Estimation Methods for Data Subject to Detection Limits

Ryan C. May

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

Chapel Hill
2012

Approved by:

Dr. Joseph G. Ibrahim

Dr. Haitao Chu

Dr. Stephen Cole

Dr. Pei Fen Kuan

Dr. John S. Preisser

Abstract

RYAN C. MAY: Estimation Methods for Data Subject to Detection Limits (Under the direction of Dr. Joseph G. Ibrahim and Dr. Haitao Chu)

Data subject to detection limits appear in a wide variety of studies. Data subject to detection limits are usually left-censored at the detection limit, often due to limitations in the measurement procedure being used. This thesis addresses three issues common to the analysis of data subject to detection limits. The first of these is the estimation of the limit of detection using repeated measurements from known analyte concentrations. An innovative change-point model is proposed to more accurately model the standard deviation of measured analyte concentrations, resulting in improved estimation of the limit of detection. The proposed methodology is applied to copy number data from an HIV pilot study. The second topic concerns estimation using generalized linear models when multiple covariates are subject to a limit of detection. We propose a Monte Carlo version of the EM algorithm similar to that in Ibrahim, Lipsitz, and Chen to handle a large number of covariates subject to detection limits in generalized linear models. Censored covariate values are sampled using the Adaptive Rejection Metropolis Algorithm of Gilks, Best, and Tan. This procedure is applied to data from the National Health and Nutrition Examination Survey (NHANES), in which values of urinary heavy metals are subject to a limit of detection. Through simulation studies, we show that the proposed approach can lead to a significant reduction in variance for parameter estimates in these models, improving the power of such studies. The third and final topic addresses the joint modeling of longitudinal and survival data using time-varying covariates that are both intermittently missing and subject to a limit of detection. The model is motivated by data from the Multicenter AIDS Cohort Study (MACS), in which HIV+ subjects have viral load and CD4 cell counts measured at repeated visits along with survival data. The viral load data is subject to both left-censoring due to detection limits (17%) and intermittent missingness (27%). A Bayesian analysis is conducted on the MACS data using the proposed joint

model. The proposed method is shown to improve the precision of estimates when compared to alternative methods.

Acknowledgments

I would first like to thank both of my advisors, Dr. Joseph Ibrahim and Dr. Haitao Chu, for their guidance with this work. Their immense knowlege and assistance has been a tremendous help to me, and I greatly appreciate the amount of time and effort they committed to helping me with this project.

I would also like to thank Dr. Stephen Cole, Dr. Pei Fen Kuan, and Dr. John Preisser for serving on my committee. I greatly appreciate the time they have given, and their insightful comments have significantly ($p<.05$) improved the quality of this work.

I would like to also thank my fellow classmates for their friendship and help through these many years at UNC. In particular, I would like to thank Allison Deal, Annie Green Howard, Dave Kessler, Dustin Long, Leann Long, Virginia Pate, Shangbang Rao, and Vonn Walter.

Many thanks to my parents, Jeff and Ann May, for their unwavering support of a son who wants to spend 11+ years of his life in college.

And finally, I would like to thank my beautiful new bride, Jeanine May, for the love and affection that makes every new day even better than the last.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

- ARMS - Adaptive Rejection Metropolis Sampling
- HIV - Human Immunodeficiency Virus
- LOD - Limit of Detection
- LOQ - Limit of Quantitation
- MLE - Maximum Likelihood Estimate
- MACS - Multicenter AIDS Cohort Study
- NHANