beta) +$$

$$\sum_{j=1}^{m}\sum_{r=1}^{R}\frac{1}{R}\log[f(u_j^{(r)};\tau)] +$$

$$\sum_{j=1}^{m}\sum_{r=1}^{R}\frac{1}{R}\log[f(a_j^{(r)} \mid c_j, \underline{w}_j;\gamma)]$$

where $\underline{u}^{(r)} = (u_1^{(r)},...,u_m^{(r)})$ and $u_j^{(r)}$ is a pseudo-random variable generated from $f(u_j \mid \underline{y}, X, \underline{a}^*;\theta^{(t)})$ and $\underline{a}^{(r)} = (a_1,...,a_{m_c}, a_{m_c+1}^{(r)}..., a_m^{(r)})$, where the first m_c elements are known ages of first observation and where the remaining elements, $a_j^{(r)}$, are pseudo-random variables generated from $f(a_j \mid \underline{w}_j, \underline{y}, X, a_j^*;\theta^{(t)})$.

The EM algorithm is an iterative algorithm that, at each iteration, updates the expectations $Q(\theta \mid \theta^{(t)})$ based on current parameter estimates and then computes updated estimates as those values that maximize $Q(\theta \mid \theta^{(t)})$. The following steps describe the algorithm:

1. Choose starting values, $\theta^{(0)} = (\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}, \tau^{(0)})$. Set t = 1.
2. Repeat until convergence:

a. Generate R vectors $(\underline{u}^{(r)}, \underline{a}^{(r)})$ from the conditional distribution of $(\underline{u}, \underline{a})$ given $\underline{y}$, X, and $\underline{a}^{*}$ with unknown parameters θ in the distribution replaced by the "current" estimates $\theta^{(t-1)}$

b. Calculate $\alpha^{(t)}$ and $\beta^{(t)}$ as those values that maximize

$$\sum_{i=1}^{n}\sum_{r=1}^{R}\frac{1}{R}\log[f(y_i \mid \underline{z}_i^T\underline{u}^{(r)}, \underline{z}_i^T\underline{\tilde{a}}^{(r)}, \underline{x}_i; \alpha, \beta)]$$

c. Calculate $\tau^{(t)}$ as the value that maximizes

$$\sum_{j=1}^{m}\sum_{r=1}^{R}\frac{1}{R}\log[f(u_j^{(r)}; \tau)]$$

d. Calculate $\gamma^{(t)}$ as the value that maximizes

$$\sum_{j=1}^{m}\sum_{r=1}^{R}\frac{1}{R}\log[f(a_j^{(r)} \mid \underline{w}_j; \gamma)]$$

e. Set $R = R + \lfloor R/c \rfloor$, for some $c > 0$

f. Set $t = t+1$.

Each of the pieces (b), (c), and (d) can be accomplished with formulas or routines that would be appropriate in the absence of censored explanatory variables and random effects, but based on the augmented data set corresponding to the R pseudo values for $(\underline{u}, \underline{a})$. Notice, in particular, that (b) can be accomplished with a GLM estimation algorithm, specifying the generated $\underline{z}_i^T\underline{u}$ terms as "offsets."

The random number generation in 2a can be accomplished with the Metropolis-Hastings algorithm as follows (see, for example, McCulloch and Searle, 2001, Sect. 10.3):

1. Generate $\underline{u}^{(0)} = (u_1^{(0)},...,u_m^{(0)})$ with $u_j^{(0)}$ generated from $f(u_j;\tau^{(t-1)})$ and $\underline{a}^{(0)} = (a_1,...,a_{m_c}, a_{m_c+1}^{(0)}...,a_m^{(0)})$ with $a_m^{(0)}$ generated from $f(a_j \mid \underline{w}_j, a_j^*;\gamma^{(t-1)})$. Set *r* = 1.
2. For *k* from 1 to (*R* + *R**):
   a. Generate $\underline{\tilde{u}} = (\tilde{u}_1,...,\tilde{u}_m)$ with $\tilde{u}_j$ generated from $f(u_j;\tau^{(t-1)})$

      Generate $\underline{\tilde{a}} = (a_1,...,a_{m_c},\tilde{a}_{m_c+1}...,\tilde{a}_m)$ with $a_m^{(k)}$ element generated from $f(a_j \mid \underline{w}_j, a_j^*;\gamma^{(t-1)})$

   b. Compute the acceptance criterion:

$$p_k = \min\left\{1, \frac{\prod_{i=1}^{n} f(y_i \mid \underline{z}_i^T\underline{\tilde{u}}, \underline{z}_i^T\underline{\tilde{a}}, \underline{x}_i;\alpha^{(t-1)},\beta^{(t-1)})}{\prod_{i=1}^{n} f(y_i \mid \underline{z}_i^T\underline{u}^{(k-1)}, \underline{z}_i^T\underline{a}^{(k-1)}, \underline{x}_i;\alpha^{(t-1)},\beta^{(t-1)})}\right\}$$

   c. Generate *v*, a Bernoulli(p_k) random variable:

   d. If *v* =1 set $\underline{u}^{(k)} = \underline{\tilde{u}}$ and $\underline{a}^{(k)} = \underline{\tilde{a}}$. Otherwise, set $\underline{u}^{(k)} = \underline{u}^{(k-1)}$ and $\underline{a}^{(k)} = \underline{a}^{(k-1)}$

3. Retain the final *R* of each of these vectors as the simulated sample. (*R** is the burn-in number.)

To speed up convergence, several authors (Levine and Fan, 2003 and Levine and Casella, 2001), recommend using importance weights instead of drawing a new MCMC sample at each iteration. The use of importance weights can greatly decrease the convergence time because generating the pseudo-random variables via the

Metropolis Algorithm is computationally more intensive than generating the importance weights. It is recommended that a burn-in period of regular Monte Carlo EM iterations is used before switching to the importance weights. The burn-in allows the target and the candidate distribution to be ‘closer’ and therefore helps decrease the convergence time.

In the above algorithm, we increase the Monte Carlo sample size *R* using Booth and Hobert’s (1999) recommendation of $R = R + \lfloor R/c \rfloor$, for some $c > 0$. This method is used because at early iterations, when the “current” parameter estimates are likely far from the MLE, one does not need a large sample size. However, as the “current” parameter estimates get closer to the MLE, one needs more precision and therefore a larger Monte Carlo sample size. Instead of using a naïve increase of the Monte Carlo sample size, Levine and Fan (2003), and Levine and Casella (2001) suggest automated algorithms that increase the Monte Carlo sample size after checking if the Monte Carlo error overwhelmed the EM estimate. For our specialized algorithm, we simply used the naïve increase.

2.5 Computing Standard Errors

Approximate standard errors can be calculated using McLachlan and Krishnan’s (1997) method. This method uses only first-order derivatives to find the approximate information matrix:

$$I\left(\hat{\theta}\right) \approx \sum_{j=1}^{m}\sum_{r=1}^{R}\frac{1}{R}S_{jr}\left(\hat{\theta}\right)S_{jr}^{T}\left(\hat{\theta}\right), \text{ where } S_{jr}\left(\hat{\theta}\right) = \left.\frac{\partial l_c^{(r)}\left(\theta; \underline{y}_j, \underline{x}_j,\ u_j^{(r)}, a_j^{(r)}\right)}{\partial \theta}\right|_{\theta=\hat{\theta}}$$

The standard errors for the MLE can then be calculated by taking the square root of the diagonal elements of $I^{-1}\left(\hat{\theta}\right)$.

There are several other methods for calculating the information matrix, including calculating the observed Fisher Information from the observed data log-likelihood, as well as Oakes (1999) and Louis' (1982) method of writing the observed data log-likelihood as functions of the complete data log-likelihood. Both of these methods would be rather complicated in our situation.

2.6 Analysis of Spotted Owl Data

The Monte Carlo EM algorithm for censored covariates in GLMMs will now be used to analyze one of the study areas from the Spotted Owl example from Section 1.2. The Oregon Coast Range study area is located in the central Coast Ranges of Oregon. There are 108 known age owls and 165 owls with censored ages. The known age owls have 542 owl/year observations, while the censored owls have 839 owl/year observations.

This analysis will examine two particular questions that the wildlife biologists are interested in gaining insight on. First, the biologists would like to know the age at peak reproductive success, after accounting for differences between years. The biologists are also interested in possible declines in reproductive success after reaching peak performance.

To answer these questions, we fit a Poisson log-linear model for mean number fledged, with linear and quadratic effects of male parent age, with year as a factor with

15 levels, and with random intercepts for the 273 male owls. The intercepts are treated as a random sample from a normal distribution. The age of first observed nesting is assumed to be normally distributed conditional on the number of years that the owl was observed.

Starting values were obtained by fitting a GLMM with the censored ages replaced with their estimated expected value given that they exceed a lower bound. (This is the “regression calibration” estimator, which is discussed more fully in Chapter 3.) A table of the parameter estimates and standard errors is shown below.

Table 2.1 MCEM estimation results for Oregon Coast Range

	MCEM All Males		lmer Known Age Males Only	
Parameter	Estimate	SE	Estimate	SE
β_0	-0.6263	0.2315	-1.8092	0.6796
β_{age}	0.0776	0.0164	0.4824	0.0944
β_{age^2}	-0.0029	0.0005	-0.0269	0.0058
β_{1991}	-1.4385	0.3206	-15.8338	2211.3104
β_{1992}	-0.0214	0.2592	-15.9970	1157.3002
β_{1993}	-1.3692	0.3044	-0.9729	0.9513
β_{1994}	-0.1321	0.2601	-0.7218	0.7130
β_{1995}	-1.6351	0.3024	-1.7309	0.7672
β_{1996}	0.2211	0.2482	-0.0933	0.6490
β_{1997}	-0.9423	0.2576	-1.1617	0.6750
β_{1998}	-0.3310	0.2439	-0.3259	0.6477
β_{1999}	-1.8897	0.3255	-2.1913	0.7342
β_{2000}	-0.4901	0.2617	-0.6369	0.6564
β_{2001}	0.3157	0.2426	-0.0490	0.6461
β_{2002}	-0.8662	0.2847	-0.8955	0.6704
β_{2003}	-2.7019	0.4643	-2.9424	0.8631
β_{2004}	-0.0407	0.2470	-0.1555	0.6507
β_{2005}	-0.6701	0.2662	-0.9596	0.6745
α_0	1.6630	0.0100		

α_{years}	-0.0145	0.0010	
σ_u	0.2709	0.0091	0.3388
σ_a	0.6304	0.0003	

Based on the MCEM analysis that considers all 273 owls, the age at peak reproductive success is estimated to be 13.38 years. Based on the GLMM analysis using only the 108 known age owls, the age at peak reproductive success is estimated to be 8.97 years. Both analyses suggest a decline in reproductive success after peak reproductive success is reached (one sided p-value for β_{age^2} < 0.0001 for both analyses). Even though both of the analyzes suggest a decline after reaching peak reproductive success, Figure 2.3 shows that the analysis using all of the owls (heavier lines) is much flatter than the analysis that just uses the known age owls (lighter lines). The dashed lines show a 95% confidence bands for the mean fixed effect of male age. The heavier set of lines is for the analysis that considers all owls and the lighter set are for known age owls only. These confidence bands include between-owl variability in intercepts. The analysis that considers all owls has a much tighter confidence interval, especially around the peak of the known age curve. See Appendix A2 for plots and tables of estimation results for the other study areas.

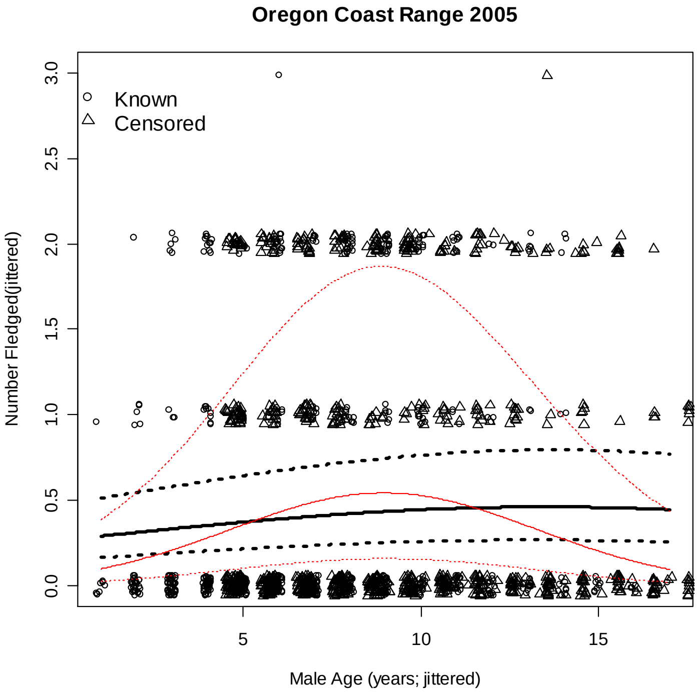


Figure 2.3 Number of owls fledged versus male parent age for the 108 known-aged owls, and GLMM model fits and approximate 95% confidence bands using only the known-age owls (thin line) and using all 273 owls (thick line) (Oregon Coast Range study area).

2.7 Discussion

The MCEM algorithm for censored covariates can suffer from slow convergence. In the models fit, the convergence time and number of iterations was a function of the number of owls and the percentage of censored age owls. The number of iterations required for convergence (with a relative convergence criterion that estimates change by less than 0.5% in successive iterations) was generally around 14. Since the algorithm uses pseudo-random variables, the convergence time and the number of iterations can vary between the same models fit with the same starting values. In addition the final estimates can vary slightly due to the convergence criteria being set so that the models do not take as long to converge.

One way to help the MCEM algorithm converge faster is to use “good” starting values. Possible choices include results from fitting only known age animals using an approximate technique for GLMMs (like lmer in R), replacing the censored ages with their conditional expected values (given that they exceed the recorded lower bound) and then fitting with lmer, or the final values from an approximate MCEM algorithm (detailed in the next chapter). The final values from the approximate MCEM algorithm are the best starting values, but require the most work to obtain. In our experience, the results from the model with the censored ages replaced by the expected value given that they exceed a lower bound are sufficient starting values that are relatively easy to obtain. Obtaining “good” starting values is also important because the MCEM algorithm is sensitive to the starting values. The standard deviation of the random effects distribution is the parameter most sensitive to the starting value.

There is evidence of a benefit in including those animals with censored ages. In the spotted owl data problem, the standard error for the coefficient of the quadratic age term, for example, was reduced by 91% over that from the known-age owls only. For other regression coefficients the percentage reduction was 66% and 82%. The simulation study in Chapter 3 shows that the MCEM parameter estimates can be biased. This bias reduces the curvature of the age-specific reproductive success curve. Therefore, in addition to the standard error reduction one needs to also consider the actual parameter estimates. Also note that in all four study areas, the MCEM curve is much flatter. This indicates that there may be issues with the method and not just a different curve for the censored ages.

2.8 References

Altwegg, R., Schaub, M., and Roulin, A.,(2007). Age-Specific Components of Temporal Variation in the Barn Owl. The American Naturalist. 169:47-61

Angelier, F., et al. (2007). Age-specific reproductive success in a long-lived bird: do older parents resist stress better?. *Journal of Animal Ecology*. 76:1181–1191

Austin, P. and Hoch J., 2004, "Estimating Linear Regression Models in the Presence of Censored Independent Variable," *Statistics in Medicine,* 23, 411-429

Bates, D.M., (2005), Fitting linear mixed models in R, *R News* **5** pp. 27–30.

Bates, D. M. (2007). *Linear mixed model implementation in lme4*. Manuscript, University of Wisconsin - Madison, January 2007.

Booth, J.G, and Hobert, J.P. (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM Algorithm *Journal of the Royal Statistical Society B.* 61, 265-285

Côté , S. D., and Festa-Bianchet, M., (2001) Offspring sex ratio in relation to maternal age and social rank in mountain goats (Oreamnos americanus). *Behavioral Ecology and Sociobiology*. 49: 260-265

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B.* 39, 1-38.

Green, D. J. (2001).The influence of age on reproductive performance in the Brown Thornbill, *Journal of Avian Biology*. 32: 6–14.

Horton, N.J. and Laird, N.M., 1998, "Maximum Likelihood Analysis of Generalized – Model with Missing Covariates," *Statistical Methods in Medical Research,* 8, 37-50

Jiming, Jiang (2007). *Linear and Generalized Linear Mixed Models and Their Applications.* Springer, New York.

Krüger, O., 2005. Age at first breeding and fitness in goshawk *Accipiter gentilis. Journal of Animal Ecology* 74:266–273

Levine, R.A., and Casella, G. (2001). Implementations of the Monte Carlo EM Algorithm. *Journal of Computational and Graphical Statsitics*. 10, 422-439.

Levine, R.A., and Fan, J. (2003). An Automated (Markov Chain) Monte Carlo EM Algorithm. *Journal of Statistical Computation and Simulation*. 74, 349-360.

Little, R., 1992, "Regression With Missing X's: A Review," *Journal of the American Statistical Association,* 87, 1227-1237

Loschl, P., (2008). Age-specific and Lifetime Reproductive Success of Known Age Northern Spotted Owls on Four Study Areas in Oregon and Washington., MS Thesis, Oregon State University

Louis, T.A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society. Series B.* 44 226-233.

McCulloch, C.E. (1997) Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association.* **92** 162-170.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models.* John Wiley and Sons, New York.

McLachlan, G.J, and Krishnan, T. (1997). *The EM Algorithm and Extensions.* John Wiley and Sons, New York.

Oakes, D. (1999). Direct Calculation of the Information Matrix via the EM Algorithm. *Journal of the Royal Statistical Society. Series B.* 61 479-482.

Newton, I., and Rothery, P., (2002) . Age-Related Trends in Different Aspects of the Breeding Performance of Individual Female Eurasian Sparrowhawks *The Auk* 119(3):735–748

Schabenberger, O. (2005). Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models, *Proceedings of the Thirtieth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc. Paper 196-30.

Schabenberger, O. (2007). Growing Up Fast: SAS 9.2 Enhancements to the GLIMMIX Procedure SAS. *Proceedings of the SAS Global Forum 2007 Conference. Cary, NC: SAS Institute Inc.*. Paper 177-2007.

Wu, K., and Wu, L. (2007) Generalized linear mixed models with informative dropouts and missing covariates. *Metrika.* 66, 1-18

Zedrosser, A. et al. (2007). Genetic estimates of annual reproductive success in male brown bears: the effects of body size, age, internal relatedness and population density. *Journal of Animal Ecology* 76: 368–375.

3. Computationally Practical Analysis of Generalized Linear Mixed Models with Censored Covariates

John Giovanini[*] and Daniel W. Schafer[**]
Department of Statistics,
Oregon State University,
Corvallis, OR 97331
[*]email: giovanin@science.oregonstate.edu
[**]email: schafer@science.oregonstate.edu

3.1 Abstract

This paper is about statistical methods for data analysis using generalized linear mixed models (GLMMs) with censored covariates. Special attention in given to the particular problem of inference about age-specific reproductive success in wild animal populations using some animals with known ages and some animals with ages only known to exceed some lower bound. GLMMs allow for non-normal response distributions, such as a Poisson distribution for the number of offspring from a parent in one year, and they account for the correlation of repeated responses from the same observational unit, such as the correlation of the number of offspring from the same parent over multiple years. Maximum likelihood analysis via the EM algorithm and Markov Chain Monte Carlo techniques was proposed in Chapter 2. This approach would not be attractive for immediate use by wildlife biologists, we suspect, because of the non-trivial programming required, the slowness, and the lack of transparency. We propose two other methods, which also incorporate animals with unknown ages but which are easier to compute. First, a Monte Carlo EM algorithm is proposed in which censored covariates are treated as "missing data" and a Laplace approximation

is made to the complete data log likelihood. The Laplace approximation permits the use of existing computing modules for GLMMs.   This results in simpler programming than the likelihood analysis proposed in Chapter 2 but falls short of being simple enough for immediate use by most wildlife biologists. Second, a simple “regression calibration” method is proposed, which uses existing software for GLMMs but simply replaces missing ages by estimated expectations from a regression model and conditional on the age exceeding the censoring value. A simulation study clarifies the degree to which these methods improve upon the common approach of ignoring animals with unknown ages and clarifies their operating characteristics relative to those of the maximum likelihood estimator.

3.2 Introduction

This work was motivated by an investigation of factors affecting individual reproductive success in a wild population of Northern Spotted Owls (*Strix occidentalis caurina*). For example, the scatterplot in Figure 3.1 shows the number of fledged spotted owls in a year versus male parent age in that year, for multi-year observations on 108 male Northern Spotted Owls. The solid points are the sample means for each age and the vertical lines are crude 95% confidence intervals for each age group (data from Pete Loschl, personal communication; see Loschl, 2008). The smooth curve is a nonparametric lowess fit, which shows an apparent increase in mean number fledged up to a maximum of 0.7 fledglings per year at about age 9, with a subsequent decrease.

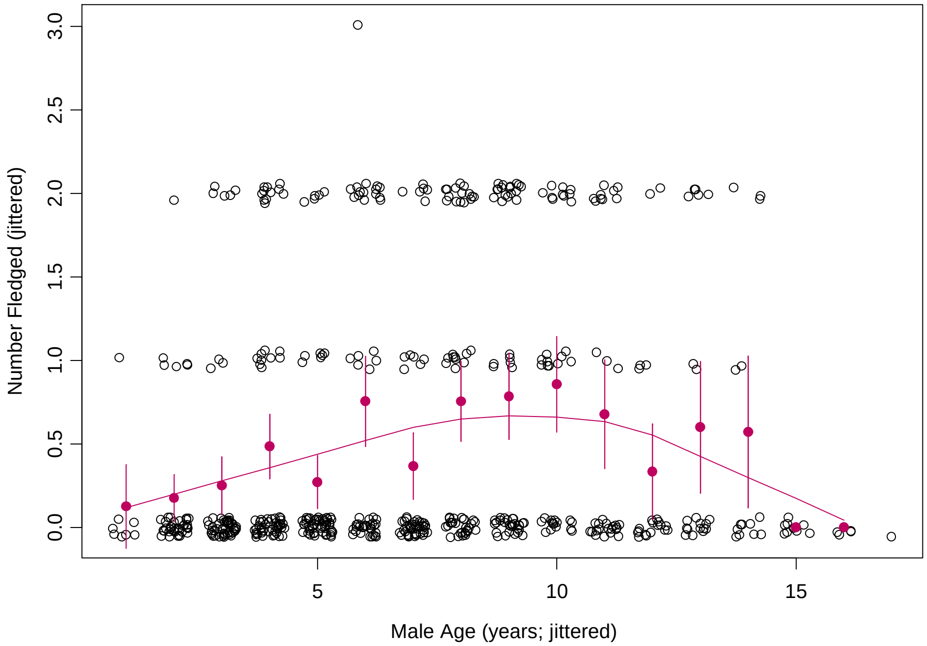


Figure 3.1 Number of fledglings versus male parent age from multiple-year observations on each of 108 Northern Spotted Owl males in the Oregon Coast Range, and lowess curve.

More formal investigation into patterns of age-specific reproductive success should account for dependence of different observations from the same male. The scatterplot below reproduces Figure 3.1 but includes the fit to a generalized linear mixed model (GLMM) that specifies a Poisson distribution for the integer count response and a regression model in which the log of the mean count is a quadratic function of age, and which accounts for the dependence of observations from the same male via random intercepts for the 108 different males. The solid curve is the approximate maximum likelihood estimate of this GLMM. The dashed lines show a 95% confidence band for the mean fixed effect of male age. (This confidence band includes between-owl variability in intercepts.)

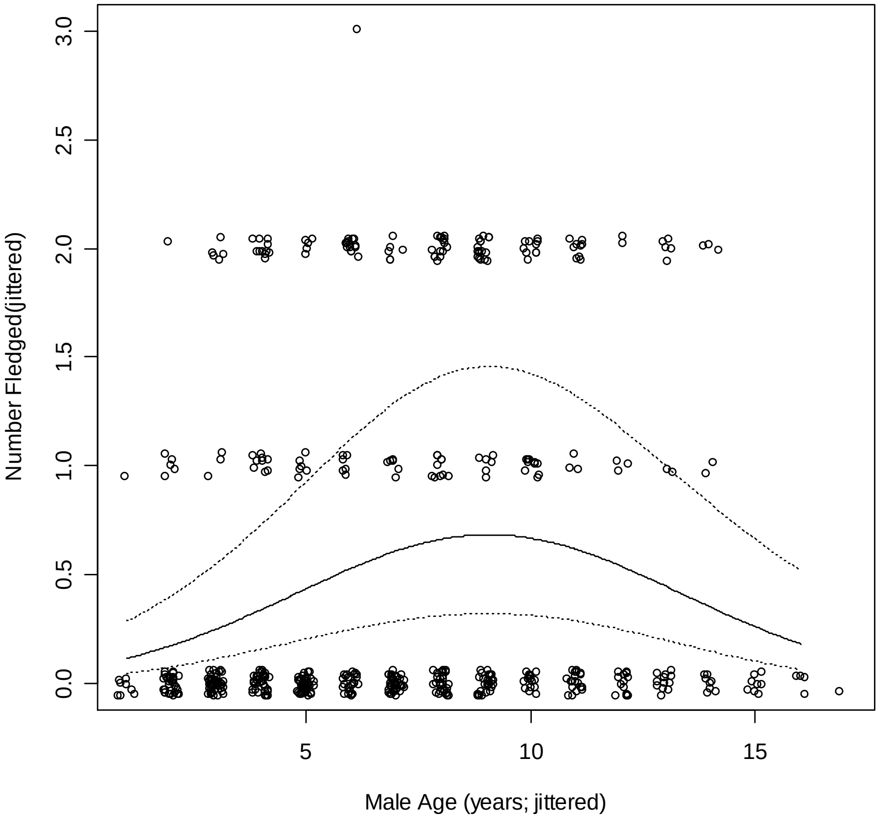


Figure 3.2 Number of fledglings versus male age as in Figure 3.1, with GLMM fit and approximate 95% confidence band.

This type of model is useful for investigating several scientific questions about age-specific reproductive success: 1.What evidence is there that male reproductive success decreases in older ages? 2. What is the age at which maximal mean reproductive success is achieved? 3. What proportion of variability in reproductive success can be explained by between-male differences after accounting for effects of age? 4. What evidence is there that various landscape and climate variables affect mean number fledged, after accounting for the effects of parent age?

When number of fledglings or some other small integer count is used as a measure of reproductive success, these questions (and similar questions for females) can be addressed with GLMM analysis (McCulloch & Searle, 2001, and Jiming, 2007). The measure of reproductive success is taken to have a Poisson (or possibly binomial) distribution with a mean that depends on parent age and other explanatory variables, but with the inclusion of random effects to account for variable reproductive successes between males (or females). Standard GLMM modules also allow for Poisson-like and binomial-like distributions with an additional dispersion parameter.

Because GLMM modules have been added to standard statistical software packages only recently, some wildlife biologists have relied on ordinary mixed linear model analysis (Loschl, 2008, Coltman et al, 2002, and Reid et al 2003). Some have used GLMMs though. This has been the case for age-reproductive success studies of barn owls (Altwegg et al. 2007), snow petrels (Angelier et al. 2007), brown thornbills (Green 2001), goshawks (Kruger 2005), brown bears (Zedrosser, et al. 2007), mountain goats (Côté, Festa-Bianchet 2001), and sparrowhawks (Newton and Rothery 2002).

Our interest is in the use of GLMMS for this purpose when a substantial number of the animals in the data set have ages that are only known to exceed some lower bound. The plots and fitted models in Figures 3.1 and 3.2, for example, are based on 542 observations from 108 male owls whose ages were known exactly. Also available are 839 observations from 165 additional male owls whose exact ages on their first season of observation were only known to exceed 3 years. (They were known to exceed the age of three because of their adult characteristics). In the year after their first observation, therefore, the researchers could be sure these owls were at least 4 years old. After 10 years of observation, they were at least 13 years old, and so on.

It is common for researchers to exclude the animals with unknown ages from the statistical analysis of age-specific reproductive success. Examples include Loschl, 2008 (spotted owls); Angelier et al, 2007 (snow petrels); Altwegg et al., 2007 (barn owls); Kruger 2005 (goshawks); and Newton and Rothery, 2002 (sparrowhawks). While the common practice of excluding animals of unknown ages isn’t likely to induce any bias into the scientific conclusions, the incorporation of information from the owls with censored ages may provide important gains in efficiency and power. One does need to consider the possible that the age-specific reproductive success curves are different for the known age vs. the censored animals. This is especially true if the animals that are excluded from the analysis are the older animals that were first observed as adults at the beginning of a study.

Notice, for example, that there is some visual indication from Figure 3.2 that the mean number of fledglings decreases with older ages. The fairly wide confidence band at that end of the graph, though, suggests that the evidence for the decrease is not convincing. Including the additional owls with censored ages may result in more precise model estimation and therefore more resolution to this and other scientific questions of interest.

Our goals are to clarify the potential efficiency gains and to provide techniques for incorporating animals with censored ages into the GLMM analysis for age-specific reproductive success. We are particularly interested in finding easy-to-use solutions, if possible. Our statistical interest is in the Spotted Owl study in particular, but the same data characteristics are found in other studies of age-specific reproductive success. We can imagine that there are other applications of GLMMs in which a covariate is censored and our results would pertain to those. Austin and Hoch (2004) report a regression problem, for example, in which the explanatory variable “household income” is obtained from a survey questionnaire with highest category “greater than $80,000.” The main effort here, though, is directed towards the particular problems encountered in the study of age-specific reproductive success from wild animal populations.

A full maximum likelihood estimator was proposed in Chapter 2. In this paper, we pursue simpler methods that wildlife biologists could use immediately with minimal extra programming. We first propose an MCEM algorithm that uses a Laplace approximation in order to make use of existing software modules. While simpler than the full likelihood approach, we would not characterize the necessary computations as “simple.” A “regression calibration” approach is also proposed, which makes use of an existing module for GLMM analysis but with unknown ages replaced by predicted values.

This paper proceeds as follows. Section 3.3 describes the model. Section 3.4 describes the approximate maximum likelihood methods for censored covariates in GLMMs. Section 3.5 describes a regression calibration estimator for censored covariates. Finally, in Section 3.6 a simulation study is described to investigate the properties of several methods for fitting GLMMs with censored covariates.

3.3 Model

We consider a model that specifies repeated measures on each of *m* subjects (or clusters) with responses that follow a generalized linear model with random intercepts for each subject, with time-dependent and time-independent explanatory variables, and with a time-independent explanatory variable that is censored. Let y_i represent the response observed for observational unit *i*, for *i = 1,…,n*. In the spotted owl example, the response is the number of young fledged and the “observational unit” is an “owl year” of observation.

Let $\underline{z}_i = (z_{i1},..., z_{im})$ where z_{ij} = 1 if observational unit *i* is associated with subject or cluster *j* and 0 if not; for *i* = 1,…,*n* and *j* = 1,…,*m*. In the example, this variable indicates the particular male associated with observational unit *i*.

Let a_j represent the explanatory variable that is censored on some of the subjects or clusters, for subject or cluster *j*, for *j* = 1,…, *m*, and let $\underline{a} = (a_1,...,a_m)$. In the example, a_j is the age of male owl *j* at the time it was first observed.

Let c_j be a censoring indicator, that takes on the value 0 if a_j is observed and takes on the value 1 if it is only known that a_j is greater than or equal to some known value. Let a_j^* be the true age at first observation, a_j, for those owls with known ages and the lower bound for age at first observation otherwise. Let $\underline{a}^* = (a_1^*,...,a_m^*)$. Let $\underline{x}_i$ be a vector of explanatory variables associated with observational unit *i*, which may be time-variant or time- invariant. Let *X* be the matrix whose ith row is $\underline{x}_i^T$. Let $\underline{u} = (u_1,...,u_m)$ represent "random effects" associated with the *M* clusters or subjects.

We suppose that the $y_i\,'s$ are conditionally independent, given $\underline{u}$, with probability density or mass function

$$f(y_i \mid \underline{a}, \underline{u}, \underline{x}_i) = f(y_i \mid \underline{z}_i^T \underline{a}, \underline{z}_i^T \underline{u}, \underline{x}_i) \text{ with mean } \mu_i \text{, where}$$

$$g(\mu_i) = \underline{x}_i^T \beta + h(\underline{z}_i^T \underline{a}; \alpha) + \underline{z}_i^T \underline{u}\ ,$$

where α and β are *p*- and *q*-vectors of unknown parameters and *g*( ) is a known "link function." In the owl example, the response distribution is taken to be Poisson and the link function is the logarithm. A dispersion parameter can be added since we will be using GLMM modules that allow a dispersion parameter.

The term $h(\underline{z}_i^T \underline{a}; \alpha)$ is of unspecified form to permit the incorporation of nonlinear effects of the censored explanatory variable, such as a quadratic effect of age for the model displayed in Figure 2.2. The need for this in the owl example is evident in the figures above. Note that if s_i represents the number of years since the male associated with observational unit *i* was first observed, then $\underline{z}_i^T \underline{a} + s_i$ is its current age associated with observational unit *i*. One possible model, for example, is

$$h(\underline{z}_i^T \underline{a} + s_i ; \alpha) = \alpha_1 (\underline{z}_i^T \underline{a} + s_i) + \alpha_2 (\underline{z}_i^T \underline{a} + s_i)^2 .$$

Suppose also that the random effects are independent and identically distributed, and independent of explanatory variables:

$$u_j \sim f(u_i \mid \underline{a}, X; \tau) = f(u_i ; \tau) .$$

In GLMMs, it is convenient to take this distribution to be normal with mean 0 and variance τ .

It is also necessary to assume some distributional model for the marginal distribution of *a*. In the owl example, there is good reason to believe that those animals that are censored have larger values of *a* then those that aren't—because most of the latter, presumably, were observed in their first year of life and then included in the study. Let

$$a_j \sim f(a_j \mid \underline{w}_j ; \gamma) ,$$

where $\underline{w}_j$ is a vector of explanatory variables that would be useful for predicting a_j . In the owl example, the total number of years that the owl was observed would be such an explanatory variable. We assume that a_j is independent of u_j and of $a_{j'}$ for $j \neq j'$.

3.4 Approximate Maximum Likelihood for GLMMs with Censored Covariates

Even without censored explanatory variables, the likelihood for all but the simplest GLMMs involves a multi-dimensional integral over the random effects. This integral can be high dimensional with no closed-form solution. Several of the currently popular methods for GLMM estimation involve approximations to the integral or other similar modifications to the likelihood.

There are several such methods that are justified differently but which use essentially the same algorithms. One method, by Schall (1991), applies the link function to the response, linearizes the regression using a first order Taylor's approximation, and then repeatedly fits linear mixed models to the working dependent variables. Breslow and Clayton (1993) included a penalty to the quasi-likelihood function to derive the penalized quasi-likelihood (PQL). Wolfinger (1993) showed how the Laplace approximation of the log-likelihood in the GLMM can be used to find estimates. All three of these methods lead to the same computational algorithm that repeatedly fits linear mixed models to working dependent variables. A good general reference on all three methods is McCulloch & Searle (2001). For the problem of this paper, we wish to apply some of the same approximation techniques in the hopes of leading to an approximate maximum likelihood analysis for GLMMs when there are censored explanatory variables. Although there will still be an additional part of the likelihood that involves the censored explanatory variables, the approximation will permit the use of currently available GLMM computing modules as part of the likelihood analysis. In particular, we can make use of existing modules for GLMM estimation by using a Laplace approximation in the “complete data” likelihood specification for using the EM Algorithm (Dempster, Laird, & Rubin 1977 and McLachlan & Krishnan 1997).

The “observed data” are $\underline{y}$ and $\underline{a}^*$. The “complete data” are taken to be $\underline{y}$ and $\underline{a}$. This differs from the setup of Chapter 2, in which the random effects, $\underline{u}$, were also specified as part of the complete data. Here, the complete data likelihood is based on the marginal distribution of $\underline{y}$, obtained from the specified model by integrating out $\underline{u}$. The complete data log likelihood is:

$$l_c(\theta; \underline{y}, \underline{a}) = \log\left[\int f(\underline{y}, \underline{u}, \underline{a} \mid X; \theta) d\underline{u}\right]$$

$$= \log\left[\int f(\underline{y} \mid \underline{u}, \underline{a}, X; \theta) f(\underline{u} \mid \underline{a}, X; \theta) d\underline{u}\right] + \log\left[f(\underline{a} \mid X; \theta)\right]$$

$$= \log\left[\int \prod_{i=1}^{n} f(y_i \mid \underline{z}_i^T \underline{u}, \underline{z}_i^T \underline{a}, x_i; \alpha, \beta, \tau) \prod_{j=1}^{m} f(u_j; \tau) d\underline{u}\right] + \sum_{j=1}^{m} \log[f(a_j \mid \underline{w}_j; \gamma)]$$

The Laplace approximation is applied to the integral in the first term. We may write the resulting approximate complete data log likelihood as

$$\tilde{l}_c(\theta; \underline{y}, \underline{a}) = l_{glmm}(\alpha, \beta, \tau; \underline{y}, \underline{a}) + \sum_{j=1}^{m} \log[f(a_j \mid \underline{w}_j; \gamma)],$$

where $l_{glmm}(\alpha, \beta, \tau; \underline{y}, \underline{a})$ is the approximate log likelihood that is maximized by the Wolfinger approach if all the ages in a were available. We do not need to specify this approximation in more detail; for our purposes, it is enough to know that routines to maximize it are available. We will make use of those routines as part of the M-step in an EM algorithm that treats the unknown ages as missing data.

The E-step (expectation) requires the expectation of the complete data log likelihood given the observed data and with unknown parameters in the expectation replaced by their estimates after *t* iterations. Let $\theta^{(t)}$ denote the estimate of θ after *t* iterations of the EM algorithm. Then the expectation is:

$$Q(\theta \mid \theta^{(t)}) = E\left\{ l_{glmm}(\alpha, \beta, \tau; \underline{y}, \underline{a}) \mid \underline{y}, X, \underline{a}^{*}\ ; \theta^{(t)} \right\} +$$

$$\sum_{j=1}^{m} E\left\{ \log[f(a_j \mid \underline{w}_j; \gamma)] \mid \underline{y}, X, \underline{a}^{*}\ ; \theta^{(t)} \right\}$$

The expectations are with respect to the distribution of $\underline{a}$ given $\underline{y}$, $\underline{a}^{*}$, and *X*. In general, the expectations are intractable, but they can be approximated by Monte Carlo methods (McCulloch and Searle, 2001, Sect. 10.3):

$$Q(\theta \mid \theta^{(t)}) \approx \sum_{r=1}^{R} \frac{1}{R}\, l_{glmm}(\alpha, \beta, \tau; \underline{y}, \underline{a}^{(r)}) + \sum_{j=1}^{m} \sum_{r=1}^{R} \frac{1}{R} \log[f(a_j^{(r)} \mid \underline{w}_j; \gamma)]$$

where $\underline{a}^{(r)} = (a_1, ..., a_{m_c}, a_{m_c+1}^{(r)} ..., a_m^{(r)})$, where the first m_c elements are known ages of first observation and where the remaining elements, $a_j^{(r)}$, are pseudo-random variables generated from $f(a_j \mid \underline{w}_j, \underline{\tilde{y}}, X, a_j^{*}; \theta^{(t)})$.

The EM algorithm is an iterative algorithm that, at each iteration, updates the expectations $Q(\theta \mid \theta^{(t)})$ based on current parameter estimates and then computes updated estimates as those values that maximize $Q(\theta \mid \theta^{(t)})$. The following steps describe the algorithm:

1. Choose starting values, $\theta^{(0)} = (\alpha^{(0)}, \beta^{(0)}, \gamma^{(0)}, \tau^{(0)})$. Set *t* = 1.
2. Repeat until convergence:

a. Calculate the adjusted response $\underline{\tilde{y}} = g(\mu) + (\underline{y} - \mu)g'(\mu)$ as in Schall (1991). This is needed to generate the pseudo random variables in the Metropolis step

b. Generate *R* vectors $\underline{a}^{(r)}$ from the conditional distribution of $\underline{a}$ given $\underline{\tilde{y}}$, *X*, and $\underline{a}^{*}$ with unknown parameters θ in the distribution replaced by the "current" estimates $\theta^{(t-1)}$

c. Calculate $\alpha^{(t)}$, $\beta^{(t)}$, and $\tau^{(t)}$ as those values that maximize

$$\sum_{r=1}^{R} \frac{1}{R}\, l_{glmm}(\alpha, \beta, \tau; \underline{y}, \underline{a}^{(r)})$$

d. Calculate $\gamma^{(t)}$ as the value that maximizes

$$\sum_{j=1}^{m}\sum_{r=1}^{R} \frac{1}{R} \log[f(a_j^{(r)} \mid \underline{w}_j; \gamma)]$$

e. Set $R = R + \lfloor R/c \rfloor$, for some $c > 0$

f. Set $t = t+1$.

Each of the pieces (c) and (d) can be accomplished with weighted formulas or routines that would be appropriate in the absence of censored explanatory variables, but based on the augmented data set corresponding to the *R* pseudo values for $\underline{a}$. In particular, piece (c) can calculated with a standard routine for GLMMs that uses the Laplace or penalized quasi-likelihood approach (such as lmer in R).

The random number generation in 2b can be accomplished with the Metropolis-Hastings algorithm as follows (see, for example, McCulloch and Searle, 2001 Sect. 10.3):

1. Generate $\underline{a}^{(0)} = (a^{(0)}_{m_c+1} \ldots, a^{(0)}_{m})$ with $a^{(0)}_{m}$ generated from $f(a_j \mid \underline{w}_j, a^*_j; \gamma^{(t-1)})$. Set *r* = 1.
2. For *k* from 1 to (*R* + *R**):
   a. Generate $\underline{\tilde{a}} = (\tilde{a}_{m_c+1} \ldots, \tilde{a}_{m})$ with $a^{(k)}_{m}$ element generated from $f(a_j \mid \underline{w}_j, a^*_j; \gamma^{(t-1)})$
   b. Compute the acceptance criterion:
   $$p_k = \min\left\{1, \frac{f(\underline{\tilde{y}} \mid \underline{u}, \underline{\tilde{a}}, X; \alpha^{(t-1)}, \beta^{(t-1)})}{f(\underline{\tilde{y}} \mid \underline{u}, \underline{a}^{(k-1)}, X; \alpha^{(t-1)}, \beta^{(t-1)})}\right\}$$
   c. Generate *v*, a Bernoulli(p_k) random variable:
   d. If *v* =1 set $\underline{a}^{(k)} = \underline{\tilde{a}}$. Otherwise, set $\underline{a}^{(k)} = \underline{a}^{(k-1)}$
3. Retain the final *R* of each of these vectors as the simulated sample. (*R** is the burn-in number.)

To speed up convergence several authors, (Levine and Fan 2003) and (Levine and Casella 2001), recommend using importance weights instead of drawing a new MCMC sample at each iteration. The use of importance weights can greatly decrease the convergence time because generating the pseudo-random variables via the Metropolis Algorithm is computationally more intensive than generating the importance weights. It is recommended that a burn-in period of regular Monte Carlo EM iterations is used before switching to the importance weights. The burn-in allows the target and the candidate distribution to be ‘closer’ and therefore helps decrease the convergence time.

In the above algorithm, we increase the Monte Carlo sample size *R* using Booth and Hobert’s (1999) recommendation of $R = R + \lfloor R/c \rfloor$, for some $c > 0$. This method is used because at early iterations, when the “current” parameter estimates are likely far from the MLE, one does not need a large Monte Carlo sample size. However, as the “current” parameter estimates get closer to the MLE, one needs more precision and therefore a larger Monte Carlo sample size. Instead of using a naïve increase of the Monte Carlo sample size, (Levine and Fan 2003) and (Levine and Casella 2001) suggest automated algorithms that increase the Monte Carlo sample size after checking if the Monte Carlo error overwhelmed the EM estimate. For our specialized algorithm, we simply used the naïve increase.

Approximate standard errors can be calculated using McLachlan and Krishnan’s (1997) method. This method uses only first-order derivatives to find the approximate information matrix:

$$I\left(\hat{\theta}\right) \approx \sum_{j=1}^{m}\sum_{r=1}^{R}\frac{1}{R}S_{jr}\left(\hat{\theta}\right)S_{jr}^{T}\left(\hat{\theta}\right), \text{ where } S_{jr}\left(\hat{\theta}\right) = \left.\frac{\partial l_c^{(r)}\left(\theta; \underline{y}_j, \underline{x}_j,\ u_j^{(r)}, a_j^{(r)}\right)}{\partial\theta}\right|_{\theta=\hat{\theta}}$$

The standard errors for the MLE can then be calculated by taking the square root of the diagonal elements of $I^{-1}\left(\hat{\theta}\right)$.

There are several other methods for calculating the information matrix, including calculating the observed Fisher Information from the observed data log-likelihood and Oakes (1999) & Louis’ (1982) method of writing the observed data log-likelihood as functions of the complete data log-likelihood. Both of these methods would be rather complicated in our situation.

3.5 A Regression Calibration Estimator

While the approximate maximum likelihood estimator of Section 3.4 avoids some of the complexity involved in the full maximum likelihood estimator, the need for the Metropolis-Hastings algorithm in the E-step voids any notion that the approach is simple. A second alternative, which is much more transparent than either of the maximum likelihood solutions, is a regression calibration estimator in which the unknown ages are replaced by predicted values.

Regression calibration (Carroll, Ruppert, and Stefanski, 1995) is an approach usually associated with regression estimation in the presence of imprecisely measured explanatory variables. The idea is popular because it can be used in many different kinds of regression models and because it is particularly transparent. The idea is to use the regression techniques that would have been appropriate if the explanatory variables were available, but to replace the missing values by their expectations given the observed measurement and the remaining explanatory variables. The details and the performance of the regression calibration method differ depending on the degree of nonlinearity of the regression with respect to the mismeasured explanatory variable and the way in which the expectations are estimated (see, for example, Schafer and Gilbert, 2006).

While the problem of interest in this paper is not thought of as a problem of measurement errors in explanatory variables, it can be cast that way. The measurement of the explanatory variable *age,* if unknown, is *measured* by the lower bound for age. The regression calibration estimator replaces the unknown ages by their expectations given this measurement and the other available explanatory variables (including all those variables that would be useful for predicting age).

In the measurement error terminology, the particular form of the regression calibration estimator for the spotted owl data problem would be described as regression calibration with internal validation, meaning that the data for estimating the expectation of the unknown ages given the other variables is a subset of the data with which the regression of interest will be estimated. This requires a two-stage process: (1) Using the owls with known ages, fit a fully parametric regression model for predicting age at first observation from other available explanatory variables. (By fully parametric, we mean in particular that conditional distribution of age, or some transformation of age, given the other explanatory variables is normal so that the conditional expectation given some lower bound can be deduced.) Using these results, find predicted ages at first observation for those owls whose exact ages are unknown. (2) Fit the regression of interest using all owls and replacing unknown ages at first observation with these predicted values.

It should be noted that the predicted ages in step (2) are themselves imprecise measurements of the explanatory variable of interest, so that the problem of imprecisely measured explanatory variables is still present. The predominant form of the imprecision, though, follows the Berkson error model (see Carroll, Ruppert, and Stefanski, 1995), which does not induce bias in the same way that the classical measurement error model does. The effect of sampling error in the estimation of the inserted predicted values in regression calibration has been examined by Monleon (2006) and Schafer and Gilbert (2006). The effects differ depending on several conditions of the particular data problem. Here, we use a simulation study to investigate the potential bias and other operating characteristics of the regression calibration estimator.

The following are the specific steps for regression calibration for estimating GLMMs with censored explanatory variables:

1. Find the expected value of the censored observations given that they exceed a lower bound

    a. Calculate $\hat{\gamma}$ and $\hat{\sigma}_a$ that maximizes

    $$\sum_{j=1}^{m_c} \log[f(\log(a_j) \mid \underline{w}_j; \gamma, \sigma_a)]$$

    b. Using the estimates from (a), calculate the predicted value of the censored ages at first observed breeding using the formula for the mean of a truncated normal distribution.

    $$\hat{\mu}_{a_j} = \hat{\gamma}_0 + \hat{\gamma}_1 years_j$$

    $$\tilde{a}_j = \exp\left[\frac{\mu_{a_j} + \hat{\sigma}_a \phi\left(\frac{\log(a_j^*) - \hat{\mu}_{a_j}}{\hat{\sigma}_a}\right)}{1 - \Phi\left(\frac{\log(a_j^*) - \hat{\mu}_{a_j}}{\hat{\sigma}_a}\right)}\right]$$

    c. Create the new age at first observed nesting vector

    $$\underline{\tilde{a}} = \left(a_1, \ldots, a_{m_c}, \tilde{a}_{m_c+1}, \ldots, \tilde{a}_m\right)$$

2. Fit the GLMM with the new age at first observation vector $\underline{\tilde{a}}$

a. Calculate $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\tau}$ as those values that maximize

$l_{glmm}(\alpha, \beta, \tau; \underline{y}, \underline{\tilde{a}})$

using with a standard routine for GLMMs that uses the Laplace or penalized quasi-likelihood approach (such as lmer in R).

The regression calibration method also allows the inclusion of multiple random effects. In the Northern Spotted Owl study, one might include the year as another random effect. This would be extremely difficult to do in the MCEM methods. The regression calibration method also allows one to easily fit a dispersion parameter.

Finally, in *linear* regression when there is no sampling variability in the expected values used for replacing the imprecisely measured explanatory variables, it is appropriate to use the usual inference procedures that would have been used if the actual explanatory variables were available. In particular, inference based on approximate normality of estimators and with reported standard errors is justified. For nonlinear regression models and using estimates to estimate the conditional expectations, the reported standard errors are too small. Sandwich formulas for adjusting approximate standard errors are available for some models (see Carroll, Rupert, and Stefanski, 1995), but we have not pursued those here. The simulation study that follows shows that the desirably simple procedure, including the use of the reported standard errors, is likely to be a very satisfactory approach.

3.6 Simulation Study

A simulation study was used to examine the relative operating characteristics of these approaches: (1) maximum likelihood estimator via the MCEM algorithm for censored covariates, (2) the approximate MCEM estimator, (3) the regression calibration estimator, (4) the naïve estimator in which censored ages are replaced by their lower bounds (to demonstrate the unsuitability of this approach, which may seem tempting to wildlife biologists), and (5) the GLMM estimator using only owls with known ages.

The conditions for the simulation study were based on estimated parameters from the Northern Spotted Owl study. In particular, a quadratic-in-age log-linear model was used as the mean of a Poisson response. Rather than specifying a distribution of ages at first observation, we randomly sampled owls with known ages from the Spotted Owl study (with replacement) and used their ages at first observation in the simulation. We randomly selected a subset of simulated subjects (owls), of a specified percentage, to have known ages and the rest to have censored ages. Those simulated subjects that were deemed to have censored ages were taken to be censored at age 3 (because that was the predominant lower bound for owl ages in the data set). We used the age parameters and the random effects distribution roughly matching those from the real data set to generated linear predictors and then number fledged. We investigated several sample sizes (total number of owls) and several values for the percentage of subjects with censored ages. The simulation conditions are further detailed in Appendix A5.

Table 3.1 shows descriptive statistics of estimates of estimates of β_2 from the 6 settings and 5 fitting methods. The statistics for each condition are based on 200 Monte Carlo samples. See the Appendix for similar tables for estimators of β_0 and β_1.

Table 3.1 Descriptive statistics of estimates of β_2 (true value -0.0235)

Sample size	50			400		
Proportion censored	.10	.25	.50	.10	.25	.50
MCEM						
mean	-0.0246	-0.0227	-0.0214	-0.0225	-0.0216	-0.0199
bias	-0.0011	0.0008	0.0021	0.0010	0.0019	0.0036
variance	7.55E-05	0.0001	5.76E-05	7.43E-06	7.58E-06	5.90E-06
MSE	7.64E-05	0.0001	6.16E-05	8.43E-06	1.12E-05	1.87E-05
Monte Carlo SD	0.0087	0.0082	0.0076	0.0027	0.0028	0.0024
Mean reported SE	0.0092	0.0086	0.0083	0.0026	0.0026	0.0025
Approx MCEM						
mean	-0.0245	-0.0223	-0.0207	-0.0228	-0.0217	-0.0200
bias	-0.0010	0.0013	0.0028	0.0007	0.0018	0.0035
variance	7.22E-05	0.0001	5.13E-05	7.52E-06	7.17E-06	5.37E-06
MSE	7.29E-05	0.0001	5.89E-05	8.02E-06	1.03E-05	1.78E-05
Monte Carlo SD	0.0085	0.0080	0.0072	0.0027	0.0027	0.0023
Mean reported SE	0.0044	0.0041	0.0038	0.0010	0.0010	0.0010
Regression Calibration						
mean	-0.0251	-0.0240	-0.0241	-0.0232	-0.0232	-0.0228
bias	-0.0015	-0.0005	-0.0006	0.0003	0.0004	0.0007
variance	7.93E-05	7.69E-05	6.91E-05	7.84E-06	8.44E-06	7.27E-06
MSE	8.14E-05	7.67E-05	6.91E-05	7.91E-06	8.54E-06	7.70E-06
Monte Carlo SD	0.0089	0.0088	0.0083	0.0028	0.0029	0.0027
Mean reported SE	0.0087	0.0086	0.0090	0.0029	0.0029	0.0030
Naïve Replace						
mean	-0.0235	-0.0212	-0.0207	-0.0221	-0.0204	-0.0193
bias	-2.30E-05	0.0023	0.0028	0.0015	0.0031	0.0043
variance	7.06E-05	6.37E-05	6.97E-05	7.05E-06	7.45E-06	5.76E-06
MSE	7.03E-05	6.88E-05	7.70E-05	9.15E-06	1.72E-05	2.38E-05
Monte Carlo SD	0.0084	0.0080	0.0083	0.0027	0.0027	0.0024
Mean reported SE	0.0084	0.0081	0.0081	0.0028	0.0027	0.0027
Known Only						
mean	-0.0259	-0.0244	-0.0266	-0.0237	-0.0238	-0.0240
bias	-0.0024	-0.0009	-0.0031	-0.0002	-0.0003	-0.0005
variance	8.48E-05	0.0001	0.0002	8.81E-06	1.04E-05	1.74E-05
MSE	9.02E-05	0.0001	0.0002	8.83E-06	1.04E-05	1.76E-05
Monte Carlo SD	0.0092	0.0102	0.0145	0.0030	0.0032	0.0042
Mean reported SE	0.0092	0.0100	0.0129	0.0030	0.0034	0.0041

Some of the features in Table 3.1 will be highlighted and clarified for further emphasis. Figures 3.3 and 3.4 show the Monte Carlo sampling distributions for the estimators of β_2 and β_0 respectively for the settings with a sample size of 400 and with 50% of the observations censored. The vertical line represents the true value of parameter.

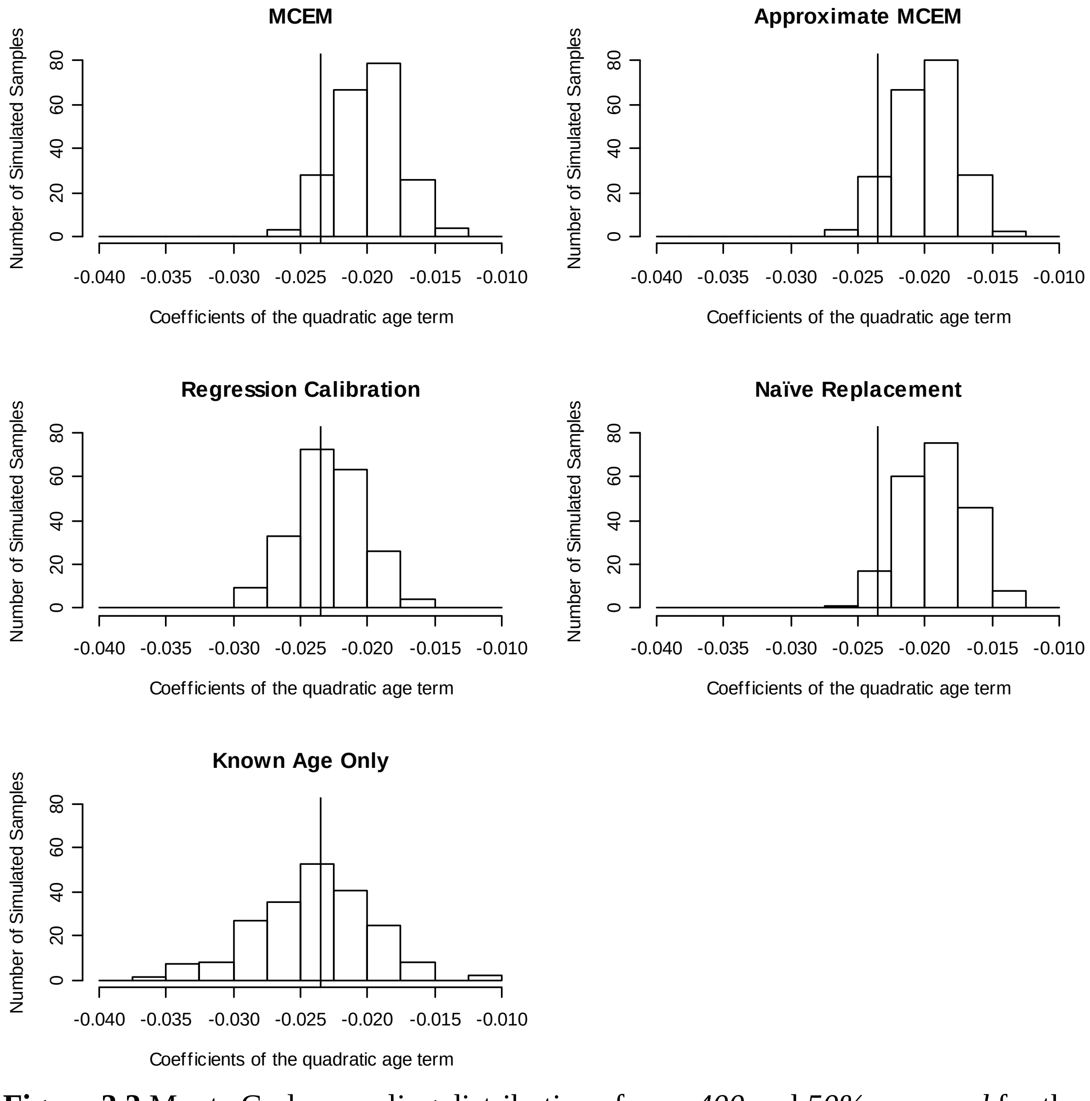


Figure 3.3 Monte Carlo sampling distributions for *n=400* and *50% censored* for the estimators of β_2, the age^2 term, for the 5 different fitting methods.

In Figure 3.3 the MCEM, approximate MCEM, and the naïve replacement are all biased. The known age only and the regression calibration estimates appear unbiased. Notice how the regression calibration estimates are much less variable than the known age only estimates.

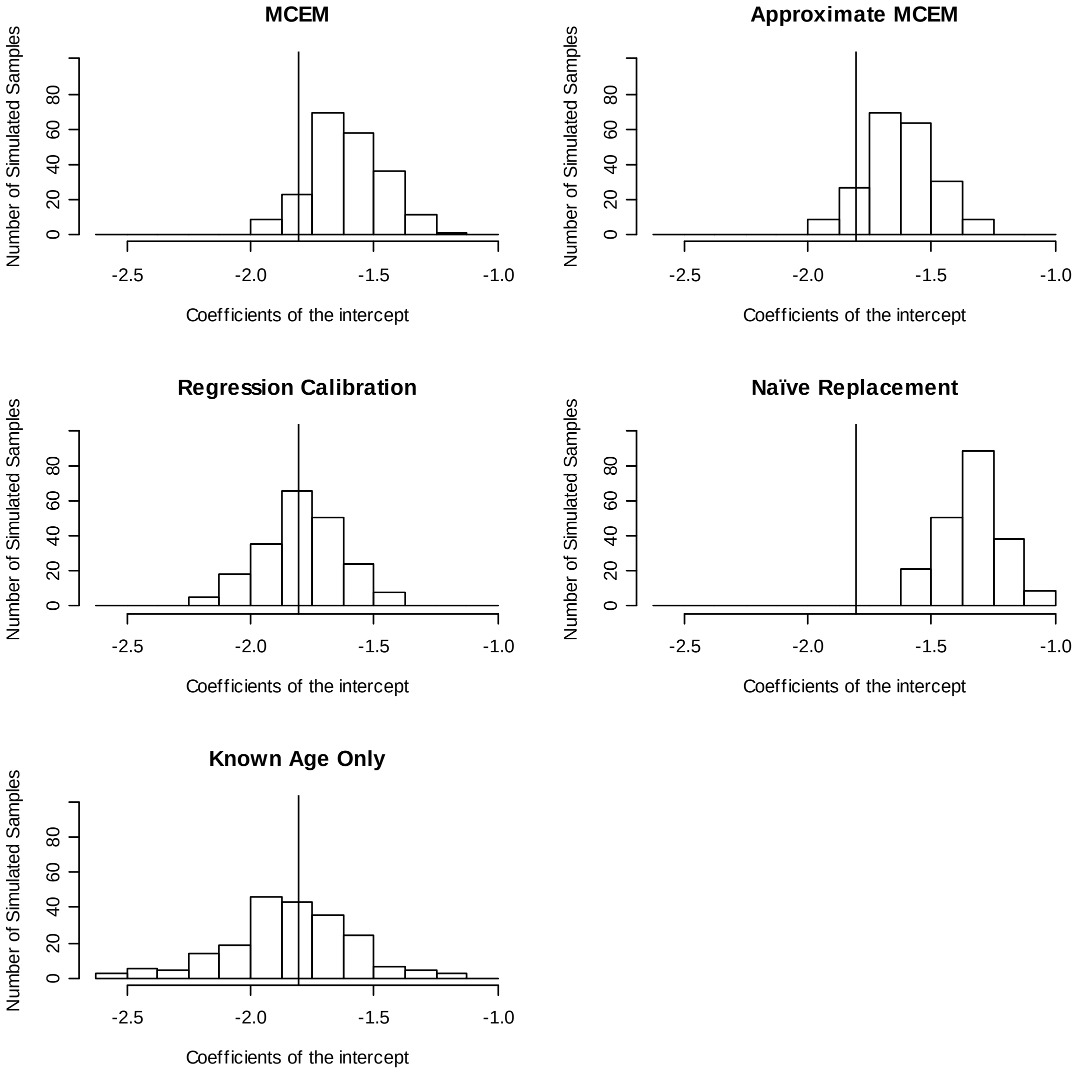


Figure 3.4 Monte Carlo sampling distributions for *n=400* and *50% censored* for the estimators of β_0, the intercept, for the 5 different fitting methods.

The above histograms show the Monte Carlo sampling distribution for the estimators of β_0 for the setting with a sample size of 400 and with 50% of the observations censored. The MCEM and approximate MCEM show an obvious bias, while the naïve replacement estimates are extremely biased. Both the regression calibration and the known age only estimates seem to be unbiased. The regression calibration estimator has smaller MSE than the known age only estimator. See Appendix A6 for the Monte Carlo sampling distributions for the other parameters and simulation conditions.

Since the estimators of the linear and quadratic terms are correlated, it helps somewhat to see at least one picture of the estimated regression curve. The solid line in Figure 3.6 shows the curve of the mean number fledged that was used in the simulation (the log of the mean is taken to be $-1.8070 + 0.3855\, age - 0.0235\, age^2$). The dashed line represents the mean MCEM fit. The dotted line represents mean fit from the naïve replacement method. Even though there appears to be some bias in the estimates from the MCEM algorithm, the curve is very close to the target curve. However, the curve from the fit obtained by replacing the censored age with the lower bound departs from the target curve to a greater degree. Note how the curve based on the naïve replacement decreases the age at peak reproductive success by 0.80 years. The plot below illustrates that simply using the naïve replacement method for censored ages can yield biased results.

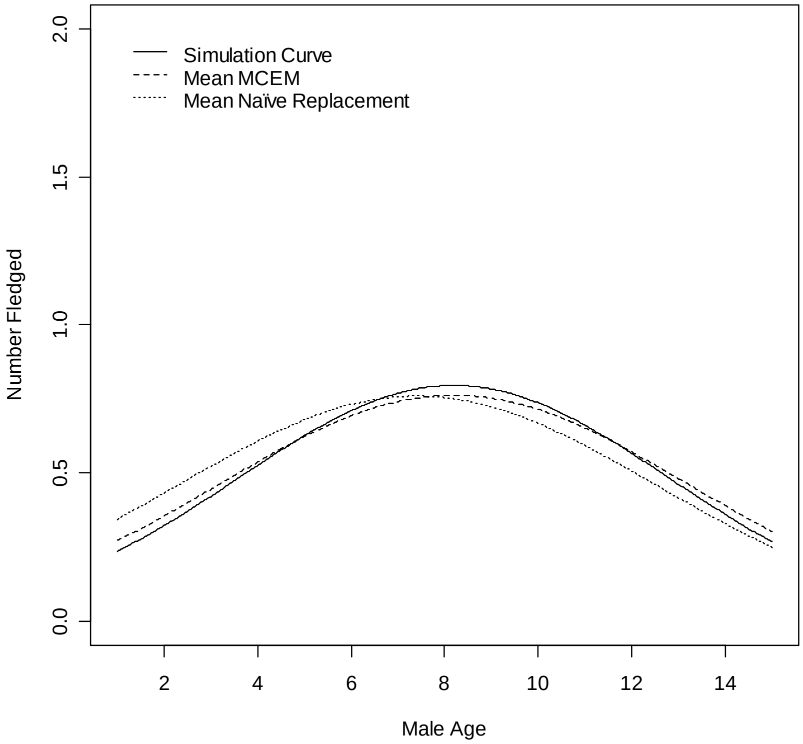


Figure 3.5 Curves based on the mean values of $\hat{\beta}$ from the simulation

In evaluating these simulations, especially those based on the conditions that roughly match the actual data set (n=400 and censoring percentage = 50%), we are particularly interested in these questions: Is there substantial precision gained by including the owls with censored ages, does the simple regression calibration method achieve this gain nearly as well as the maximum likelihood estimator based on the MCEM algorithm, and is inference based on the standard errors and approximate normality of the regression calibration estimator supported by the simulation results? Figure 3.6 shows the MSEs for three estimators of β_2: the MCEM, the regression calibration estimator, and GLMM estimator using the known age owls only. When the sample size is 50, both the MCEM and the regression calibration estimator perform much better than the estimator based on known age owls only. However, when the sample size is 400, the regression calibration estimator out performs the MCEM. This outperformance is due to the bias in the estimates for the MCEM.

With 50% censoring and a sample size of 50, the MSE of the regression calibration estimator is 35% of the MSE of the GLMM estimator based on known owls only. With 50% censoring and a sample size of 400, which most closely matched the spotted owl data set, the regression calibration MSE is 43% of that for the GLMM estimator based on only known age owls. (Similar statements are true for the estimators of the other regression coefficients. The MSE of the regression calibration estimator of β_0 is 37% and 40% of the MSE of the GLMM estimator with known age owls only with 50% censoring and sample sizes of 50 and 400 respectively. For β_1 the MSE of the regression calibration estimator is 35% and 43% of the MSE of the GLMM estimator with known age owls only.)

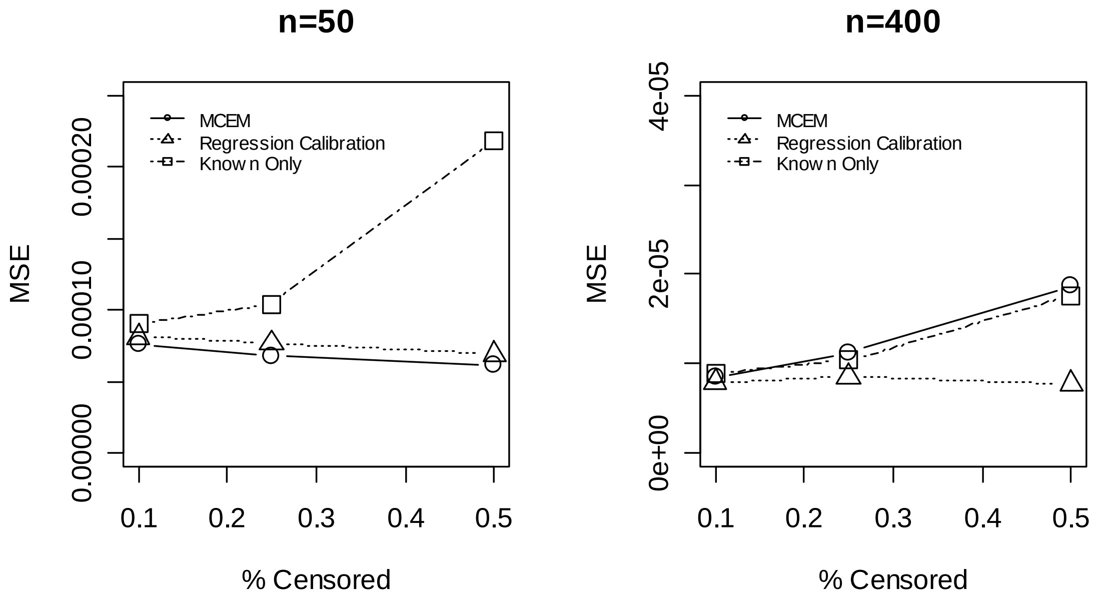


Figure 3.6 Mean Squared Error for three estimators of β_2

We also wish to see if the computed standard errors adequately approximate the standard deviations of the sampling distribution. Figure 3.7 shows the mean reported standard error vs. the Monte Carlo standard deviation for three estimators of β_2. In all cases, it appears that the reported standard error is a good approximation. This seems particularly important for the regression calibration estimator, in which the reported standard error is the usual one obtained from the GLMM fitting procedure without any further adjustment. This fact, with the evidence from Figures 3.3 and 3.4 that the sampling distribution is roughly normally-shaped, indicates no obvious problems with usual inferences based on approximate normality and reported standard errors.

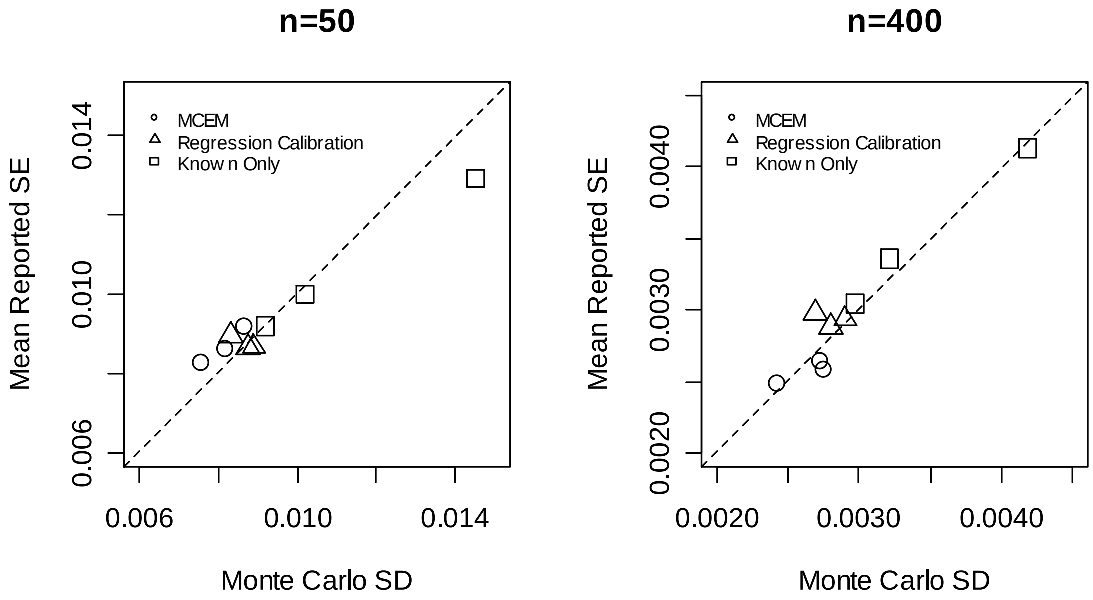


Figure 3.7 Reported Mean SE vs. Monte Carlo SD for three estimators of β_2

3.7 Analysis of Spotted Owl Data

The regression calibration method for censored covariates will now be used to analyze one of the study areas from the Spotted Owl example from Section 1.2. The Oregon Coast Range study area is located in the central Coast Ranges of Oregon. There are 108 known age owls and 165 owls with censored ages. The known age owls have 542 owl/year observations, while the censored owls have 839 owl/year observations.

This analysis will examine two particular questions that the wildlife biologists are interested in gaining insight on. First, they would like to know the age at peak reproductive success, after accounting for differences between years. The biologists are also interested in possible declines in reproductive success after reaching peak performance. The table below shows the parameter estimates and standard errors. See Appendix A7 for plots and tables of estimation results for the other study areas.

Table 3.2 Regression calibration estimation results for Oregon Coast Range

	Regression Calibration All Males		lmer Known Age Males Only	
Parameter	Estimate	SE	Estimate	SE
β_0	-1.3797	0.2680	-1.8092	0.6796
β_{age}	0.3037	0.0546	0.4824	0.0944
β_{age^2}	-0.0154	0.0029	-0.0269	0.0058
β_{1991}	-1.4978	0.3664	-15.8338	2211.3104
β_{1992}	-0.0947	0.2178	-15.9970	1157.3002
β_{1993}	-1.5129	0.3260	-0.9729	0.9513
β_{1994}	-0.2715	0.2193	-0.7218	0.7130
β_{1995}	-1.7880	0.3197	-1.7309	0.7672

β_{1996}	0.0568	0.2108	-0.0933	0.6490
β_{1997}	-1.1195	0.2507	-1.1617	0.6750
β_{1998}	-0.5004	0.2272	-0.3259	0.6477
β_{1999}	-2.0656	0.3398	-2.1913	0.7342
β_{2000}	-0.6539	0.2398	-0.6369	0.6564
β_{2001}	0.1743	0.2187	-0.0490	0.6461
β_{2002}	-0.9505	0.2716	-0.8955	0.6704
β_{2003}	-2.7693	0.4940	-2.9424	0.8631
β_{2004}	-0.0647	0.2315	-0.1555	0.6507
β_{2005}	-0.7043	0.2646	-0.9596	0.6745
α_0	1.2189	0.0879		
α_{years}	-0.0175	0.0144		
σ_u	0.2769		0.3388	
σ_a	0.5654			

Based on the regression calibration analysis that considers all 273 owls, the age at peak reproductive success is estimated to be 9.86 years. Based on the GLMM analysis using only the 108 known age owls, the age at peak reproductive success is estimated to be 8.97 years. Both analyses suggest a decline in reproductive success after peak reproductive success is reached (one sided p-value for $\beta_{age^2} < 0.0001$ for both analyses). The solid lines in the plot below show the estimated curve for the mean number fledged. The regression calibration analysis using all of the owls has the heavier lines, while the analysis that just uses the known age owls uses the lighter lines. The dashed lines show a 95% confidence bands for the mean fixed effect of male age. The heavier set of lines is for the analysis that considers all owls and the lighter set are for known age owls only. These confidence bands include between-owl variability in intercepts. The analysis that considers all owls has a much tighter confidence interval, especially around the peak of the known age curve.

The extreme negative estimates and large standard error for the years 1991 and 1992 in the known age only analysis are due to 3 observations and 10 observations that were all 0. The results of the analysis would probably not change if these years were removed.

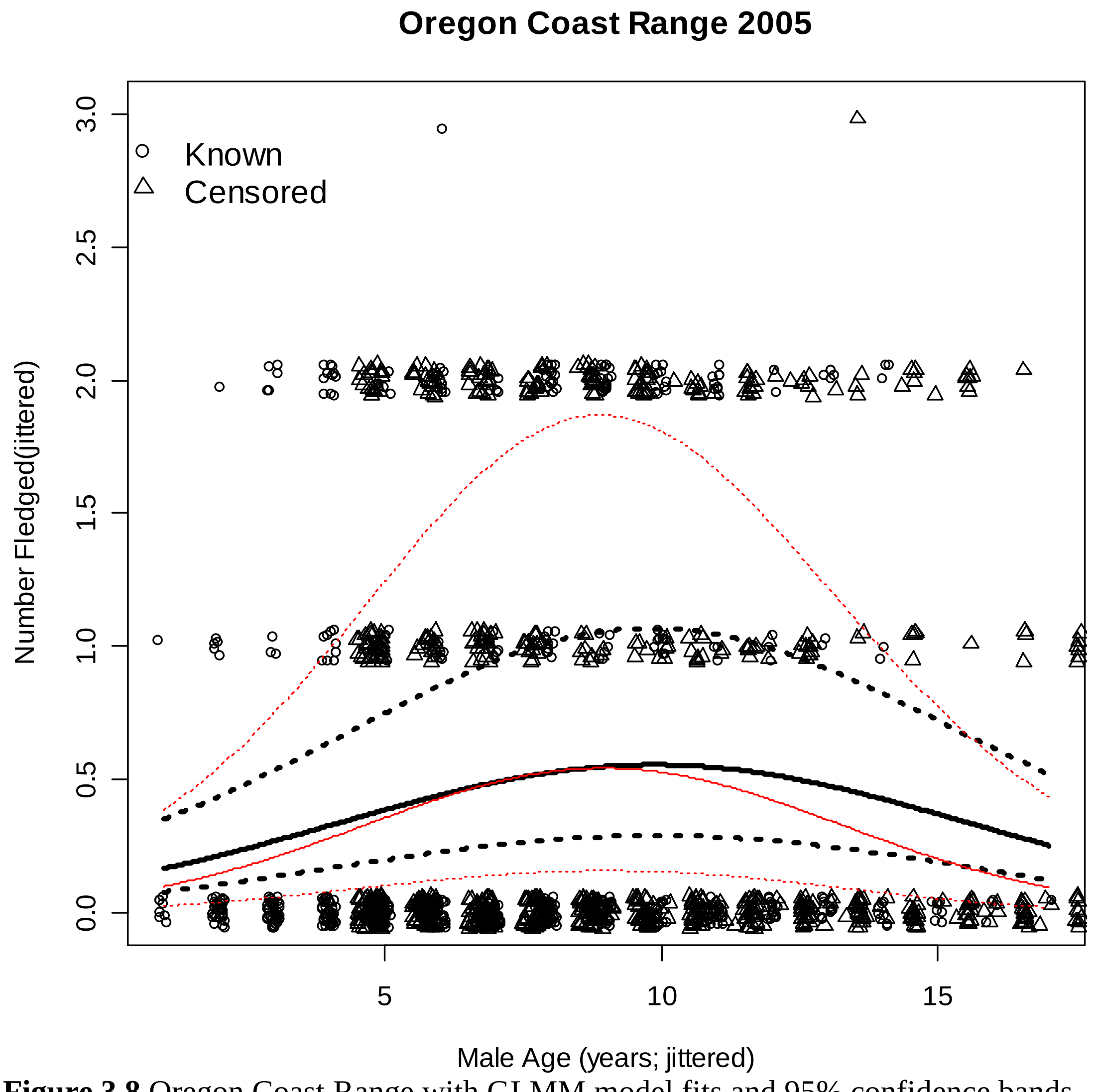


Figure 3.8 Oregon Coast Range with GLMM model fits and 95% confidence bands using only 108 known-age owls (thin line) and using all 273 owls (thick line). The owls with censored ages are plotted at their conditional expected ages.

3.8 Discussion

While the approximate MCEM algorithm is slightly faster than the MCEM algorithm of chapter 2, it is still a computationally intensive method. The number of iterations required for convergence (with a relative convergence criterion that estimates change by less than 0.5% in successive iterations) was generally around 14. The computational time required to generate the pseudo-random variables in the Metropolis step is less than the MCEM method of chapter 2 because we only need to generate first observed ages for the censored animals. However, the M-step takes longer because we are fitting a GLMM instead of a GLM with the random effects as offsets.

Both the MCEM and the approximate MCEM algorithms suffer from bias. It is not clear why, but it is possible that the convergence criterion that was used is too large. The EM algorithm in general can suffer from slow convergence, and the MCEM algorithm has the added complexity of the Monte Carlo estimate in the E-step. Unfortunately, the extremely large Monte Carlo sample sizes that are needed to insure that the Monte Carlo estimate of the E-step is “close” to the actual intractable integral can often cause memory failures in R. As computing power and memory increase, it is likely that one would be able to use a smaller convergence criterion. This may solve the bias issue in the MCEM algorithms, but more research may be needed to clarify exactly why these methods are biased.

The approximate MCEM algorithm would be difficult for a biologist to implement without further computer programming skills. The regression calibration method of Section 3.5 seems to be a practical approach for including censored age animals in the analysis, but without the computational hurdles of the MCEM and approximate MCEM algorithms. We feel that the regression calibration method could be easily implemented by a researcher who has had some formal statistics training because this method is relatively simple extension of the methods that would be used if all of the ages were known.

Our simulation study demonstrates that the regression calibration method performs substantially better than either using the naïve replacement method or using only known age animals in the analysis. With respect to bias in the parameter estimates, this method performs better than the MCEM algorithm and nearly as well in reducing the variance of the estimates.

3. 9 References

Altwegg, R., Schaub, M., and Roulin, A.,(2007). Age-Specific Components of Temporal Variation in the Barn Owl. The American Naturalist. 169:47-61

Angelier, F.,et al. (2007). Age-specific reproductive success in a long-lived bird: do older parents resist stress better? *Journal of Animal Ecology*.76:1181–1191

Austin, P.C. and Hoch J.S., 2004, Estimating Linear Regression Models in the Presence of Censored Independent Variable, *Statistics in Medicine,* 23, 411-429

Booth, J.G, and Hobert, J.P. (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM Algorithm *Journal of the Royal Statistical Society B.* 61, 265-285

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25

Coltman, D.W., Festa-Bianchet, M, Jorgenson, J. T. and Strobeck, C. (2002) Age dependent sexual selection in bighorn rams. *Proceeding of the Royal Society London B.* 269, 165-172

Côté , S. D., and Festa-Bianchet, M., (2001) Offspring sex ratio in relation to maternal age and social rank in mountain goats (Oreamnos americanus). *Behavioral Ecology and Sociobiology*. 49: 260-265

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B.* 39, 1-38.

Green, D. J. (2001).The influence of age on reproductive performance in the Brown Thornbill, *Journal of Avian Biology*. 32: 6–14.

Horton, N.J. and Laird, N.M., 1998, "Maximum Likelihood Analysis of Generalized – Model with Missing Covariates," *Statistical Methods in Medical Research,* 8, 37-50

Jiming, Jiang (2007). *Linear and Generalized Linear Mixed Models and Their Applications.* Springer, New York.

Krüger, O., 2005. Age at first breeding and fitness in goshawk *Accipiter gentilis. Journal of Animal Ecology* 74:266–273

Levine, R.A., and Casella, G. (2001). Implementations of the Monte Carlo EM Algorithm. *Journal of Computational and Graphical Statsitics*. 10, 422-439.

Levine, R.A., and Fan, J. (2003). An Automated (Markov Chain) Monte Carlo EM Algorithm. *Journal of Statistical Computation and Simulation*. 74, 349-360.

Little, R., 1992, "Regression With Missing X's: A Review," *Journal of the American Statistical Association,* 87, 1227-1237

Loschl, P., (2008). Age-specific and Lifetime Reproductive Success of Known Age Northern Spotted Owls on Four Study Areas in Oregon and Washington., MS Thesis, Oregon State University

Louis, T.A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society. Series B.* 44 226-233.

McCulloch, C.E. (1997) Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association.* **92** 162-170.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models.* John Wiley and Sons, New York.

McLachlan, G.J, and Krishnan, T. (1997). *The EM Algorithm and Extensions.* John Wiley and Sons, New York.

Monleon, V.J., (2006). Regression Calibration and Maximum Likelihood Inference for Measurement Error Models, PhD Dissertation, Oregon State University.

Newton, I., and Rothery, P., (2002) . Age-Related Trends in Different Aspects of the Breeding Performance of Individual Female Eurasian Sparrowhawks *The Auk* 119(3):735–748

Oakes, D. (1999). Direct Calculation of the Information Matrix via the EM Algorithm. *Journal of the Royal Statistical Society. Series B.* 61 479-482.

Reid, J.M., Bignal E.M., Bignal, S. McCracken, D.I., and Monaghan, P. (2003). Age-specific reproductive performance in red-billed choughs *Pyrrhocorax pyrrhocorax*: patterns and processes in a natural population. Journal of Animal Ecology. 72 (5) , 765–776 776

Schafer, D.W. and Gilbert, E.S, (2006) Statistical Implications of Dose Uncertainties in Radiation Dose-Response Analyses of Epidemiological Data, *Radiation Research*, 166, 303-312.

Schall, R. (1991) Estimation in generalized linear models with random effects. *Biometrika* **78**, 719–727.

Wolfinger, R. (1993). Laplace’s Approximation for Nonlinear Mixed Models. *Biometrika.* 80 791-795.

Wu, K., and Wu, L. (2007) Generalized linear mixed models with informative dropouts and missing covariates. *Metrika.* 66, 1-18

Zedrosser, A. et al. (2007). Genetic estimates of annual reproductive success in male brown bears: the effects of body size, age, internal relatedness and population density. *Journal of Animal Ecology* 76: 368–375.

4. Conclusions

This dissertation proposes three algorithms for including censored age individuals in an age specific reproductive success analysis. Special attention is given to a specific Northern Spotted Owl study (Loschl 2008), but the problem of censored age individuals is common in studies of age specific reproductive success. Recent age-specific reproductive studies that excluded censored age individuals and used GLMMs include barn owls (Altwegg et al. 2007), snow petrels (Angelier et al. 2007), brown thornbill (Green 2001), goshawk (Kruger 2005), brown bears (Zedrosser, et al. 2007), mountain goats (Côté, Festa-Bianchet 2001), and sparrowhawks (Newton, Rothery 2002). As evident in the Northern Spotted Owl study, the percentage of animals with unknown ages can be substantial.

Two of the algorithms are Monte Carlo EM algorithms that require generating pseudo-random numbers to calculate the intractable expectation step. The first algorithm treats both the censored ages and the random effects as “missing data.” The second method only treats the censored ages as “missing data.” The pseudo-random number generation is accomplished via Metropolis algorithms. Both of these methods are computationally intensive and require the researcher to write a significant amount of computer code. In addition, the algorithms can be slow to converge and additional computations must be done to obtain standard errors.

Since our goal was to develop a practical method that would be easy to implement and given the above issues with the MCEM algorithms, a third method was developed. The regression calibration method simply replaces the censored observations with their estimated expected values given appropriate covariates for predicting age and given the lower bound. Then the age specific reproductive success analysis can proceed using the methods that would be appropriate if all of the ages were known. Researchers with a working knowledge of applied statistical methods (like regression, ANOVA, and generalized linear models) should be able to easily implement this method using standard computer routines. We have found no evidence of any problem with using reported standard errors and approximate normality for inferences. Figure 4.1 shows the estimated mean curves for the known age only analysis (lighter lines) and the regression calibration analysis (heavier lines). The confidence interval for the regression calibration method is much tighter than the method that only uses known age owls.

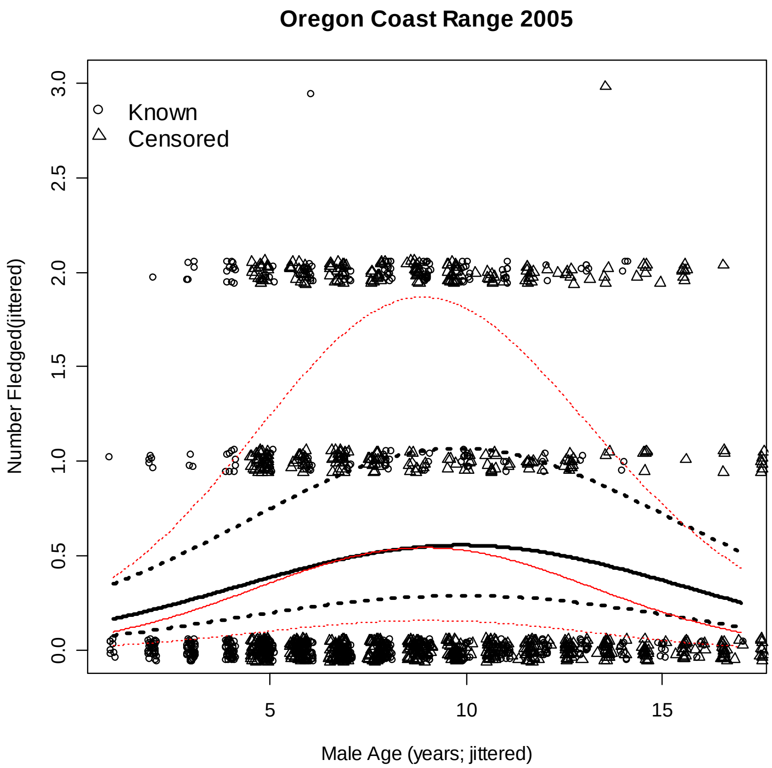


Figure 4.1 Oregon Coast Range with GLMM model fits and 95% confidence bands using only 108 known-age owls (thin line) and using all 273 owls (thick line).

Chapter 3 showed the results of a simulation study to examine how 5 different fitting methods performed on 6 different setting. The settings were chosen to cover a range of possible sample sizes and proportion of censored age individuals. Based on the simulation study, it is clear that the naïve replacement method that uses the lower bound as the true age causes biased estimates of the regression parameters. This bias increases as both the sample size increase and the proportion of censored age individuals increase. For the biologists studying age specific reproductive success, this result should be noted and more sophisticated models should be considered when including censored individuals in the analysis.

The rest of the methods that include the censored observations succeed in increasing the precision of the estimates. Both of the MCEM algorithms show some bias in the regression estimates. The estimates from the regression calibration method have slightly more variability than the MCEM estimates, but do not exhibit the bias of the MCEM estimates. The regression calibration estimates are much less variable than the estimates that are obtained by restricting attention to known age individuals only. This increased precision from the regression calibration method will help researchers answer important biological questions.

Bibliography

Altwegg, R., Schaub, M., and Roulin, A.,(2007). Age-Specific Components of Temporal Variation in the Barn Owl. The American Naturalist. 169:47-61

Angelier, F., et al. (2007). Age-specific reproductive success in a long-lived bird: do older parents resist stress better?. *Journal of Animal Ecology*. 76:1181-91

Austin, P.C. and Hoch J.S., (2004), "Estimating Linear Regression Models in the Presence of Censored Independent Variable," *Statistics in Medicine,* 23, 411-429

Bates, D.M., (2005), Fitting linear mixed models in R, *R News* **5** pp. 27–30.

Bates, D. M. (2007). *Linear mixed model implementation in lme4*. Manuscript, University of Wisconsin - Madison, January 2007.

Booth, J.G, and Hobert, J.P. (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM Algorithm *Journal of the Royal Statistical Society B.* 61, 265-285

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25

Coltman, D.W., Festa-Bianchet, M, Jorgenson, J. T. and Strobeck, C. (2002) Age dependent sexual selection in bighorn rams. *Proceeding of the Royal Society London B.* 269, 165-172

Côté , S. D., and Festa-Bianchet, M., (2001) Offspring sex ratio in relation to maternal age and social rank in mountain goats (Oreamnos americanus). *Behavioral Ecology and Sociobiology*. 49: 260-265

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B.* 39, 1-38.

Green, D. J. (2001).The influence of age on reproductive performance in the Brown Thornbill, *Journal of Avian Biology*. 32: 6–14.

Horton, N.J. and Laird, N.M., 1998, “Maximum Likelihood Analysis of Generalized – Model with Missing Covariates,” *Statistical Methods in Medical Research,* 8, 37-50

Jiming, Jiang (2007). *Linear and Generalized Linear Mixed Models and Their Applications.* Springer, New York.

Krüger, O., 2005. Age at first breeding and fitness in goshawk *Accipiter gentilis. Journal of Animal Ecology* 74:266–273

Levine, R.A., and Casella, G. (2001). Implementations of the Monte Carlo EM Algorithm. *Journal of Computational and Graphical Statsitics*. 10, 422-439.

Levine, R.A., and Fan, J. (2003). An Automated (Markov Chain) Monte Carlo EM Algorithm. *Journal of Statistical Computation and Simulation*. 74, 349-360.

Little, R., 1992, “Regression With Missing X’s: A Review,” *Journal of the American Statistical Association,* 87, 1227-1237

Loschl, P., (2008). Age-specific and Lifetime Reproductive Success of Known Age Northern Spotted Owls on Four Study Areas in Oregon and Washington., MS Thesis, Oregon State University

Louis, T.A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society. Series B.* 44 226-233.

McCulloch, C.E. (1997) Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association.* **92** 162-170.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models.* John Wiley and Sons, New York.

McLachlan, G.J, and Krishnan, T. (1997). *The EM Algorithm and Extensions.* John Wiley and Sons, New York.

Monleon, V.J., (2006). Regression Calibration and Maximum Likelihood Inference for Measurement Error Models, PhD Dissertation, Oregon State University.

Newton, I., and Rothery, P., (2002) . Age-Related Trends in Different Aspects of the Breeding Performance of Individual Female Eurasian Sparrowhawks *The Auk* 119(3):735–748

Oakes, D. (1999). Direct Calculation of the Information Matrix via the EM Algorithm. *Journal of the Royal Statistical Society. Series B.* 61 479-482.

Reid, J.M., Bignal E.M., Bignal, S. McCracken, D.I., and Monaghan, P. (2003). Age-specific reproductive performance in red-billed choughs *Pyrrhocorax pyrrhocorax*: patterns and processes in a natural population. Journal of Animal Ecology. 72 (5) , 765–776 776

Schabenberger, O. (2005). Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models, *Proceedings of the Thirtieth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc. Paper 196-30.

Schabenberger, O. (2007). Growing Up Fast: SAS 9.2 Enhancements to the GLIMMIX Procedure SAS. *Proceedings of the SAS Global Forum 2007 Conference. Cary, NC: SAS Institute Inc.*. Paper 177-2007.

Schafer, D.W. and Gilbert, E.S, (2006) Statistical Implications of Dose Uncertainties in Radiation Dose-Response Analyses of Epidemiological Data, *Radiation Research*, 166, 303-312.

Schall, R. (1991) Estimation in generalized linear models with random effects. *Biometrika* **78**, 719–727.

Wolfinger, R. (1993). Laplace’s Approximation for Nonlinear Mixed Models. *Biometrika.* 80 791-795.

Wu, K., and Wu, L. (2007) Generalized linear mixed models with informative dropouts and missing covariates. *Metrika.* 66, 1-18

Zedrosser, A. et al. (2007). Genetic estimates of annual reproductive success in male brown bears: the effects of body size, age, internal relatedness and population density. *Journal of Animal Ecology* 76: 368–375.

APPENDICES

A1 R Code for fitting MCEM with Censored Covariates

```
#### MODEL #####
# y|x,u ~ Poisson(mu)
# log(mu) = beta[0] + beta[1]*age + beta[2]*age^2 + u
# u ~ N(0,sdu)
# where age = Z %*% age.first + years.since.first
# where age.first is the true age at first observation (unknown for the censored owls)
# log(age.first|years.of.obs) ~ N(alpha[0]+alpha[1]*years.of.obs,sda)


# Metropolis-Hastings Generation




#[1]"male"    "year"    "age"     "fledged" "area"    "censor"
n       <- length(male)

# PUT OWL-SPECIFIC VARIABLES IN m-VECTORS

malem          <- unique(male)       # mx1 vector of unique male IDs
m              <- length(malem)      # 118 # number of males
age.first.obs  <- rep(0,m)    # age at first observation or lower bound, for each male
censorm        <- rep(0,m)    # mx1 vector: 1 if age is censored for owl j, 0 if not
years.of.obs   <- rep(0,m)    # total years of observation on owl j
number.of.obs  <- rep(0,m)    # number of observations on owls j; j = 1,...,m

Z              <- matrix(rep(0,n*m),n,m)     # Design matrix with male indicators

for (j in 1:m) {
        age.first.obs[j]      <- min(age[male==malem[j]])
        censorm[j]            <- mean(censor[male==malem[j]])
        years.of.obs[j]       <- max(age[male==malem[j]])-min(age[male==malem[j]])
        number.of.obs[j]      <- length(fledged[male==malem[j]])
        Z[,j]          <- ifelse(male==malem[j],1,0)
}

mc <- m - sum(censorm)        # males with known ages
m-mc                          # with censored ages
years.since.first <- age - Z %*% age.first.obs  # Years since first observation (nx1
                                                # vector)

####################################################################################

#### METROPOLIS-HASTINGS SIMULATION FUNCTION ####

### UTILITY FUNCTIONS ###

# Function to generate left-truncated normals

        rtnorm  <- function(mu,sd,lo) {
        n              <- length(mu)
        mu + sd*qnorm(runif(n,pnorm((lo-mu)/sd),1))
}


# Function to generate random ages given lower bound, from truncated lognormal

acandidate <- function(mua,sda,lowest.age) {
        # mua is mx1 vector of means for log age at first observation
        # sda is 1x1 constant standard deviation for log age at first observation
        # lowest.age is mx1 vector of lower bounds for age
        m  <- length(mua)
        lacandidate    <- rtnorm(mua,sda,log(lowest.age)) # simulated log ages at first
                                                          # obs
        exp(lacandidate)      # simulated ages (not logged aged) at first observation
        }
```

```
metropolis <- function(theta,fledged,censor,age.first.obs, years.since.first, R) {
        n              <- length(fledged)
        m              <- length(age.first.obs)
        U      <- matrix(0,nrow=R,ncol=m) # Rxm Matrix to be filled in with simulated
                                           # random effects, u
        A      <- matrix(0,nrow=R,ncol=m) # Rxm matrix to be filled in with simulated
                                           # ages at first obs.
        alpha  <- theta$alpha
        beta           <- theta$beta
        sdu            <- theta$sdu
        sda            <- theta$sda

        # Initialize
        critu  <- rep(0,m) # to store the acceptance criterion for candidates for u
        crita  <- rep(0,m) # to store the acceptance criterion for candidates for age
                           #  at first observation
        ones   <- matrix(rep(1,n),n,1) # nx1 vector of 1's

        Y      <- matrix(rep(fledged,m),n,m)  # nxm; m copies of the response
        Years.since <- matrix(rep(years.since.first,m),n,m)  # nxm; m copies of years
                                                             # since first observation

        mua    <- alpha[1] + alpha[2]*years.of.obs[censorm==1] # mean of lognormal
                                                       # regression for censored ages
        U[1,]  <- rnorm(m,0,sdu)     # initial simulation of random effects
    A[1,] <-
c(age.first.obs[censorm==0],acandidate(mua,sda,age.first.obs[censorm==1]))
                              # The first mc rows of A[r,] are the ages at first
                              # observation for owls with known ages. The remainder
                              # are simulated values from the truncated lognormal
                              # regression.

        for (r in 2:R) {
               # Generate candidate vectors for u and a
               u.candidate <- rnorm(m,0,sdu)
               a.candidate <-
        c(age.first.obs[censorm==0],acandidate(mua,sda,age.first.obs[censorm==1]))

               ### CALCULATE DENOMINATOR OF ACCEPTANCE CRITERION VECTORS ####
               age <- Z %*% A[r-1,] + years.since.first   nx1; age = "age at first
                                                     observation" + years since first
               fixed   <- beta[1] + beta[2]*age + beta[3]*(age^2)   # nx1; fixed
                                                                    # effects in linear
                                                                    # predictor using
                                                                    # previous
                                                                    # simulated a's
               Fixed   <- matrix(rep(fixed,m),n,m)   # nxm; m copies of fixed effects
               umat    <- matrix(rep(U[r-1,],m),m,m) # mxm; m copies of previously
                                                             # simulated random effects
               Eta     <- Fixed + Z %*% umat         # nxm; m copies of linear
                                                     # predictor based on previously #
                                                     simulated u's, a's
               Density <- dpois(Y,exp(Eta)) # nm x 1; Poisson density at the linear
                                                     # predictor
               Density <- matrix(Density,n,m) # nxm; m copies of Poisson density at
                                              # linear predictor; previous simulations
               denom   <- exp(t(Density) %*% ones)   # mx1; jth element is the product
                                                     # of Poisson pmfs at previously
                                                     # simulated values of uj and aj

               ### GET ACCEPTANCE VECTOR FOR U ###
               umat2   <- umat   # mxm; m copies of previously simulated u's
               diag(umat2) <- u.candidate  # mxm; Replace diagonal elements with new
                                            # candidates
               Eta <- Fixed + Z %*% umat2    # nxm
               Density <- dpois(Y,exp(Eta))  #  nxm; jth column has Poisson pmfs with
                                             # all u's equal to the previously
                                             # simulated values, except with uj
                                             # replaced by new candidate
               numer  <- exp(t(Density) %*% ones)   # mx1; jth element is the product
                                                    # of the n Poisson pmfs at
                                                    # previously simulated values of u
                                                    # and a, except with uj replaced
                                                    # by new candidate

               critu  <- ifelse(numer<denom,numer/denom,1)  # mx1; jth element is
                                                            # acceptance probability
                                                            # for candidate uj

               ### GET ACCEPTANCE VECTOR FOR A ###
               amat   <- matrix(rep(A[r-1,],m),m,m) # mxm; m copies of previously
                                                    # simulated a's (age at first
                                                    # observation)
               diag(amat) <- a.candidate     #mxm; Replace diagonal elements with new
                                                   # candidates
               Age <- Z %*% amat + Years.since # nxm; Form ages from ages at first
                                             # observation plus years since first
               Eta    <- beta[1] + beta[2]*Age + beta[3]*(Age^2) + Z %*% umat
               Density <- dpois(Y,exp(Eta))  #  nxm; jth column has n Poisson pmfs
                                             #   with all u's and a's equal to the
                                             #   previously simualted values, except
                                             #   with aj replaced by new candidate
               numer  <- exp(t(Density) %*% ones)   # mx1; jth element is the product
                                                    # of the n Poisson pmfs at
                                                    # previously simulated values of u
                                                    # and a, except with aj replaced
                                                    # by new candidate
               crita  <- ifelse(numer<denom,numer/denom,1)  # mx1; jth element is the
                                                            # acceptance probability
                                                            # for candidate aj

               ### CARRY OUT ACCEPTANCE ###
               bernu <- rbinom(m,1,critu)  # mx1; generate m Bernoullis with the
                                             # candidate u acceptance probabilities
               berna <- rbinom(m,1,crita)  # mx1; generate m Bernoullis with the
                                             # candidate a acceptance probabilities
               U[r,] <- bernu*u.candidate + (1-bernu)*U[r-1,]      # mx1; new u is
                                                                   # either previous u
                                                                   #or new candidate
               A[r,] <- berna*a.candidate + (1-berna)*A[r-1,]      # mx1; new a is
                                                                   #either previous a
                                                                   #or new candidate
                      }
        list(U=U,A=A)  # Return R simulated u vectors and a vectors
}
####################################################################################


# Likelihood function used to compute importance weights

like    <- function(theta) {

    Density1 <- dpois(fledgedvector, exp(xmat %*%theta$beta + uvector))  # glm part
    Density2 <-  dnorm(uvector2,0, theta$sdu)               # random effects part
    Density3 <-
dnorm(log(avector), theta$alpha[1]+ theta$alpha[2]*years.of.obsvector,theta$sda)
                                                            # age at first obs part
    MCnum <- rep(1:R,n)
    MCnum2 <- rep(1:R,m)
    s1 <- tapply(Density1, INDEX=MCnum, sum)        # summing over the MC samples
    s2 <- tapply(Density2, INDEX=MCnum2, sum)       # summing over the MC samples
    s3 <- tapply(Density3, INDEX=MCnum2, sum)       # summing over the MC samples
    return(s1+s2+s3)                                # returning the overall sum

}

#############################################################################
```

```
# MCEM algorithm for censored ages



time.check <-date()
###############################
convergence.criterion <- .005
iter <- 1 # iteration number
maxdif <- 1 # initial setting for max abs. value

# STARTING VALUES
beta.old <-c(-0.403156, 0.141645, -0.008767) # from lmer with censored ages replaced
                                             # by expected given above threshold
alpha.old<- c(0.553586294,  0.005887441) # from lm fit of log agefirst on yearsobsm
                                         # from owls with known ages
sda.old        <- 0.4560279

sdu.old        <-  0.2268224             # from lmer with censored ages replaced



theta.old <- list(beta=beta.old, alpha=alpha.old, sdu=sdu.old, sda=sda.old)



c <-3  # proportion to increase the Monte Carlo sample size by
R <- 8 # starting Monte Carlo sample size


print(c("iter", "beta0", "beta1", "beta2","alpha1", "alpha2",  "sdu", "sda"),
quote=F, sep="\t")
print(round(c(0, unlist(theta.old)),3), sep="\t")



# keep track of the estimates


beta1.iter <- beta.old[1]
beta2.iter <- beta.old[2]
beta3.iter <- beta.old[3]
alpha1.iter <-alpha.old[1]
alpha2.iter <- alpha.old[2]
sdu.iter   <- sdu.old
sda.iter   <- sda.old
maxdif.iter <- maxdif
R.iter <-R


S <- 3 # burn in for regular MCEM before switching to importance weighting



##############################
while (maxdif>convergence.criterion) {
#if (iter==1) {

      # Monte Carlo E-step
        R0 <- R
        R <- R0 + floor(R0/c)

       # Obtain R psuedo random variables using the Metropolis Alg

       if (iter<=S) {
                    sample <- metropolis(theta.old,fledged,censor,age.first.obs,
                    years.since.first, R=R)
                    w <- 1
                    imp.weight1 <- rep(1,R*n)
                    imp.weight2 <- rep(1,R*m)}
```

```
  # Re-initializing theta to use importance weights
 if (iter==S) {
              theta.init <- theta.old
              sample <- metropolis(theta.init,fledged,censor,age.first.obs,
              years.since.first, R=R)
              }


if (iter >S){
    samp.incr <-floor(R0/c)
    samplek <- metropolis(theta.init,fledged,censor,age.first.obs,
               years.since.first, R=samp.incr)
    sample$U <-matrix(c(t(sample$U),t(samplek$U)), nrow=R0+samp.incr, byrow=T)
    sample$A <-matrix(c(t(sample$A),t(samplek$A)), nrow=R0+samp.incr, byrow=T)
        }



# replicating for the augmented data set for the n observations


agevector <-  matrix(t(Z%*%t(sample$A)), ncol=1,
              byrow=T)+rep(years.since.first,rep(R,n))
agevector2 <- agevector^2
uvector <-   matrix(t(Z%*%t(sample$U)), ncol=1, byrow=T)
fledgedvector <- rep(fledged,rep(R,n))
intercept <- rep(1,n*R)
xmat <- cbind(intercept, agevector, agevector2) # glm.fit needs the x's in
                                                # matrix form

#   male      year        R
#   1           98        1
#   .           98        2
#   .           .         .
#   1           98        R
#   1           99        1
#   .           99        2
#   .           .         .
#   1           .         R



# replicating for the augmented data set for each owl


avector <- matrix(sample$A, ncol=1, byrow=T)     # making the a matrix into an
                                                  # m*r vector
uvector2 <- matrix(sample$U, ncol=1, byrow=T)     # making the u matrix into
                                                # an m*r vector
years.of.obsvector <-rep(years.of.obs ,rep(R,m))



# compute the importance weights as in Levine and Casella

if (iter >=S) {w <- (like(theta.old)/like(theta.init))
               imp.weight1 <- rep(w*R,n)
               imp.weight2 <- rep(w*R,m)}



# M-step

##  using glm.fit to obtain estimates for the generalized linear model part
##  this fits the n x R observations in the augmented data set


fit1 <- glm.fit(xmat, fledgedvector, offset=uvector,
               weights=(imp.weight1/sum(w)), family = poisson(), intercept = F)
      beta.new <- fit1$coef

      # Finding random effects sd
      sdu.new  <- sqrt(sum(uvector2^2)/(m*R))

      ## fitting the model for the age of first observation--

      fit2 <- lm(log(avector)~ years.of.obsvector, weights=imp.weight2/sum(w))

      alpha.new  <-fit2$coef
      sda.new  <- summary(fit2)$sigma


      theta.new <- list(beta=beta.new, alpha=alpha.new , sdu=sdu.new , sda=sda.new)


      # keep track of the estimates

      beta1.iter <- c(beta1.iter,beta.new[1])
      beta2.iter <- c(beta2.iter,beta.new[2])
      beta3.iter <- c(beta3.iter,beta.new[3])
      alpha1.iter <- c(alpha1.iter,alpha.new[1])
      alpha2.iter <- c(alpha2.iter,alpha.new[2])
      sdu.iter   <- c(sdu.iter,sdu.new)
      sda.iter   <- c(sda.iter,sda.new)


      # check convergence
      maxdif <- max(abs((unlist(theta.new)-unlist(theta.old))/unlist(theta.old)))

      maxdif.iter <- c(maxdif.iter,maxdif)

      # print current estimate

      print(round(c(iter, unlist(theta.new),maxdif),3), sep="\t")

      iter <- iter+1

      R.iter <- c(R.iter,R)


      beta.old <- beta.new
      alpha.old <- alpha.new
      sdu.old  <- sdu.new
      sda.old  <- sda.new
      theta.old <- theta.new




      }

time.check <- c(time.check,date())

# print mle
mle <- theta.old
mle
round(unlist(mle),4)
```

```
######################################################################################

# Computing Standard Errors

# First Derivatives
b0 <- (-exp(xmat %*%mle$beta + uvector)+fledgedvector)
b1 <- (-agevector*exp(xmat %*%mle$beta + uvector)+fledgedvector*agevector)
b2 <- (-agevector^2*exp(xmat %*%mle$beta + uvector)+fledgedvector*agevector^2)

a0 <- (log(avector)-mle$alpha[1]-mle$alpha[2]*years.of.obsvector)/mle$sda^2
a1 <- years.of.obsvector*(log(avector)-mle$alpha[1]-mle$alpha[2]*years.of.obsvector)
/mle$sda^2

su <- (-1/mle$sdu -uvector2^2/mle$sdu^3)
sa <- ((-1/mle$sda)+((log(avector)-mle$alpha[1] mle$alpha[2]*years.of.obsvector )^2/
mle$sda^3))



# Summing up over the repeated measures to the males

male.vector <- rep(male,rep(R,n))+(rep(1:R,n)/(R+1))
b00 <- tapply(b0,male.vector,sum)
b11 <- tapply(b1,male.vector,sum)
b22 <- tapply(b2,male.vector,sum)


# Cacluating the approximate information matrix

info <-  matrix(rep(0,49),nrow=7)

for (i in 1:(m*R)){

        info <- info +c(b00[i],b11[i],b22[i], a0[i], a1[i], su[i],
sa[i])%*%t(c(b00[i],b11[i],b22[i], a0[i], a1[i], su[i], sa[i]))
}

approx.info <- 1/R*info


var.cov <-solve(approx.info)
se <- sqrt(diag(var.cov))
se
```

A2 MCEM Spotted Owl Analysis

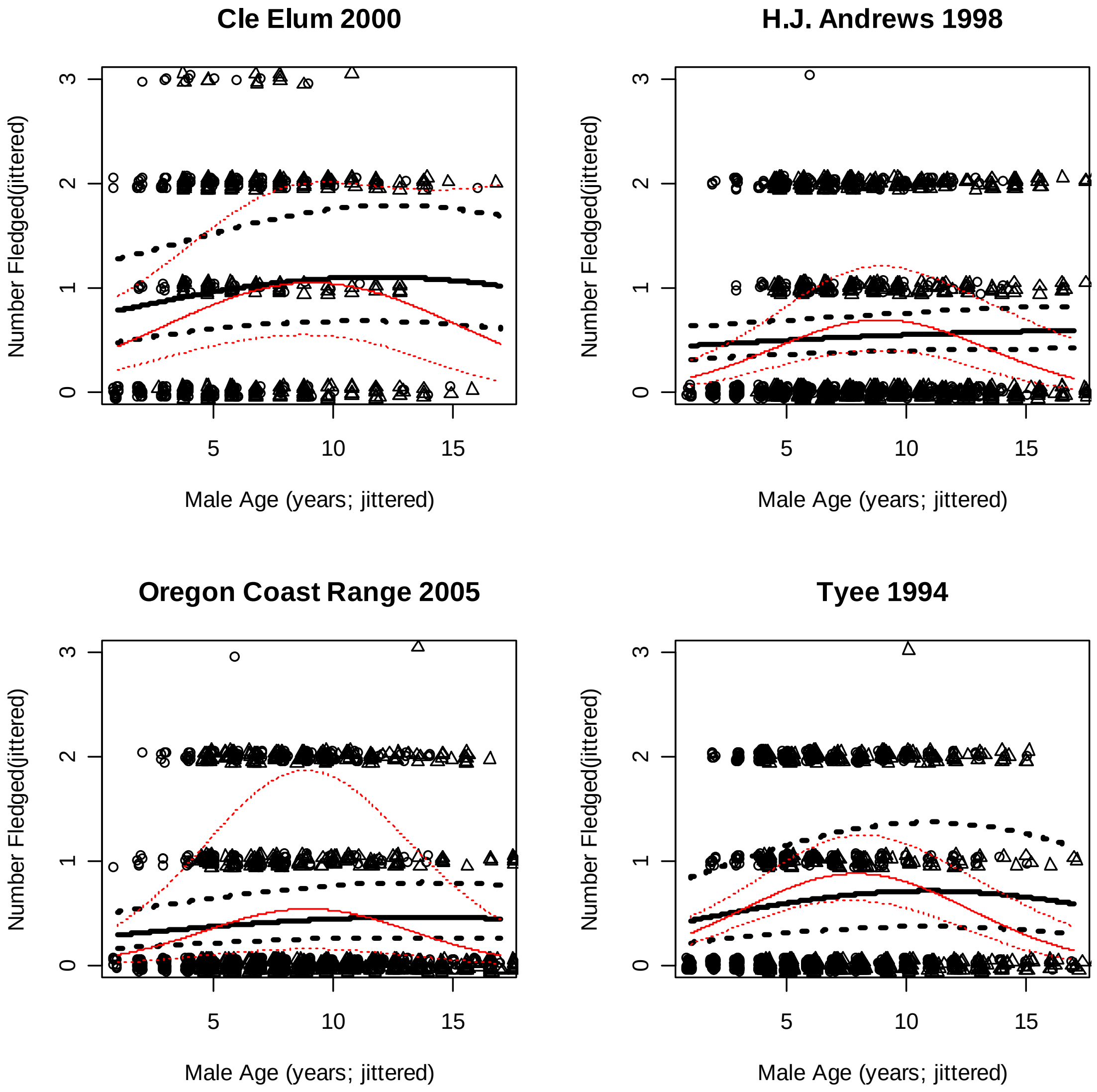


Figure A2.1 Northern Spotted Owl Model Fits. The darker lines are the full MCEM analysis and the lighter lines are the known age owls only. The dashed lines are the 95% confidence intervals.

Table A2.1 MCEM estimation results for Cle Elum

	MCEM All Males		lmer Known Age Males Only	
Parameter	Estimate	SE	Estimate	SE
β_0	0.1592	0.4525	0.2118	0.7419
β_{age}	0.0642	0.0246	0.2417	0.0880
β_{age^2}	-0.0026	0.0009	-0.0132	0.0062
β_{1990}	-0.1072	0.5646	-1.1714	1.0134
β_{1991}	-0.4952	0.5054	-0.8920	0.8327
β_{1992}	0.0410	0.5204	-0.2413	0.7657
β_{1993}	-1.8857	0.4925	19.1415	3042.1741
β_{1994}	0.0074	0.4808	-0.2921	0.7634
β_{1995}	-0.6642	0.4648	-1.1262	0.7938
β_{1996}	-0.0279	0.4942	-0.5873	0.7686
β_{1997}	-2.3662	0.6260	-19.2933	2755.8640
β_{1998}	-0.0672	0.4835	-0.7237	0.7715
β_{1999}	-0.9150	0.5114	-1.1869	0.7928
β_{2000}	-0.4616	0.4935	-1.2653	0.7885
β_{2001}	-0.3660	0.5215	-1.0528	0.7854
β_{2002}	-0.6304	0.5125	-1.6525	0.8534
β_{2003}	-0.2498	0.5065	-0.8773	0.7867
β_{2004}	-0.5228	0.5169	-0.8920	0.7847
β_{2005}	-0.5625	0.5178	-1.0588	0.7959
α_0	1.1305	0.0259		
α_{years}	0.0200	0.0031		
σ_u	8.0e-07	4.0e-08	0.1798	
σ_a	0.6773	0.0012		

Table A2.2 MCEM estimation results for H.J. Andrews

	MCEM All Males		lmer Known Age Males Only	
Parameter	Estimate	SE	Estimate	SE
β_0	0.4757	49.3804	-1.3088	0.8309
β_{age}	0.0347	0.0113	0.4531	0.1069
β_{age^2}	-0.0009	0.0002	-0.0254	0.0069
β_{1987}	-0.6842	49.3847	-17.5988	5717.5404
β_{1988}	-0.6900	49.3809	-0.4866	1.0280
β_{1989}	-1.4987	49.3804	-0.9190	0.8886
β_{1990}	-1.2441	49.3808	-0.6830	0.8155
β_{1991}	-1.5554	49.3804	-2.0862	1.0065
β_{1992}	-0.4220	49.3805	-0.0005	0.7421
β_{1993}	-17.9970	5333987.0386	-17.7043	1565.3191
β_{1994}	-1.5571	49.3804	-1.1521	0.7923
β_{1995}	-1.7569	49.3805	-1.6220	0.8175
β_{1996}	-0.6952	49.3804	-0.3271	0.7356
β_{1997}	-1.6209	49.3805	-1.3016	0.7760
β_{1998}	-1.3257	49.3805	-1.0785	0.7616
β_{1999}	-2.0027	49.3806	-1.0082	0.7657
β_{2000}	-0.9708	49.3804	-1.0164	0.7661
β_{2001}	-0.8354	49.3803	-0.4137	0.7360
β_{2002}	-0.9515	49.3804	-1.1963	0.7716
β_{2003}	-1.8307	49.3804	-1.2249	0.7745
β_{2004}	-0.6150	49.3803	-0.2838	0.7323
β_{2005}	-2.2662	49.3810	-1.5017	0.8226
α_0	1.7161	0.0084		
α_{years}	-0.0065	0.0010		
σ_u	0.0100	0.0003	0.0000	
σ_a	0.6898	0.0002		

Table A2.3 MCEM estimation results for Tyee

	MCEM All Males		lmer Known Age Males Only	
Parameter	Estimate	SE	Estimate	SE
β_0	0.0540	1.0181	14.7886	717.0570
β_{age}	0.0754	0.0125	0.3497	0.0697
β_{age^2}	-0.0029	0.0004	-0.0221	0.0046
β_{1985}	-2.5912	1.4499		
β_{1986}	-0.3203	1.0531		
β_{1987}	-0.9651	1.0913		
β_{1988}	-1.8897	1.1020		
β_{1989}	-0.7526	1.0413	13.3596	717.0573
β_{1990}	-0.8360	1.0346	13.0984	717.0571
β_{1991}	-1.3267	1.0227	12.7455	717.0571
β_{1992}	-0.5289	1.0321	13.0840	717.0570
β_{1993}	-2.1139	1.0423	11.8633	717.0571
β_{1994}	-0.8800	1.0244	13.2773	717.0570
β_{1995}	-1.5128	1.0275	12.6601	717.0570
β_{1996}	-0.2967	1.0221	13.6158	717.0570
β_{1997}	-0.8693	1.0253	12.7755	717.0570
β_{1998}	-0.8052	1.0274	13.1168	717.0570
β_{1999}	-1.2281	1.0314	12.7374	717.0570
β_{2000}	-1.0111	1.0278	12.9589	717.0570
β_{2001}	-0.2148	1.0213	13.8331	717.0570
β_{2002}	-0.8623	1.0240	13.1580	717.0570
β_{2003}	-1.5779	1.0292	12.5660	717.0570
β_{2004}	-0.8273	1.0235	13.1943	717.0570
β_{2005}	-0.7746	1.0235	13.3154	717.0570
α_0	1.3175	0.0130		
α_{years}	-0.0088	0.0015		
σ_u	0.3144	0.0098	0.2628	
σ_a	0.7109	0.0004		

A3 R Code for fitting Approximate MCEM with Censored Covariates

```
#### MODEL #####
# y|x,u ~ Poisson(mu)
# log(mu) = beta[0] + beta[1]*age + beta[2]*age^2 + u
# u ~ N(0,sdu)
# where age = Z %*% age.first + years.since.first
# where age.first is the true age at first observation (unknown for the censored owls)
# log(age.first|years.of.obs) ~ N(alpha[0]+alpha[1]*years.of.obs,sda)


# loading necessary libraries

library(mvtnorm)

n <- length(male)

# PUT OWL-SPECIFIC VARIABLES IN m-VECTORS

malem   <- unique(male)       # mx1 vector of unique male IDs
m       <- length(malem)      # 118 # number of males
age.first.obs  <- rep(0,m)    # age at first observation or lower bound, for each male
censorm <- rep(0,m)           # mx1 vector: 1 if age is censored for owl j, 0 if not
years.of.obs   <- rep(0,m)           # total years of observation on owl j
number.of.obs  <- rep(0,m)           # number of observations on owls j; j = 1,...,m

Z <- matrix(rep(0,n*m),n,m)   # Design matrix with male indicators

for (j in 1:m) {
        age.first.obs[j]      <- min(age[male==malem[j]])
        censorm[j]            <- mean(censor[male==malem[j]])
        years.of.obs[j]       <- max(age[male==malem[j]])-min(age[male==malem[j]])
        number.of.obs[j]      <- length(fledged[male==malem[j]])
        Z[,j]          <- ifelse(male==malem[j],1,0)
}

mc <- m - sum(censorm)        # 49  # males with known ages
m-mc                                         # 69 with censored ages
years.since.first <- age - Z %*% age.first.obs # Years since first observation (nx1
                                               # vector)

nc1  <- length(male[censor==1])        # number of observations from censored age owls
mc1   <- m-mc                          # number of censored age owls
Zc1 <-Z[(n-nc1+1):n, (mc+1):m]         # the subset of the Z matrix that deals with
                                      # the censored age owls
####################################################################################


#### METROPOLIS-HASTINGS SIMULATION FUNCTION ####

### UTILITY FUNCTIONS ###

# Function to generate left-truncated normals

        rtnorm  <- function(mu,sd,lo) {
        n              <- length(mu)
        mu + sd*qnorm(runif(n,pnorm((lo-mu)/sd),1))
}


# Function to generate random ages given lower bound, from truncated lognormal

acandidate <- function(mua,sda,lowest.age) {
        # mua is mx1 vector of means for log age at first observation
        # sda is 1x1 constant standard deviation for log age at first observation
        # lowest.age is mx1 vector of lower bounds for age
        m <- length(mua)
        lacandidate    <- rtnorm(mua,sda,log(lowest.age)) # simulated log ages at first
                                                          # obs
        exp(lacandidate) # simulated ages (not logged aged) at first observation
        }

metropolis <- function(theta, y.adjust, censor, age.first.obs, years.since.first,
R=1000) {


  A <- matrix(0,nrow=R,ncol=mc1) # Rxm matrix to be filled in with simulated ages at
                                 # first obs.
        alpha  <- theta$alpha
        beta   <- theta$beta
        sdu    <- theta$sdu
        sda    <- theta$sda


   sigma <- theta.old$sdu^2*Zc1%*%t(Zc1) + theta.old$sd^2*diag(nc1)   # calculating
                                                                    # the covariance
                                                                    # matrix for the
                                                                    # adjusted y-values
                                                                    # calculated using
                                                                    # the previous sd.u
                                                                    # and sd values
                                                                    # from lmer

        # Initialize
        crita  <- rep(0,mc1)  # to store the acceptance criterion for candidates for
                              # age at first observation

        Years.since <- matrix(rep(years.since.first[censor==1],mc1),nc1,mc1)
                                             # nc1x  mc1; mc1 copies of years since
                                             # first observation

        mua <- alpha[1] + alpha[2]*years.of.obs[censorm==1] # mean of lognormal
                                                     # regression for censored ages

        A[1,]  <- acandidate(mua,sda,age.first.obs[censorm==1])
               #  Simulated values from the truncated lognormal regression.

        for (r in 2:R) {

               # Generate candidate vector for a
               a.candidate <- acandidate(mua,sda,age.first.obs[censorm==1])

               ### CALCULATE DENOMINATOR OF ACCEPTANCE CRITERION VECTORS ####
               age <- Zc1 %*% A[r-1,] + years.since.first[censor==1]       # nc1x1;
                                                            # age = "age at first
                                                            # observation" + years
                                                            # since first

               fixed   <- beta[1] + beta[2]*age + beta[3]*(age^2) # nx1; fixed
                                                                  # effects in linear
                                                                  # predictor using
                                                             # previous simulated a's


        Density  <- rep(dmvnorm(as.vector(y.adjust[censor==1]),fixed, sigma), mc1)
                              # multivariate normal density at the linear predictor

               denom <- exp(Density)


               ### GET ACCEPTANCE VECTOR FOR A ###

               amat    <- matrix(rep(A[r-1,],mc1),mc1,mc1)   # mc1 x mc1 ; mc1 copies
                                                            # of previously simulated
                                                     # a's (age at first observation)
               diag(amat)     <- a.candidate  # mc1 x mc1;  Replace diagonal elements
                                              # with new candidates
               Age <- Zc1 %*% amat + Years.since # nc1 x mc1; Form ages from ages at
                                     # first observation plus years since first

               fixed   <- beta[1] + beta[2]*Age + beta[3]*(Age^2)


    y.adjust.standard <- y.adjust[censor==1]-fixed   # dmvnorm only allows matrices
                                                     # for the x, not mu


    Density    <- dmvnorm(t(y.adjust.standard),rep(0,nc1), sigma)
               # nm x 1; multivariate normal density at the linear predictor

               numer   <- exp(Density)

               crita   <- ifelse(numer<denom,numer/denom,1)  # mc1 x1; jth element is
                                                            # the acceptance
                                                            # probability for
                                                            # candidate aj

               ### CARRY OUT ACCEPTANCE ###

               berna <- rbinom(mc1,1,crita)  # mc1 x1; generate mc1 Bernoullis with
                                     # the candidate a acceptance probabilities

               A[r,] <- berna*a.candidate + (1-berna)*A[r-1,]  # mc1 x 1; new a is
                                                               # either previous a or
                                                               # new candidate
                      }


  list(A=A)                                  # Return R simulated a vectors
}

####################################################################################
```

```
#######################################################################################

# Likelihood function used to compute importance weights

like   <- function(theta) {

    # replicating the datasets for censored owls

    malevector2 <- rep(male[censor==1], rep(R,nc1))
    fledgedvector2 <-  rep(fledged[censor==1],rep(R,nc1))
    malem2 <- malem[censorm==1]
    X.mat <- cbind(rep(1,length(age.metrop)),age.metrop,age.metrop^2)
    Z.mat <- matrix(rep(0,R*nc1*mc1),R*nc1,mc1)      # Design matrix with male
                                                     # indicators
    for (j in 1:mc1) {
             Z.mat[,j] <- ifelse(malevector2==malem2[j],1,0)
      }

      # calculating the linear predictor, adjusted y, and the multivariate normal
      # density
      lin.pred <- X.mat%*%theta$beta+ Z.mat%*%u.old[censorm==1]

      y.adjust <- matrix(lin.pred +
(fledgedvector2exp(lin.pred))/exp(lin.pred),nrow=R,
                       ncol=nc1, byrow=F)

      Xbeta <- matrix(X.mat%*%theta$beta,nrow=R, ncol=nc1, byrow=F)
      y.adjust.standard <- y.adjust-Xbeta
      sigma <- theta$sdu^2*Zc1%*%t(Zc1) + theta.old$sd^2*diag(nc1)
      s1 <- dmvnorm(y.adjust.standard,rep(0,nc1) , sigma )

      # calculating likelihood for the log first age part of the model

      avector2 <- as.vector(matrix(sample$A, ncol=1, byrow=T))
      years.of.obsvector2 <-rep(years.of.obs[censorm==1] ,rep(R,mc1))

      Density2 <-
dnorm(log(avector2),theta$alpha[1]+theta$alpha[2]*years.of.obsvector2,
        theta$sda)

        MCnum2 <- rep(1:R,mc1)

    s2 <- tapply(Density2, INDEX=MCnum2, sum)

    return(s1+s2)
}

#######################################################################################
# loading necessary libraries
#install.packages("lme4", repos = "http://r-forge.r-project.org")

library(MASS)
library(nlme)
library(Matrix)
library(lme4)




# approximate MCEM algorithm for censored ages



###############################

start.time <-date()              # time that algorithm starts to see how long until
                                 # convergence
```

```
convergence.criterion <- 0.005   # converence criterion
iter <- 1                        # iteration number
maxdif <- 1                      # initial setting for max abs. value
c <-3                            # amount that the Monte Carlo sample size is
                                 # increased by: R <-R+ floor(R/c)



# STARTING VALUES


beta.old <- c(-0.70412947,  0.23793109, -0.01462246) # from glmmPQL for owls
                                                     # with known ages
alpha.old<- c(0.553586294,  0.005887441) # from lm fit of log agefirst on yearsobsm
                                        # from owls with known ages
sdu.old <-  0.2268224
sda.old <- 0.4560279
theta.old <- list(beta=beta.old, alpha=alpha.old, sdu=sdu.old, sda=sda.old)

R <- 8  # initial R value--this actual makes the intial R = 10, since 8+floor(8/3) =10
u.old <- rep(0,m)


age2 <- age^2          # creating the X matrix that will be used in calculating
X <- cbind(rep(1,n), age, age2)      # the adjusted y values for the approx


print(c("iter", "beta0", "beta1", "beta2","alpha1", "alpha2",  "sdu", "sd", "sda"),
quote=F, sep="\t")
print(round(c(0, unlist(theta.old)),3), sep="\t")


# keep track of the estimates


beta1.iter <- beta.old[1]
beta2.iter <- beta.old[2]
beta3.iter <- beta.old[3]
alpha1.iter <-alpha.old[1]
alpha2.iter <- alpha.old[2]
sdu.iter   <- sdu.old
sd.iter   <- sd.old
sda.iter   <- sda.old
maxdif.iter <- maxdif
R.iter <- R



S <- 3 # burn in for regular MCEM before switching to importance weighting

##############################
while (maxdif>convergence.criterion) {


      # Monte Carlo E-step

      # calculating the adjusted response

      lin.pred <- X%*%theta.old$beta+ Z%*%u.old

      y.adjust <- lin.pred + (fledged-exp(lin.pred))/exp(lin.pred)


      # Monte Carlo E-step
        R0 <- R
        R <- R0 + floor(R0/c)

       # Obtain R psuedo random variables using the Metropolis Alg
```

```
  if (iter<=S) {
               sample <- metropolis(theta.old,y.adjust,censor,age.first.obs,
                         years.since.first, R=R)
               w <- 1
               imp.weight1 <- c(rep(1,n-nc1), rep(1,nc1*R)/R)
               imp.weight2 <- c(rep(1,mc),rep(1,mc1*R)/R)
                }

  if (iter==S) {
               theta.init <- theta.old
               sample <- metropolis(theta.old,y.adjust,censor,age.first.obs,
                 years.since.first, R=R)
               }



 if (iter >S){
     samp.incr <-floor(R0/c)
     samplek <- metropolis(theta.old,y.adjust,censor,age.first.obs,
                 years.since.first, R=samp.incr)
     sample$A <-matrix(c(t(sample$A),t(samplek$A)), nrow=R0+samp.incr, byrow=T)
         }



 # replicating for the augmented data set for the n observations

 malevector <- c(male[censor==0],rep(male[censor==1], rep(R,nc1)))
 age.metrop <- matrix(t(Zc1 %*%t(sample$A)), ncol=1, byrow=T)+
               rep(years.since.first[censor==1],rep(R,nc1))
 agevector <- c(age[censor==0], age.metrop)
 agevector2 <- agevector^2
 fledgedvector <-  c(fledged[censor==0],rep(fledged[censor==1],rep(R,nc1)))






# replicating for the augmented data set for each owl


 avector <- c(age.first.obs[censorm==0], as.vector(matrix(sample$A, ncol=1,
 byrow=T)))     # making the a matrix into an m*r vector
 years.of.obsvector <-c(years.of.obs[censorm==0],rep(years.of.obs[censorm==1]
                         ,rep(R,mc1)))


   # compute the importance weights as in Levine and Casella

 if (iter >=S) {w <- (like(theta.old)/like(theta.init))
                imp.weight1 <- c(rep(1,n-nc1), rep(w,nc1)/sum(w))
                imp.weight2 <- c(rep(1,mc),rep(w,mc1)/sum(w))
               }
 # M-step

   fit1 <-lmer(formula= fledgedvector~agevector+agevector2
          +(1|malevector) ,weights=imp.weight1, family=poisson,
          start=list(malevector=matrix(sdu.old ,1,1)) )

 beta.new <- fixef(fit1)                           # esimated beta's
 sdu.new <- sqrt(as.vector(VarCorr(fit1)$male))    # standard devation of the
                                                    # random effects distribution

 u.new <-as(ranef(fit1)$male[,1], "vector")        # random effects coef's, need
                                                    # these to calculate the next
                                                    # y-adjusted values
```

```
      ## fitting the model for the age of first observation


      fit2 <- lm(log(avector)~ years.of.obsvector, weights=imp.weight2  )
      alpha.new  <-fit2$coef                # esimated alpha's

     sda.new  <- summary(fit2)$sigma*sqrt(fit2$df.residual)/(sqrt(m-2))   # estimated
                                                                  # sd for the first
                                                                  # observed ages

      ## Setting the new estimates

      theta.new <- list(beta=beta.new, alpha=alpha.new , sdu=sdu.new, sda=sda.new)



      # check convergence
      maxdif <- max(abs((unlist(theta.new)-unlist(theta.old))/unlist(theta.old)))


      # print current estimate

      print(round(c(iter, unlist(theta.new), maxdif),3), sep="\t")


      # keep track of the estimates

      beta1.iter <- c(beta1.iter,beta.new[1])
      beta2.iter <- c(beta2.iter,beta.new[2])
      beta3.iter <- c(beta3.iter,beta.new[3])
      alpha1.iter <- c(alpha1.iter,alpha.new[1])
      alpha2.iter <- c(alpha2.iter,alpha.new[2])
      sdu.iter   <- c(sdu.iter,sdu.new)
      sda.iter   <- c(sda.iter,sda.new)
      R.iter <- c(R.iter,R)
      maxdif.iter <-c(maxdif.iter,maxdif)



      # replacing old estimates with the new ones

      iter <- iter+1

      beta.old <- beta.new
      alpha.old <- alpha.new
      sdu.old  <- sdu.new
      sda.old  <- sda.new
      u.old <- u.new
      theta.old <- theta.new



      }
end.time <-date()      # obtaining the ending time to see how long until convergence

# print mle
mle <- theta.old
mle
round(unlist(mle),4)


######################################################################################
```

```
# Calculating the standard errors


# calculating the adjusted y-values
X.mat <- cbind(rep(1,length(agevector)),agevector, agevector2)
Z.mat   <- matrix(rep(0,length(malevector)*m),length(malevector),m) # Design matrix
                                                                    # with male
                                                                    # indicators
    for (j in 1:m) {
            Z.mat[,j]                        <- ifelse(malevector==malem[j],1,0)
      }
lin.pred <- X.mat%*%mle$beta+ Z.mat%*%u.old
y.adjust <- lin.pred + (fledgedvector-exp(lin.pred))/exp(lin.pred)
Rvector <-c(rep(0,n-nc1),rep(1:R, nc1))
censorvector <-  c(censor[censor==0],rep(censor[censor==1],rep(R,nc1)))
yearvector <-c(year[censor==0],rep(year[censor==1],rep(R,nc1)))
d4 <- data.frame(malevector,agevector,agevector2, y.adjust,Rvector,
censorvector,yearvector)

d5 <- orderBy(~censorvector+malevector+Rvector, data=d4)
attach(d5)

X.mat <- cbind(rep(1,length(agevector)),agevector, agevector2)


# Calculating first derivative information for known age owls

bmc <- matrix(rep(0,3*(mc)),3,mc)
sumc <- rep(0,mc)
for (j in 1:mc){
    nobs <- length(y.adjust[malevector==malem[j]])
    V.inv <- solve(mle$sdu^2*matrix(rep(1,nobs^2),nobs,nobs)+diag(nobs))
    X.matrix <- matrix(X.mat[malevector==malem[j]], ncol=3)
    y.adjust.vect <- y.adjust[malevector==malem[j]]

    bmc[,j] <- t(X.matrix)%*%V.inv%*%y.adjust.vect-
t(X.matrix)%*%V.inv%*%X.matrix%*%mle$beta
    sumc[j] <- .5*t(y.adjust.vect-
X.matrix%*%mle$beta)%*%V.inv%*%matrix(rep(1,nobs^2),nobs,nobs)%*%V.inv%*%(y.adjust.vec
t-X.matrix%*%mle$beta)-sum(diag(V.inv%*%matrix(rep(1,nobs^2),nobs,nobs)))
}


# Calculating first derivative information for censored age owls

bmc1 <- matrix(rep(0,3*(mc1*R)),3,(mc1*R))
sumc1 <- rep(0,mc1*R)
for (j in (mc+1):m){
        for (r in 1:R){
            index <- (j-mc-1)*R+r
            nobs <- length(y.adjust[malevector==malem[j] & Rvector==r])
            V.inv <- solve(mle$sdu^2*matrix(rep(1,nobs^2),nobs,nobs)+diag(nobs))
            X.matrix <- matrix(X.mat[malevector==malem[j]& Rvector==r], ncol=3)
            y.adjust.vect <- y.adjust[malevector==malem[j]& Rvector==r]

            bmc1[,index] <- t(X.matrix)%*%V.inv%*%y.adjust.vect-
t(X.matrix)%*%V.inv%*%X.matrix%*%mle$beta
            sumc1[index] <- .5*t(y.adjust.vect-
X.matrix%*%mle$beta)%*%V.inv%*%matrix(rep(1,nobs^2),nobs,nobs)%*%V.inv%*%(y.adjust.vec
t-X.matrix%*%mle$beta)-sum(diag(V.inv%*%matrix(rep(1,nobs^2),nobs,nobs)))
      }
}



b0 <- c(bmc[1,], bmc1[1,])
b1 <- c(bmc[2,], bmc1[2,])
b2 <- c(bmc[3,], bmc1[3,])
su <- c(sumc, sumc1)
```

```
# Calculating first derivative information for log age at fist obs

a0 <- (log(avector)-mle$alpha[1]-mle$alpha[2]*years.of.obsvector)/mle$sda^2
a1 <- years.of.obsvector*( log(avector)-mle$alpha[1]-mle$alpha[2]*
years.of.obsvector)/mle$sda^2
sa <- ((-1/mle$sda) +((log(avector)-mle$alpha[1]- mle$alpha[2]*
years.of.obsvector)^2/mle$sda^3))


# Calculating the approximate info
info <-  matrix(rep(0,49),nrow=7)
weight <-  c(rep(1,mc),rep(1/R,mc1*R))
for (i in 1:(mc+mc1*R)){

        info <- info +weight[i]*c(b0[i],b1[i],b2[i], a0[i], a1[i], su[i],
sa[i])%*%t(c(b0[i],b1[i],b2[i], a0[i], a1[i], su[i], sa[i]))


}


var.cov <-solve(info)
se <- sqrt(diag(var.cov))
se
```

A4 R Code for fitting Regression Calibration with Censored Covariates

```
# loading necessary libraries
library(lme4)

# fitting a regression model for the log age of first obs for known age owls


fit <- lm(log(age.first.obs)~ years.of.obs, subset=censorm==0  )
sda <- summary(fit)$sigma
alpha <- fit$coef


# calculating the new age based on the regression output and conditional
# on the lower bound
mu.a <- alpha[1] + alpha[2]*years.of.obs[censorm==1]

expect.age <- mu.a+  sda*dnorm((log(age.first.obs[censorm==1])-mu.a)/sda)/(1-
pnorm((log(age.first.obs[censorm==1])-mu.a)/sda))
age.first.new <- c(age.first.obs[censorm==0], exp(expect.age))


age.new <-  Z %*% age.first.new + years.since.first
age.new2 <-   age.new^2


# fitting with lmer

yearfactor <-as.factor(year)
fit.replace.expected <-lmer(formula= fledged ~  age.new + age.new2+  yearfactor +
(1|male), family=poisson)
fit.replace.expected
```

A5 Simulation Details

The simulation study looked at six different setting of sample size and proportion of censored observations. The sample sizes used were 50 and 400 owls. The proportion of censored observations that were studied was 10%, 25%, and 50%. We used a factorial structure with a Monte Carlo sample size of 200 for each setting. These setting were chosen to give a range of possible situations that would be encountered by the biologists and to match the setting for the Northern Spotted Owl study.

After randomly selecting known age owls from the Spotted Owl study, we randomly generated random effects from a $N\left(0,\sigma_u^2\right)$ distribution, with $\sigma_u = 0.3$. The linear predictor was then found using the values of $\beta_0 = \text{-}1.8070$, $\beta_1 = 0.3855$, and $\beta_2 = \text{-}0.0235$. These values all roughly match the estimated values for the Spotted Owl study. We then randomly sampled a percentage of the owls to have censored ages. Again to match the Spotted Owl study, we used 3 years as the lower bound for the first observed age for the censored owls.

The resulting data sets were then analyzed using (1) maximum likelihood estimator via the MCEM algorithm for censored covariates, (2) the approximate MCEM estimator, (3) the regression calibration estimator, (4) the naïve estimator in which censored ages are replaced by their lower bounds (to demonstrate the unsuitability of this approach, which may seem tempting to wildlife biologists), and (5) the GLMM estimator using only owls with known ages. To avoid computational difficulties resulting from large Monte Carlo sample sizes in the MCEM algorithms, convergence of these algorithms was either relative parameter convergence of 0.5% or 11 total iterations.

The starting values for the approximate MCEM algorithm were set as the final estimates from the regression calibration model. In addition, the starting values for the MCEM algorithm were the final estimates from the approximate MCEM algorithm.

A6 Simulation Results

Table A6.1 Descriptive statistics of estimates of β_0 (true value -1.8070)

Sample size	50			400		
Proportion censored	.10	.25	.50	.10	.25	.50
MCEM						
mean	-1.8276	-1.7413	-1.6524	-1.7660	-1.7077	-1.6082
bias	-0.0206	0.0657	0.1546	0.0410	0.0993	0.1988
variance	0.2605	0.2623	0.2164	0.0321	0.0260	0.0214
MSE	0.2598	0.2654	0.2393	0.0336	0.0357	0.0608
Monte Carlo SD	0.5104	0.5122	0.4652	0.1791	0.1612	0.1463
Mean reported SE	0.5327	0.5114	0.4874	0.1637	0.1612	0.1559
Approx MCEM						
mean	-1.8580	-1.7536	-1.6624	-1.7786	-1.7145	-1.6194
bias	-0.0510	0.0534	0.1446	0.0284	0.0925	0.1876
variance	0.2573	0.2522	0.2002	0.0317	0.0254	0.0197
MSE	0.2588	0.2539	0.2201	0.0324	0.0339	0.0548
Monte Carlo SD	0.5073	0.5022	0.4474	0.1781	0.1594	0.1403
Mean reported SE	0.2771	0.2672	0.2552	0.0787	0.0775	0.0759
Regression Calibration						
mean	-1.8897	-1.8489	-1.8585	-1.8053	-1.7973	-1.7905
bias	-0.0827	-0.0419	-0.0515	0.0017	0.0097	0.0165
variance	0.2774	0.2899	0.2643	0.0329	0.0291	0.0265
MSE	0.2830	0.2903	0.2656	0.0328	0.0291	0.0267
Monte Carlo SD	0.5266	0.5384	0.5141	0.1815	0.1705	0.1628
Mean reported SE	0.5204	0.5225	0.5415	0.1777	0.1807	0.1845
Naïve Replace						
mean	-1.7381	-1.5504	-1.3969	-1.6718	-1.4952	-1.3311
bias	0.0689	0.2566	0.4101	0.1352	0.3118	0.4759
variance	0.2358	0.2103	0.1728	0.0284	0.0216	0.0140
MSE	0.2395	0.2751	0.3402	0.0465	0.1188	0.2404
Monte Carlo SD	0.4856	0.4586	0.4157	0.1684	0.1471	0.1182
Mean reported SE	0.4831	0.4424	0.3964	0.1655	0.1521	0.1360
Known Only						
mean	-1.9378	-1.8681	-1.9470	-1.8252	-1.8299	-1.8438
bias	-0.1308	-0.0611	-0.1400	-0.0182	-0.0229	-0.0368
variance	0.3130	0.3802	0.6864	0.0374	0.0384	0.0662
MSE	0.3288	0.3821	0.7026	0.0376	0.0388	0.0672
Monte Carlo SD	0.5595	0.6166	0.8285	0.1933	0.1960	0.2573
Mean reported SE	0.5486	0.5942	0.7542	0.1866	0.2053	0.2522

Table A6.2 Descriptive statistics of estimates of β_1 (true value 0.3855)

Sample size		50			400	
Proportion censored	.10	.25	.50	.10	.25	.50
MCEM						
mean	0.4065	0.3783	0.3543	0.3711	0.3547	0.3267
bias	0.0210	-0.0072	-0.0311	-0.0143	-0.0307	-0.0587
variance	0.0192	0.0181	0.0154	0.0021	0.0020	0.0016
MSE	0.0195	0.0180	0.0163	0.0023	0.0029	0.0050
Monte Carlo SD	0.1384	0.1345	0.1240	0.0457	0.0443	0.0394
Mean reported SE	0.1462	0.1387	0.1328	0.0435	0.0428	0.0413
Approx MCEM						
mean	0.4003	0.3658	0.3390	0.3751	0.3564	0.3282
bias	0.0149	-0.0197	-0.0465	-0.0104	-0.0290	-0.0573
variance	0.0186	0.0172	0.0137	0.0021	0.0019	0.0014
MSE	0.0188	0.0175	0.0158	0.0022	0.0027	0.0047
Monte Carlo SD	0.1364	0.1312	0.1170	0.0457	0.0433	0.0374
Mean reported SE	0.0757	0.0709	0.0664	0.0196	0.0189	0.0185
Regression Calibration						
mean	0.4091	0.3932	0.3937	0.3823	0.3796	0.3752
bias	0.0237	0.0077	0.0083	-0.0031	-0.0059	-0.0102
variance	0.0203	0.0203	0.0184	0.0022	0.0022	0.0019
MSE	0.0208	0.0202	0.0184	0.0022	0.0022	0.0020
Monte Carlo SD	0.1426	0.1424	0.1357	0.0468	0.0468	0.0437
Mean reported SE	0.1405	0.1406	0.1458	0.0474	0.0483	0.0492
Naïve Replace						
mean	0.3768	0.3310	0.3057	0.3555	0.3171	0.2855
bias	-0.0086	-0.0545	-0.0797	-0.0300	-0.0683	-0.1000
variance	0.0179	0.0160	0.0154	0.0019	0.0018	0.0013
MSE	0.0179	0.0189	0.0216	0.0028	0.0065	0.0113
Monte Carlo SD	0.1337	0.1265	0.1239	0.0437	0.0426	0.0356
Mean reported SE	0.1335	0.1258	0.1199	0.0452	0.0430	0.0407
Known Only						
mean	0.4234	0.4000	0.4268	0.3898	0.3904	0.3937
bias	0.0380	0.0145	0.0414	0.0044	0.0049	0.0082
variance	0.0223	0.0273	0.0507	0.0025	0.0027	0.0047
MSE	0.0236	0.0273	0.0522	0.0025	0.0027	0.0047
Monte Carlo SD	0.1492	0.1651	0.2252	0.0498	0.0519	0.0685
Mean reported SE	0.1484	0.1611	0.2061	0.0499	0.0550	0.0676

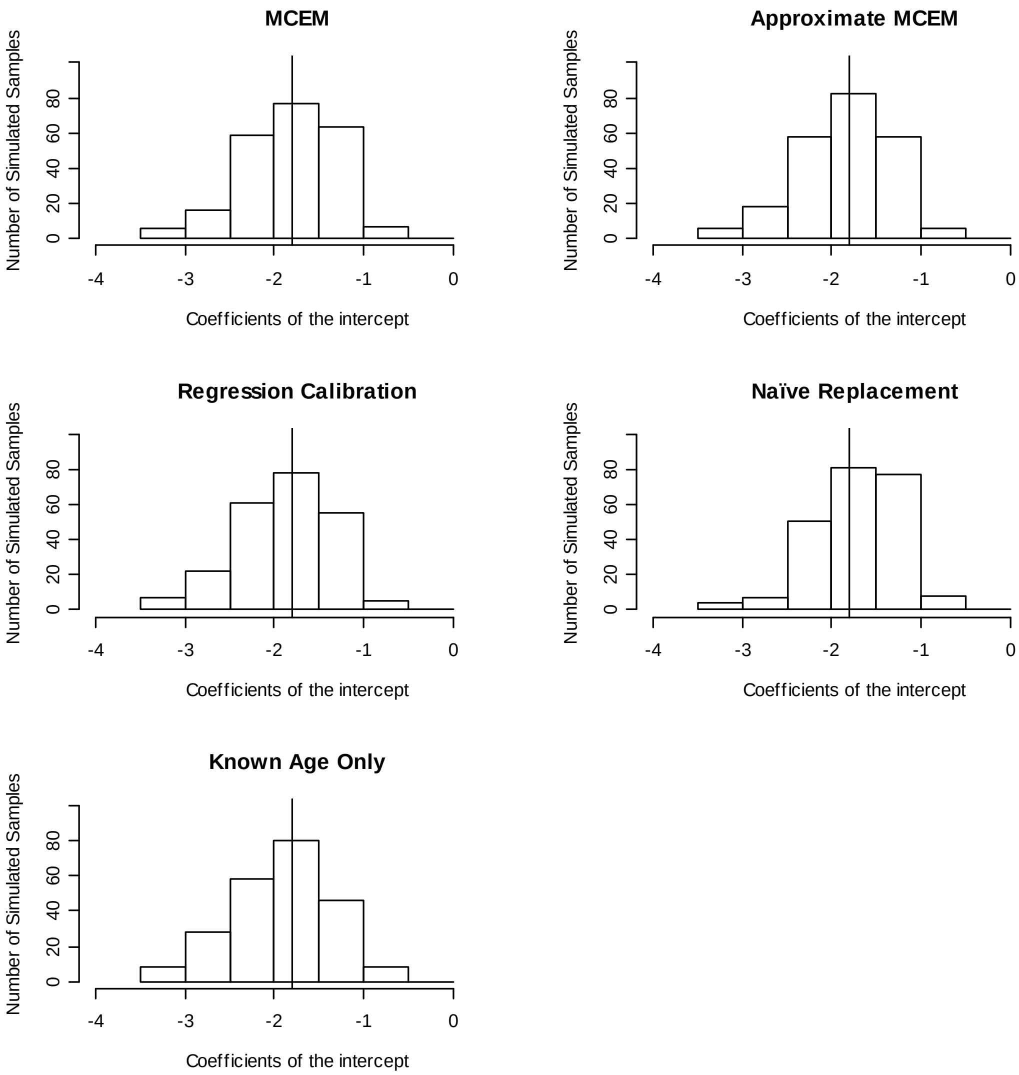


Figure A6.1 Monte Carlo sampling distributions for *n=50* and *10% censored* for the estimators of β_0, the intercept, for the 5 different fitting methods.

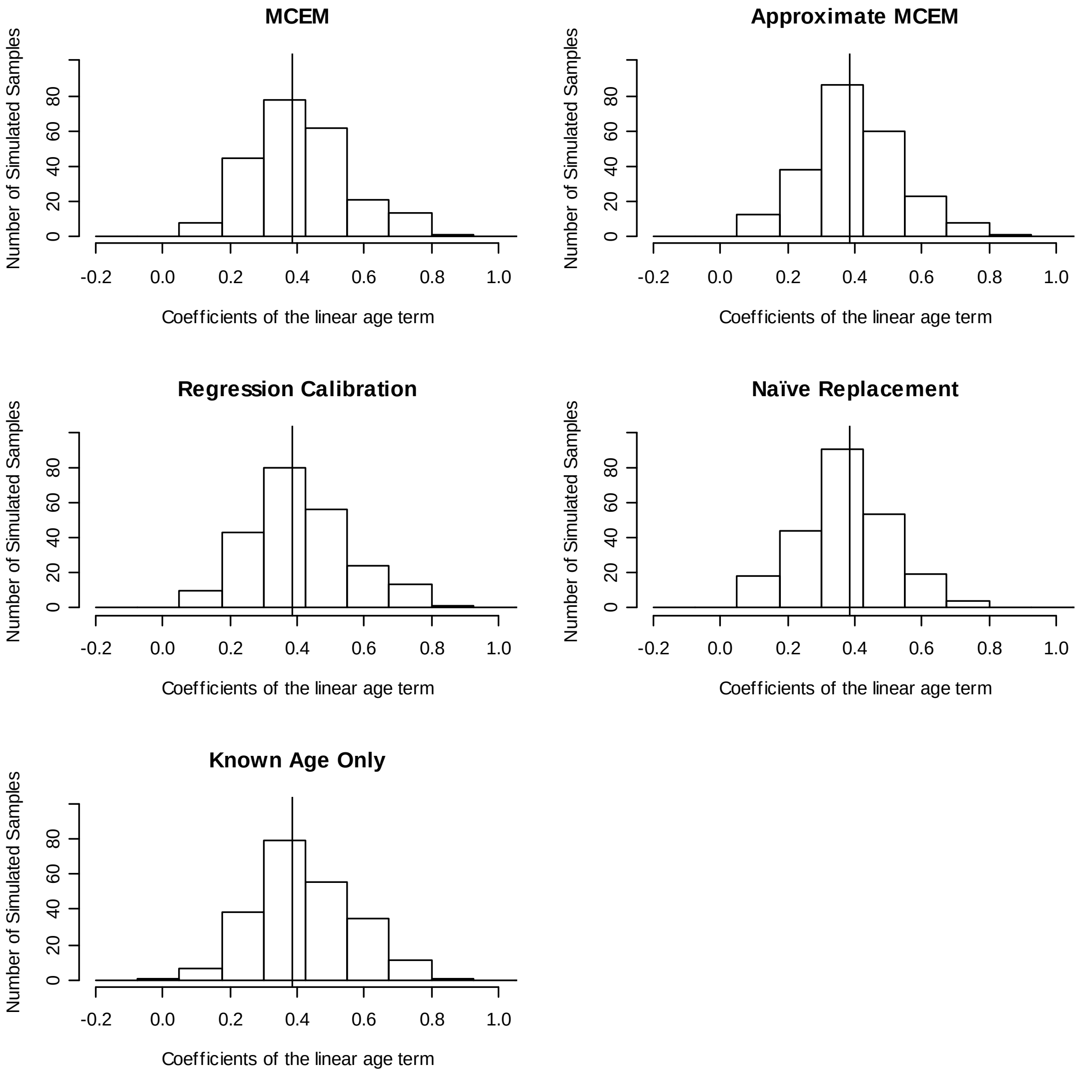


Figure A6.2 Monte Carlo sampling distributions for *n=50* and *10% censored* for the estimators of β_1, the age term, for the 5 different fitting methods.

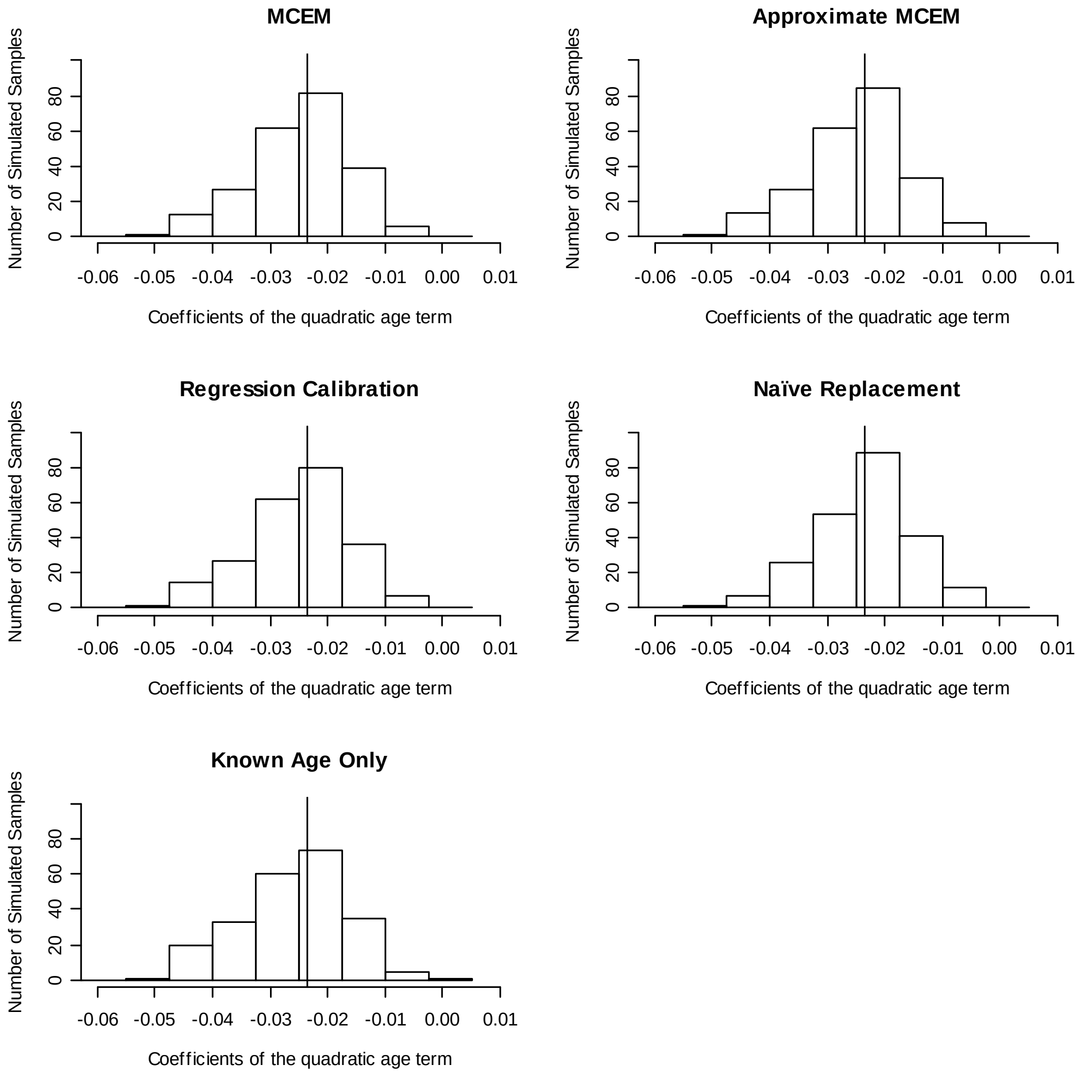


Figure A6.3 Monte Carlo sampling distributions for *n=50* and *10% censored* for the estimators of β_2, the age^2 term, for the 5 different fitting methods.

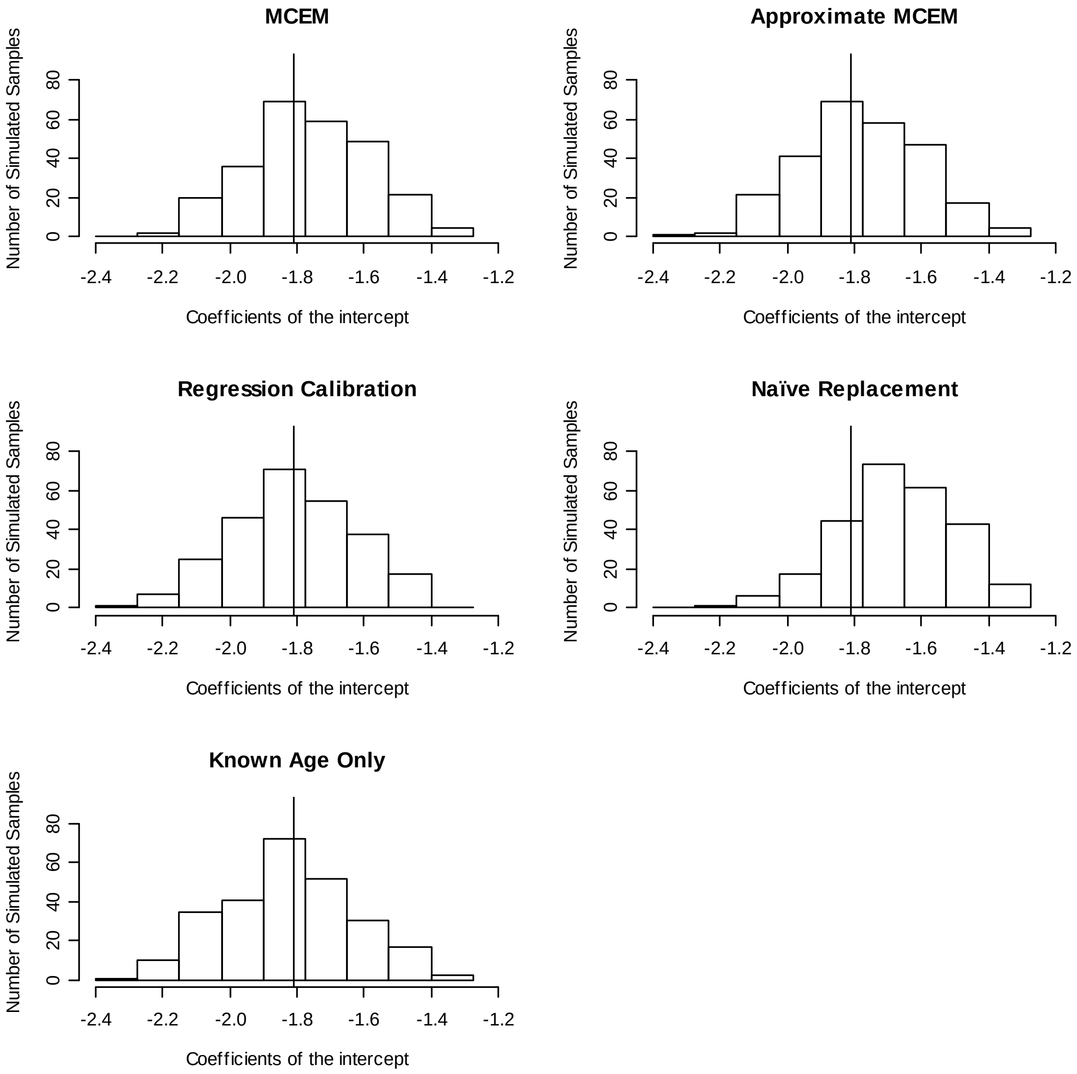


Figure A6.4 Monte Carlo sampling distributions for *n=400* and *10% censored* for the estimators of β_0, the intercept, for the 5 different fitting methods.

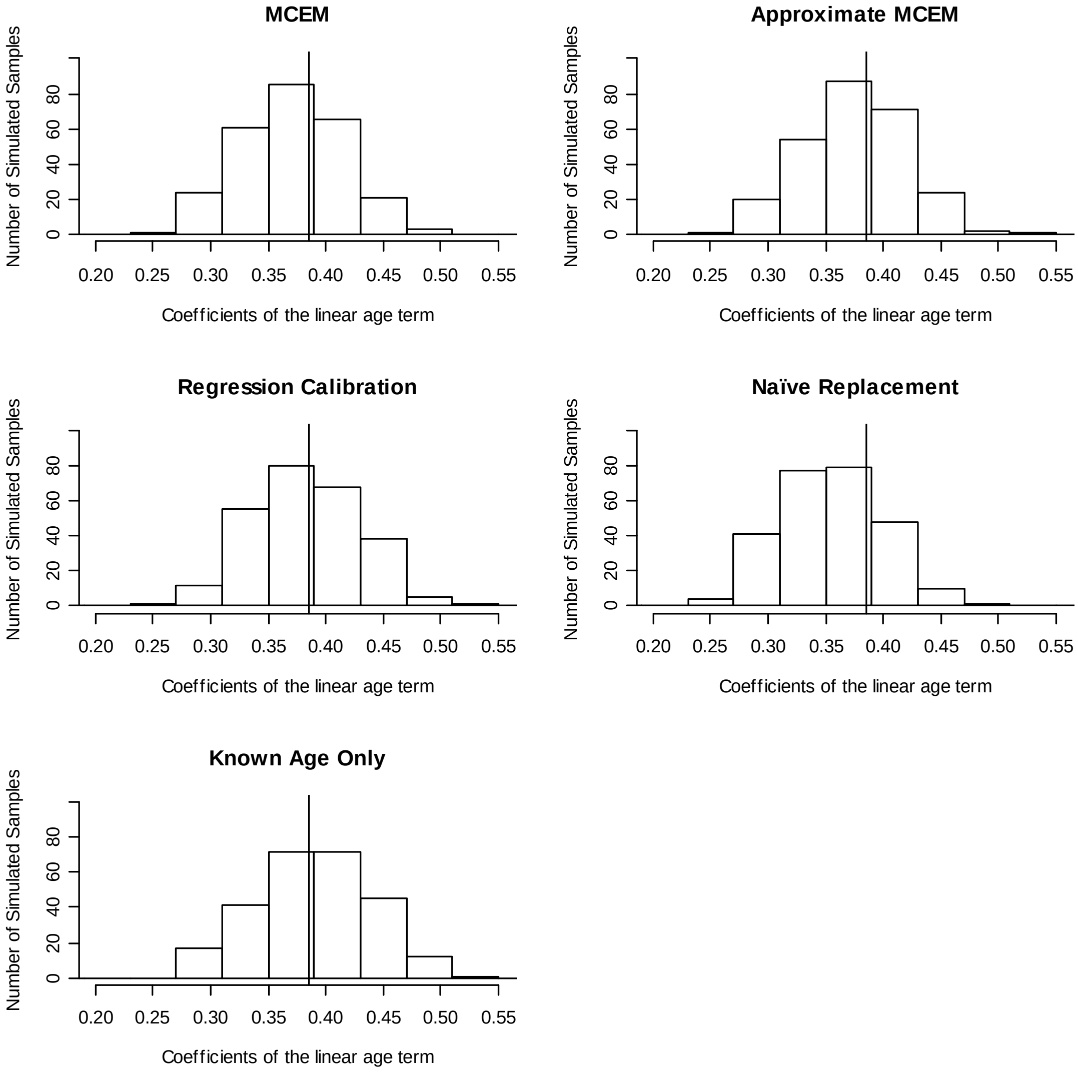
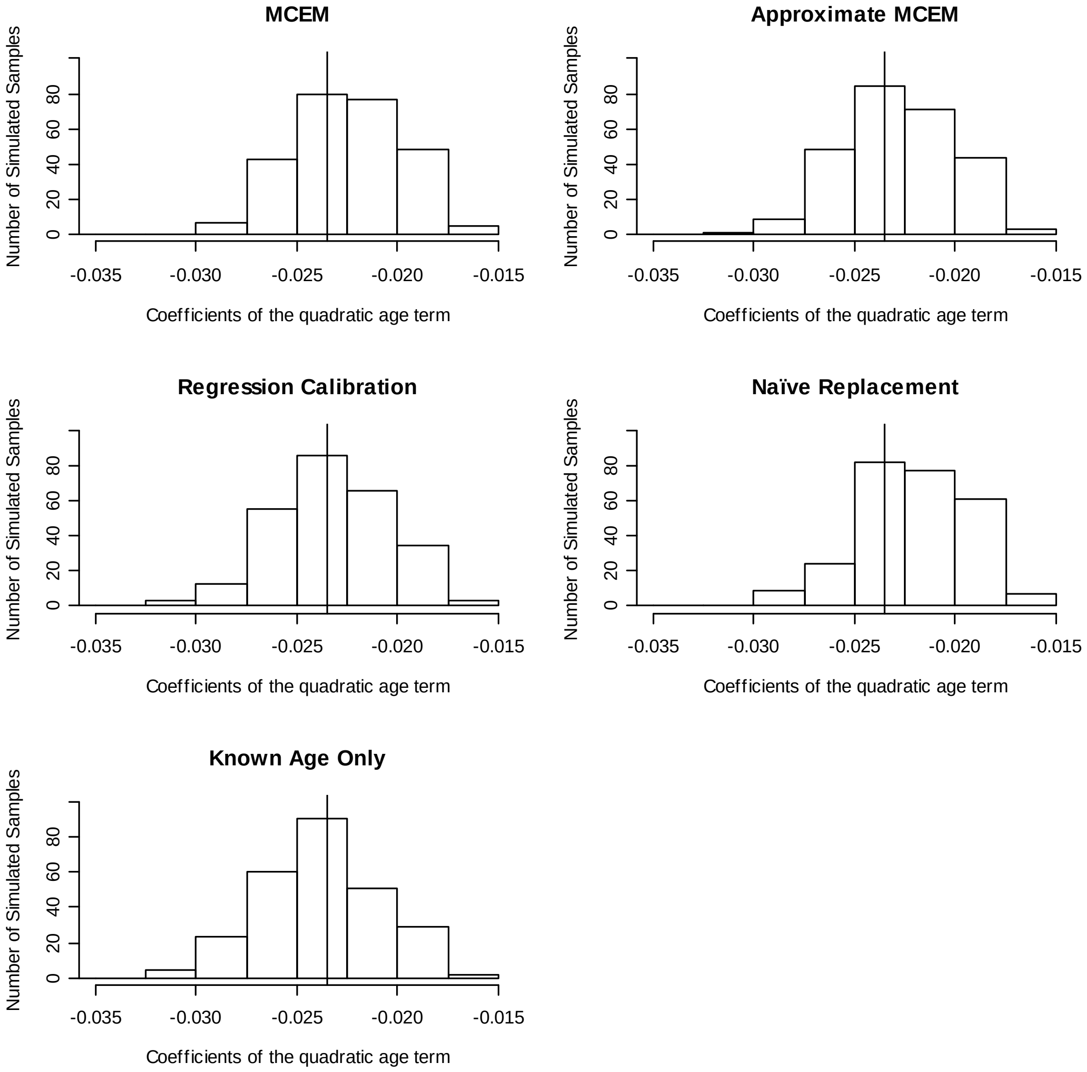


Figure A6.5 Monte Carlo sampling distributions for *n=400* and *10% censored* for the estimators of β_1, the age term, for the 5 different fitting methods.



MCEM

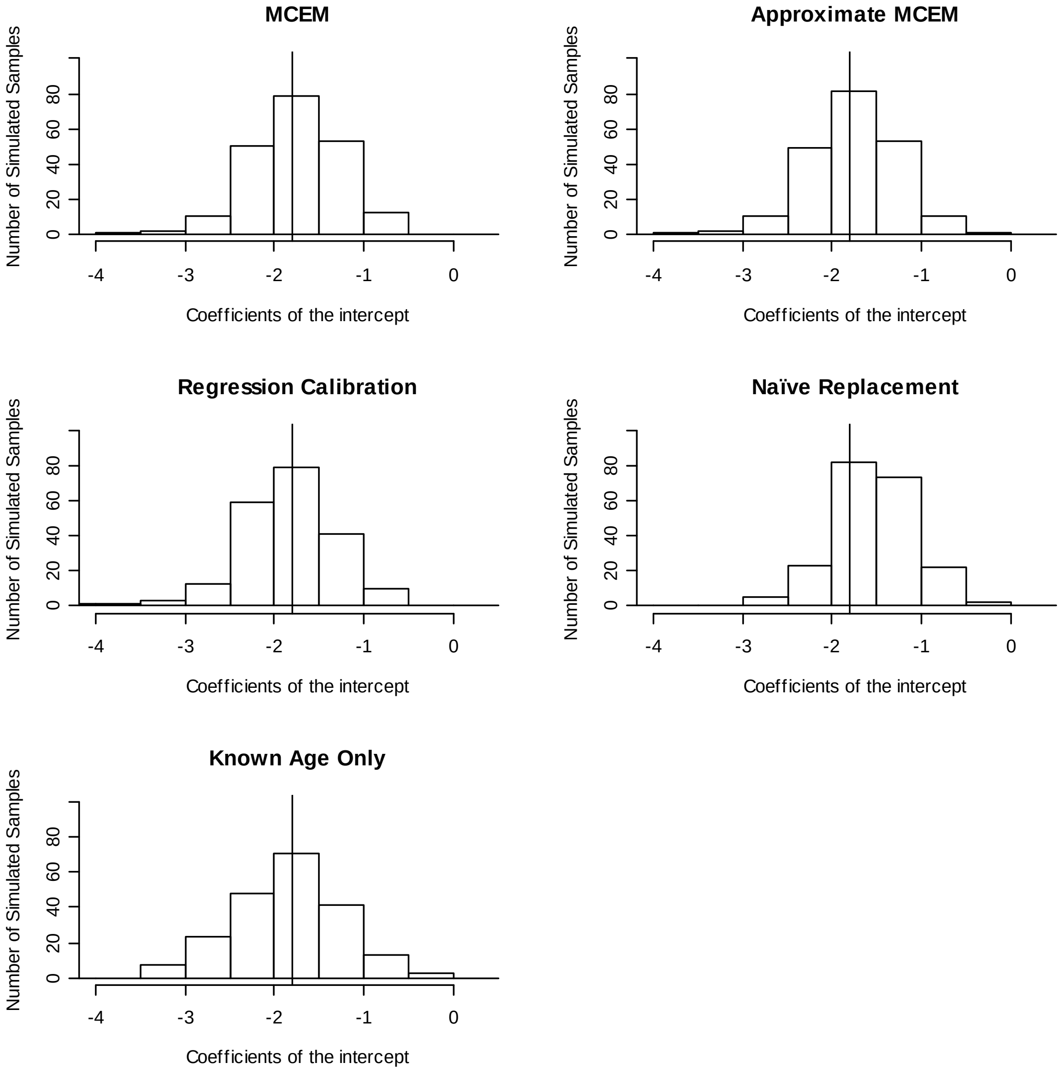
Approximate MCEM

Regression Calibration

Naïve Replacement

Known Age Only

Figure A6.6 Monte Carlo sampling distributions for *n=400* and *10% censored* for the estimators of β_2, the age^2 term, for the 5 different fitting methods.


MCEM

Approximate MCEM

Regression Calibration

Naïve Replacement

Known Age Only

Figure A6.7 Monte Carlo sampling distributions for *n=50* and *25% censored* for the estimators of β_0, the intercept, for the 5 different fitting methods.

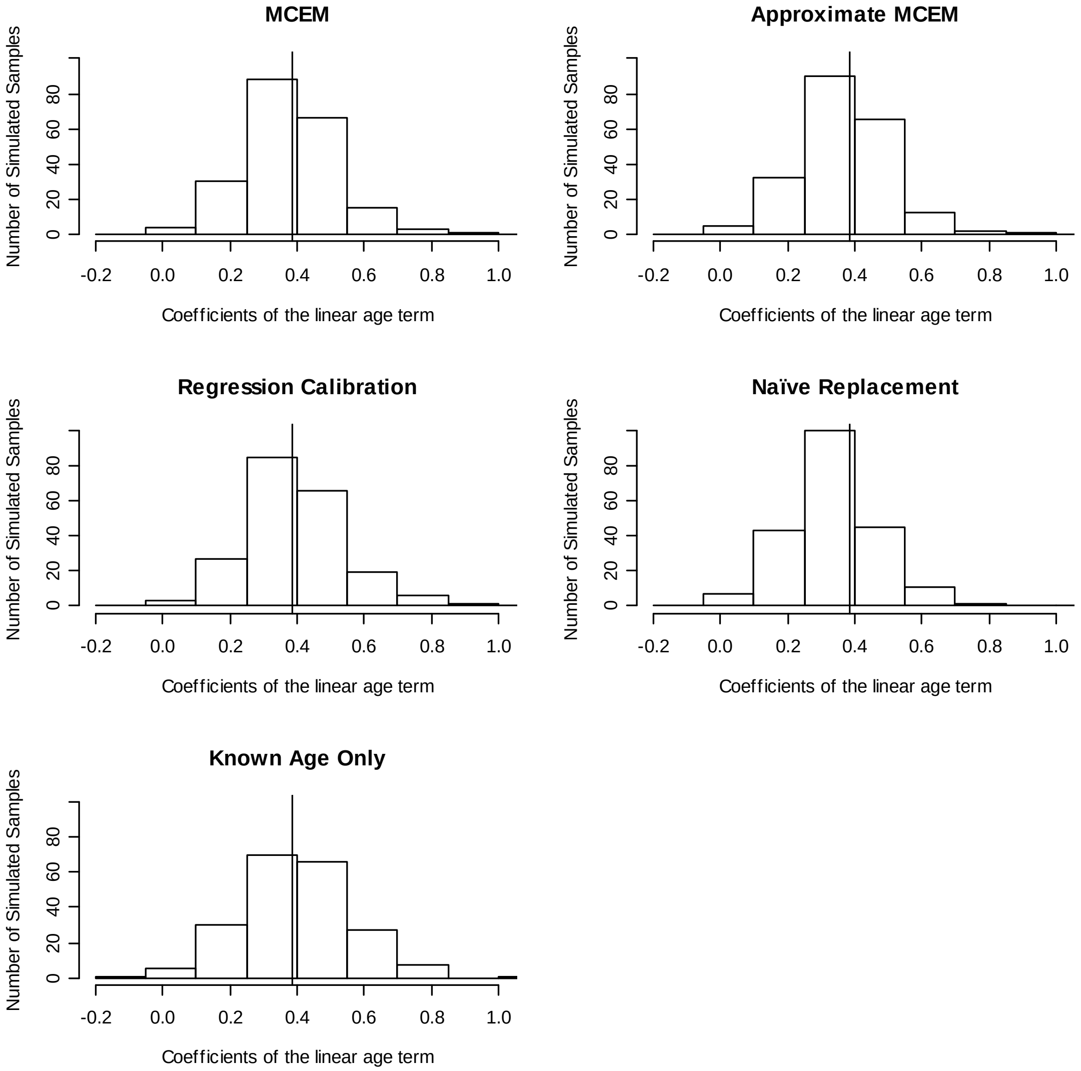


Figure A6.8 Monte Carlo sampling distributions for *n=50* and *25% censored* for the estimators of β_1, the age term, for the 5 different fitting methods.

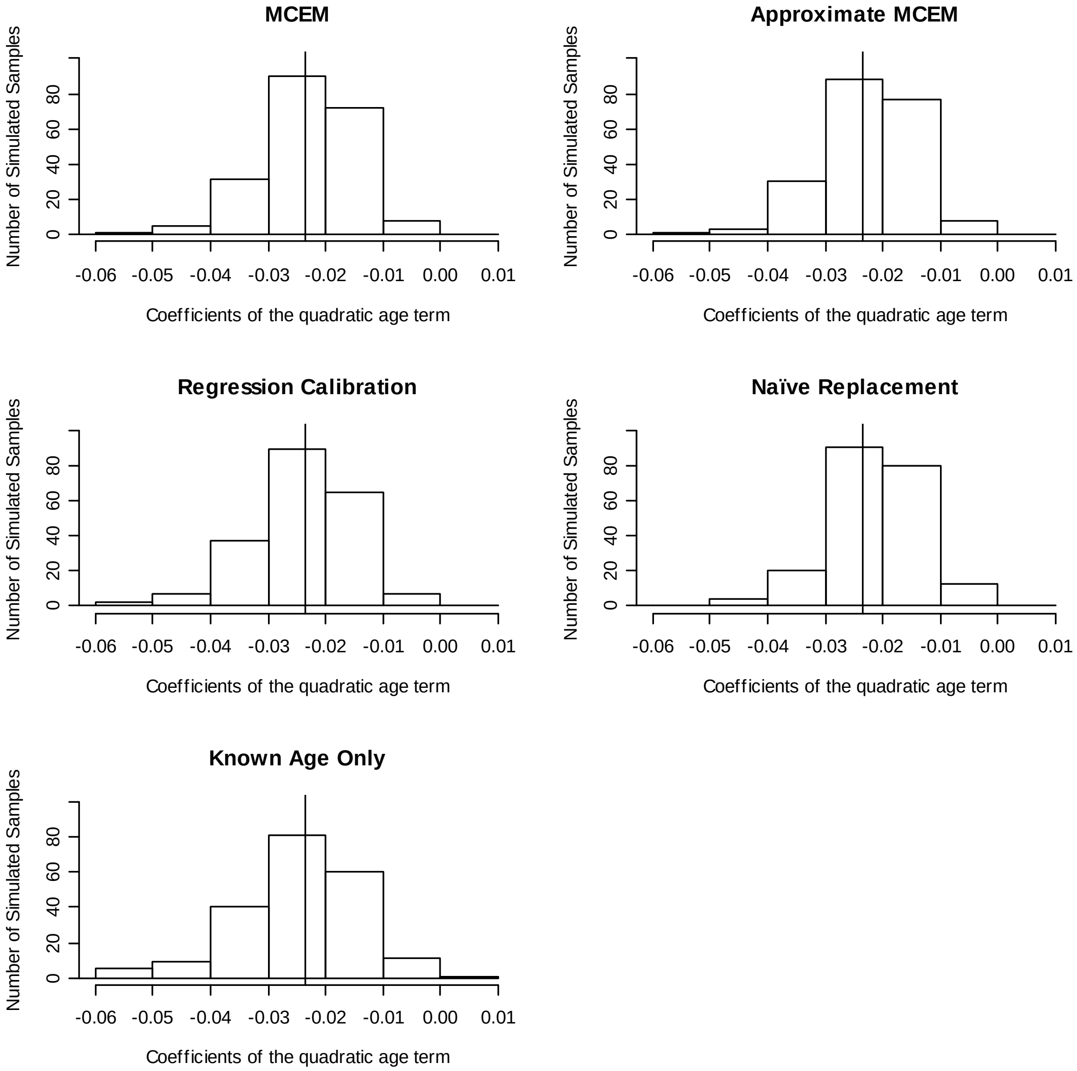


Figure A6.9 Monte Carlo sampling distributions for *n=50* and *25% censored* for the estimators of β_2, the age^2 term, for the 5 different fitting methods.

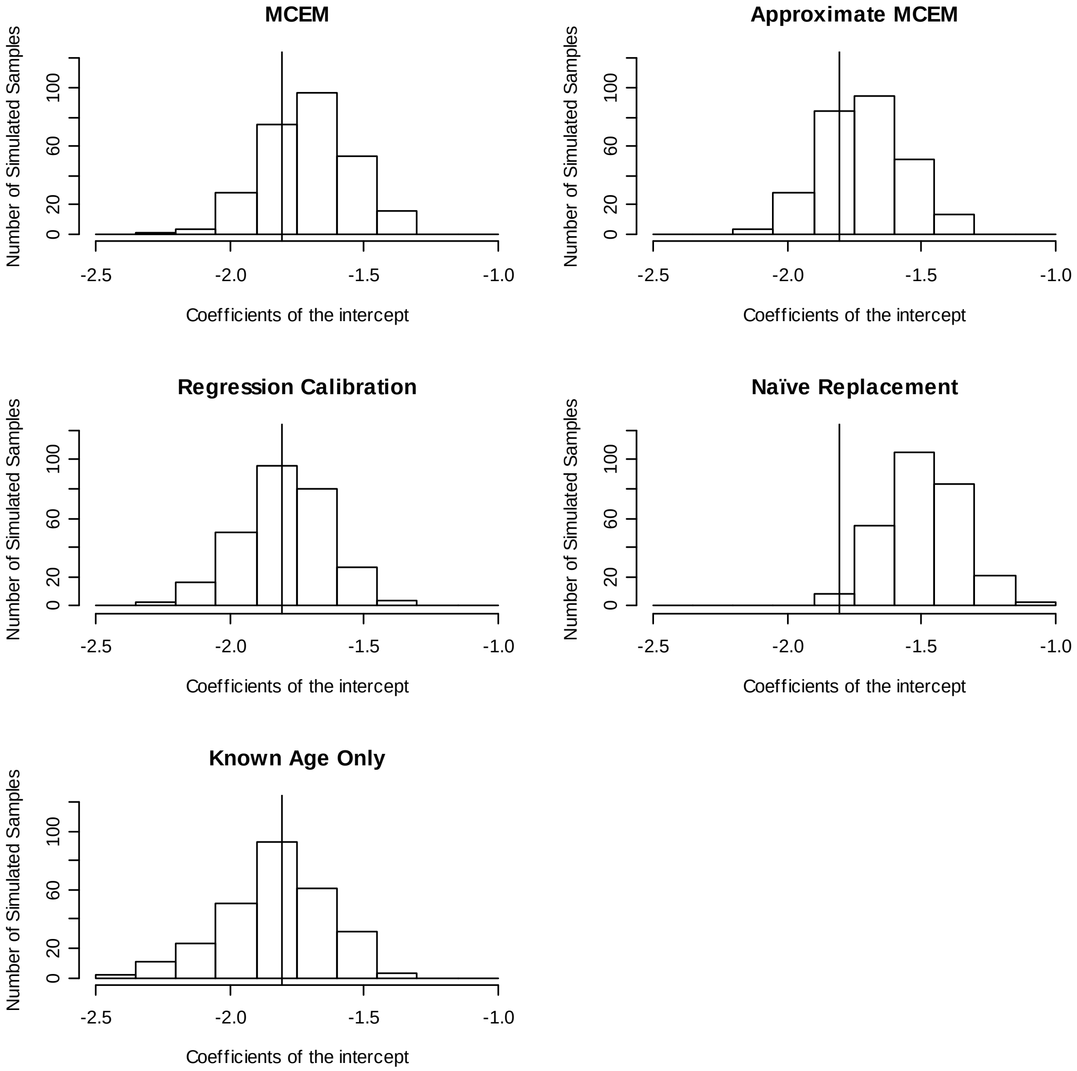


Figure A6.10 Monte Carlo sampling distributions for *n=400* and *25% censored* for the estimators of β_0, the intercept, for the 5 different fitting methods.

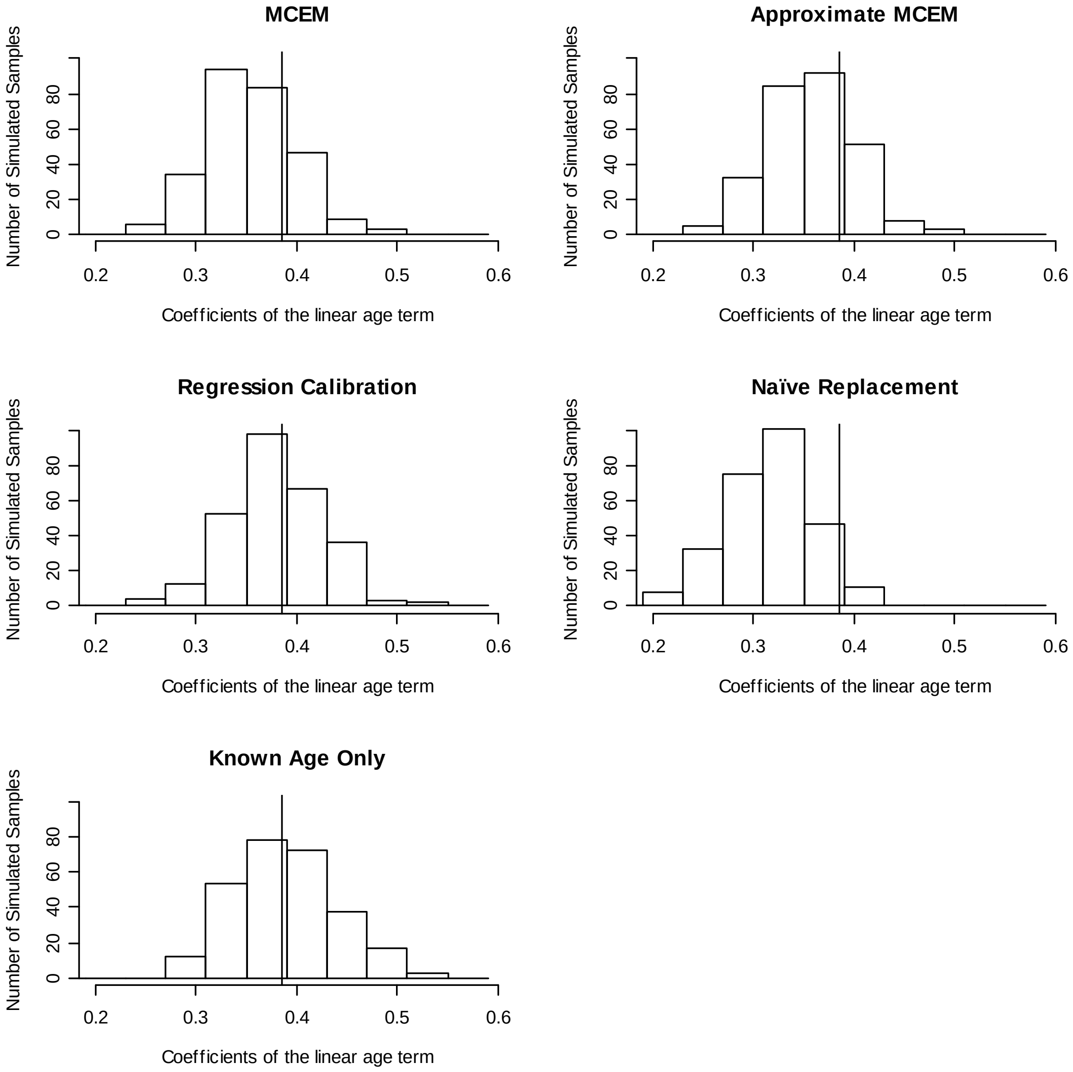


Figure A6.11 Monte Carlo sampling distributions for *n=400* and *25% censored* for the estimators of β_1, the age term, for the 5 different fitting methods.

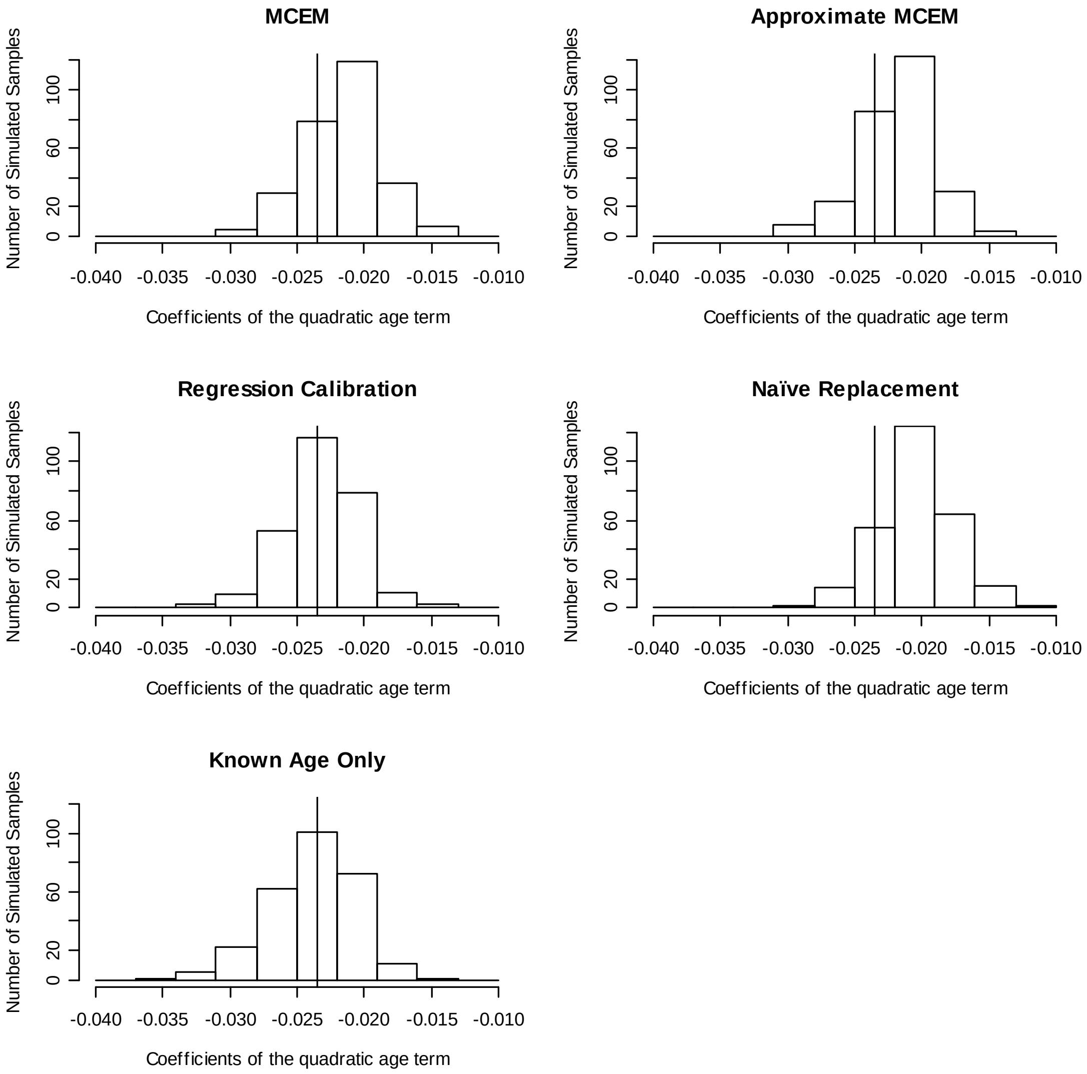


Figure A6.12 Monte Carlo sampling distributions for *n=400* and *25% censored* for the estimators of β_2, the age^2 term, for the 5 different fitting methods.

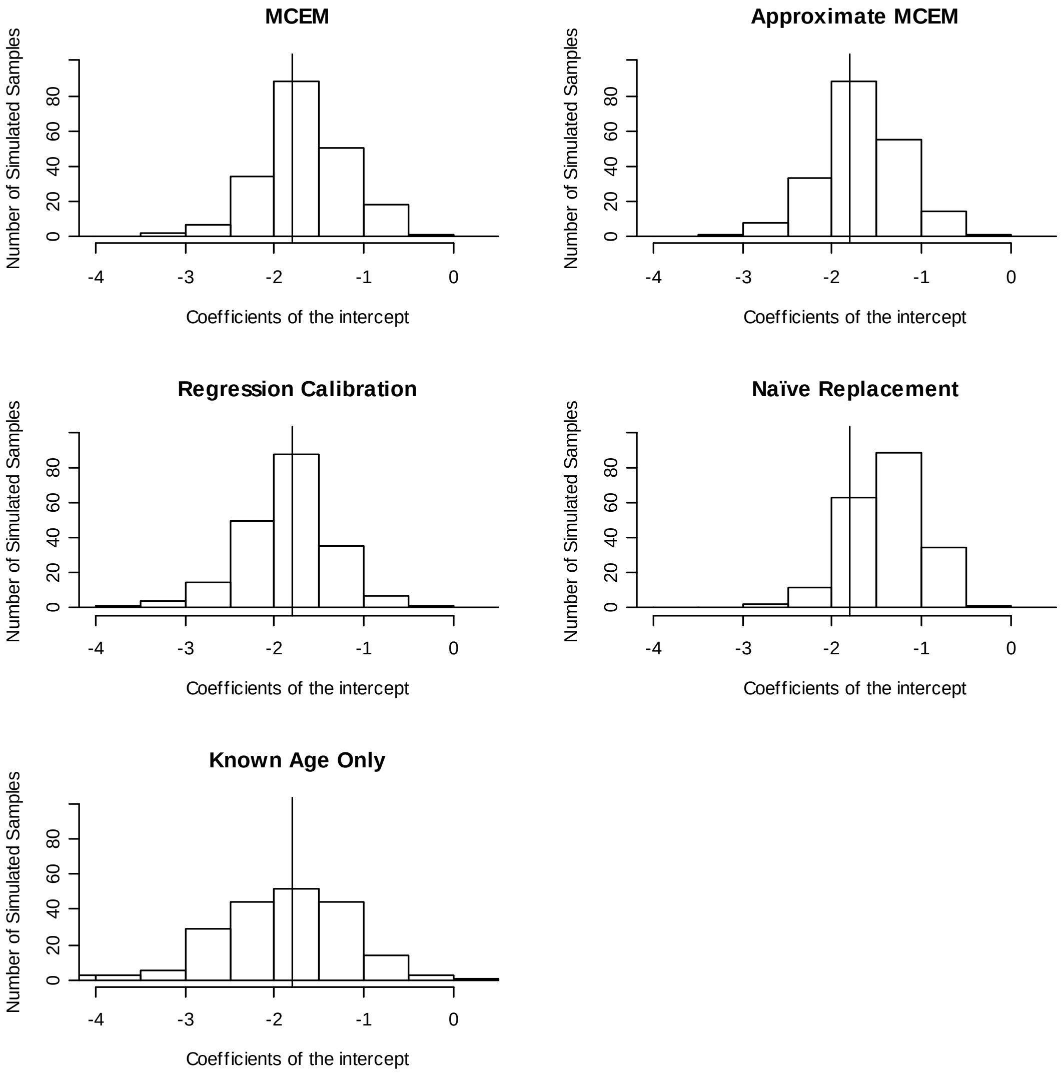


Figure A6.13 Monte Carlo sampling distributions for *n=50* and *50% censored* for the estimators of β_0, the intercept, for the 5 different fitting methods.

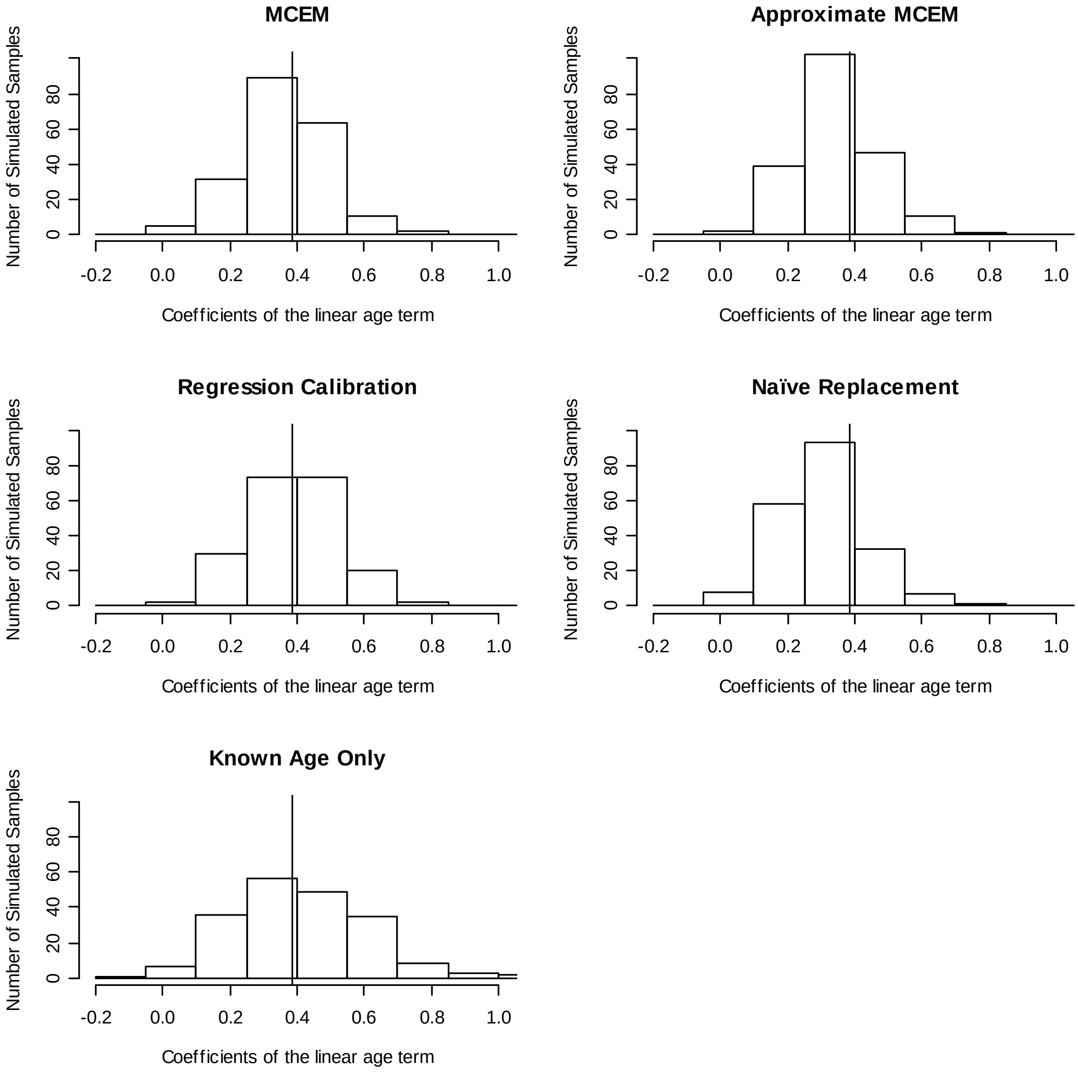


Figure A6.14 Monte Carlo sampling distributions for *n=50* and *50% censored* for the estimators of β_1, the age term, for the 5 different fitting methods.

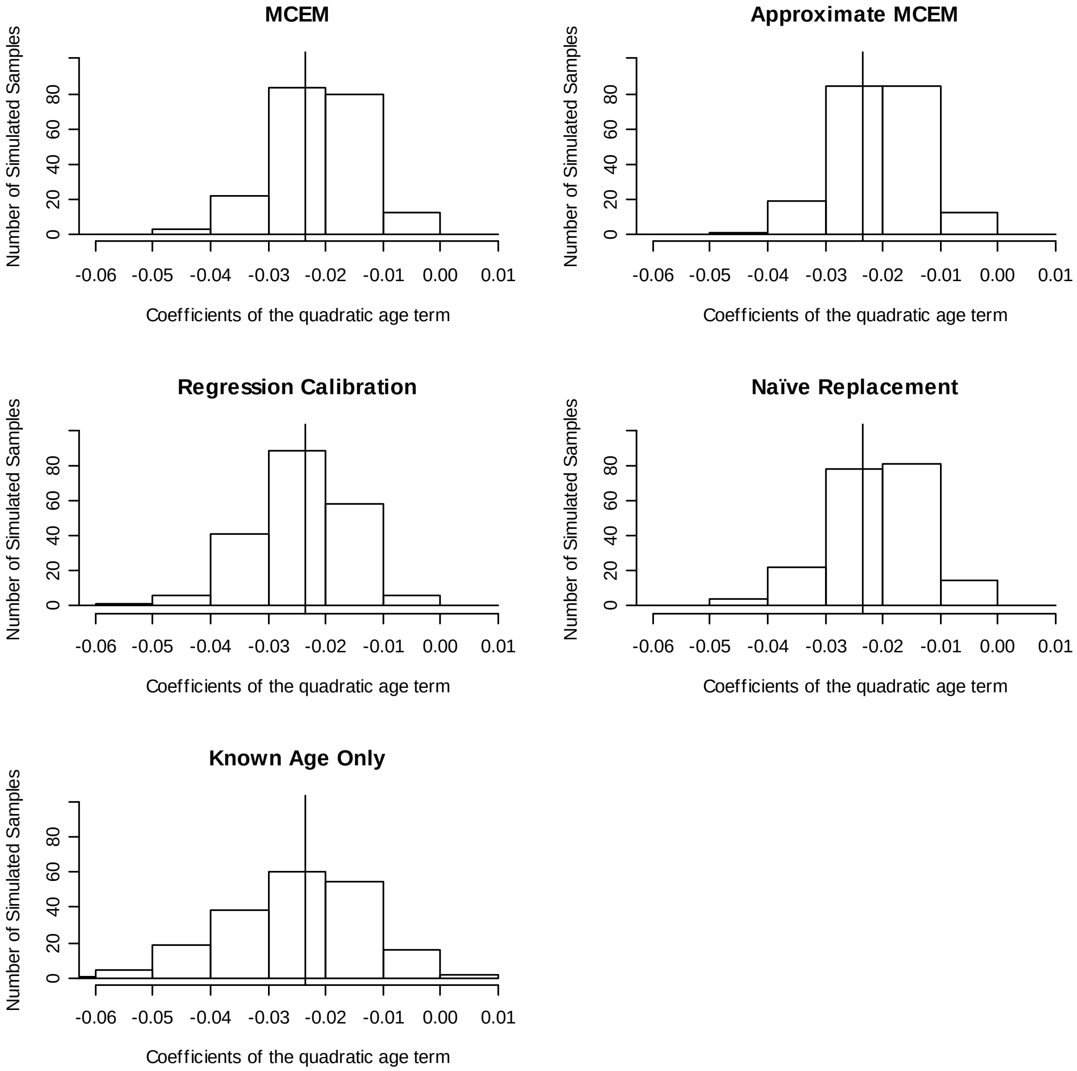


Figure A6.15 Monte Carlo sampling distributions for *n=50* and *50% censored* for the estimators of β_2, the age^2 term, for the 5 different fitting methods.

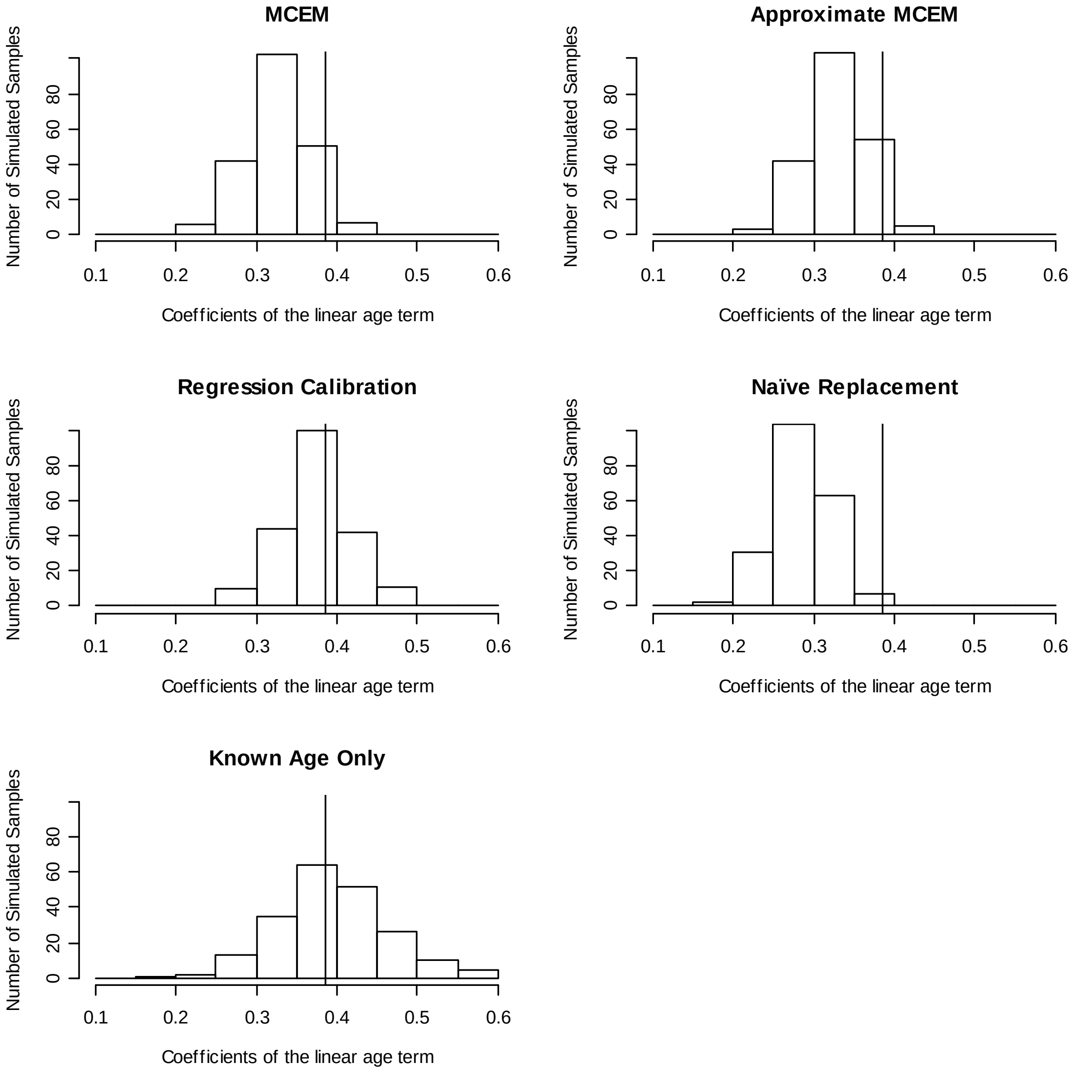
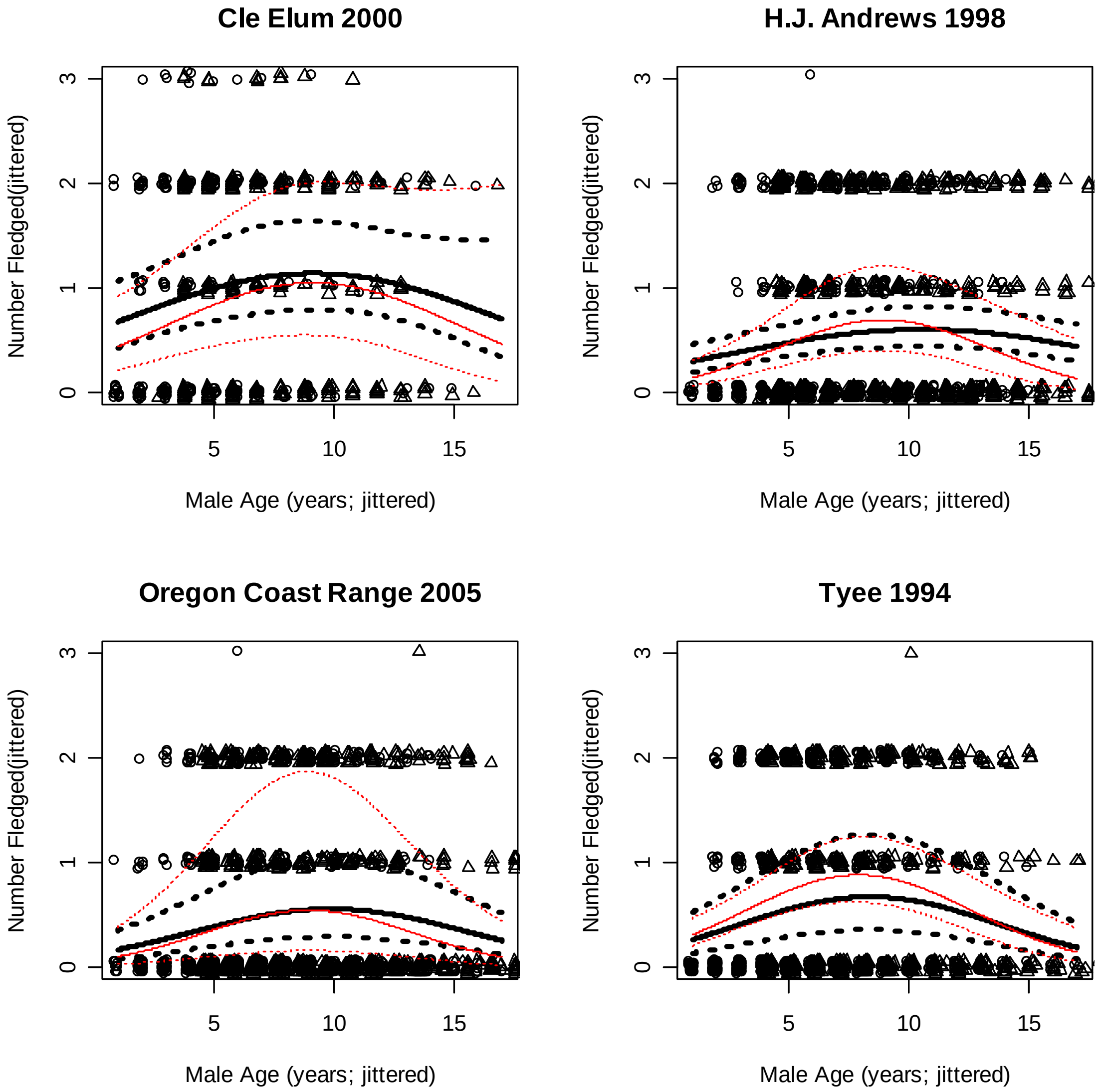


Figure A6.16 Monte Carlo sampling distributions for *n=400* and *50% censored* for the estimators of β_1, the age term, for the 5 different fitting methods.

A7 Regression Calibration Spotted Owl Analysis



Cle Elum 2000

H.J. Andrews 1998

Oregon Coast Range 2005

Tyee 1994

Figure A7.1 Northern Spotted Owl Model Fits. The darker lines are the regression calibration analysis and the lighter lines are the known age owls only. The dashed lines are the 95% confidence intervals.

Table A7.1 Regression calibration estimation results for Cle Elum

	Regression Calibration All Males		lmer Known Age Males Only	
Parameter	Estimate	SE	Estimate	SE
β_0	-0.0108	0.3071	0.2118	0.7419
β_{age}	0.1445	0.0595	0.2417	0.0880
β_{age^2}	-0.0079	0.0039	-0.0132	0.0062
β_{1990}	-0.1197	0.3050	-1.1714	1.0134
β_{1991}	-0.5350	0.3017	-0.8920	0.8327
β_{1992}	-0.0175	0.2839	-0.2413	0.7657
β_{1993}	-1.9605	0.4110	19.1415	3042.1741
β_{1994}	-0.0787	0.2907	-0.2921	0.7634
β_{1995}	-0.7547	0.3179	-1.1262	0.7938
β_{1996}	-0.1180	0.2997	-0.5873	0.7686
β_{1997}	-2.4324	0.5664	-19.2933	2755.8640
β_{1998}	-0.1361	0.3007	-0.7237	0.7715
β_{1999}	-0.9762	0.3699	-1.1869	0.7928
β_{2000}	-0.5184	0.3224	-1.2653	0.7885
β_{2001}	-0.4226	0.3285	-1.0528	0.7854
β_{2002}	-0.6659	0.3660	-1.6525	0.8534
β_{2003}	-0.2838	0.3226	-0.8773	0.7867
β_{2004}	-0.5657	0.3391	-0.8920	0.7847
β_{2005}	-0.6424	0.3472	-1.0588	0.7959
α_0	0.5536	0.0946		
α_{years}	0.0059	0.0197		
σ_u	0.0000		0.1798	
σ_a	0.4560			

Table A7.2 Regression calibration estimation results for H.J. Andrews

	Regression Calibration All Males		lmer Known Age Males Only	
Parameter	Estimate	SE	Estimate	SE
β_0	-0.1166	0.7384	-1.3088	0.8309
β_{age}	0.1625	0.0431	0.4531	0.1069
β_{age^2}	-0.0077	0.0022	-0.0254	0.0069
β_{1987}	-0.4870	0.7656	-17.5988	5717.5404
β_{1988}	-0.5153	0.7265	-0.4866	1.0280
β_{1989}	-1.3281	0.7404	-0.9190	0.8886
β_{1990}	-1.1078	0.7281	-0.6830	0.8155
β_{1991}	-1.4362	0.7356	-2.0862	1.0065
β_{1992}	-0.3166	0.7144	-0.0005	0.7421
β_{1993}	-17.9195	841.6792	-17.7043	1565.3191
β_{1994}	-1.4881	0.7321	-1.1521	0.7923
β_{1995}	-1.6981	0.7387	-1.6220	0.8175
β_{1996}	-0.6398	0.7184	-0.3271	0.7356
β_{1997}	-1.5459	0.7341	-1.3016	0.7760
β_{1998}	-1.2439	0.7245	-1.0785	0.7616
β_{1999}	-1.9178	0.7405	-1.0082	0.7657
β_{2000}	-0.8792	0.7195	-1.0164	0.7661
β_{2001}	-0.7329	0.7171	-0.4137	0.7360
β_{2002}	-0.8529	0.7194	-1.1963	0.7716
β_{2003}	-1.7204	0.7357	-1.2249	0.7745
β_{2004}	-0.4953	0.7149	-0.2838	0.7323
β_{2005}	-2.1330	0.7569	-1.5017	0.8226
α_0	1.0930	0.0843		
α_{years}	-0.0133	0.0161		
σ_u	0.0000		0.0000	
σ_a	0.5863			

Table A7.3 Regression calibration estimation results for Tyee

	Regression Calibration All Males		lmer Known Age Males Only	
Parameter	Estimate	SE	Estimate	SE
β_0	-0.5562	0.6334	14.7886	717.0570
β_{age}	0.2886	0.0482	0.3497	0.0697
β_{age^2}	-0.0172	0.0029	-0.0221	0.0046
β_{1985}	-2.5997	1.1864		
β_{1986}	-0.3834	0.6502		
β_{1987}	-1.0214	0.7691		
β_{1988}	-1.9481	0.7663		
β_{1989}	-0.8210	0.6574	13.3596	717.0573
β_{1990}	-0.9569	0.6335	13.0984	717.0571
β_{1991}	-1.4705	0.6448	12.7455	717.0571
β_{1992}	-0.6796	0.6278	13.0840	717.0570
β_{1993}	-2.2900	0.6777	11.8633	717.0571
β_{1994}	-1.0526	0.6344	13.2773	717.0570
β_{1995}	-1.6917	0.6510	12.6601	717.0570
β_{1996}	-0.4965	0.6285	13.6158	717.0570
β_{1997}	-1.0455	0.6379	12.7755	717.0570
β_{1998}	-0.9287	0.6349	13.1168	717.0570
β_{1999}	-1.3666	0.6479	12.7374	717.0570
β_{2000}	-1.1253	0.6422	12.9589	717.0570
β_{2001}	-0.3041	0.6278	13.8331	717.0570
β_{2002}	-0.9163	0.6352	13.1580	717.0570
β_{2003}	-1.6057	0.6505	12.5660	717.0570
β_{2004}	-0.8761	0.6323	13.1943	717.0570
β_{2005}	-0.7855	0.6324	13.3154	717.0570
α_0	0.9340	0.0624		
α_{years}	-0.0230	0.0113		
σ_u	0.2862		0.2628	
σ_a	0.5210			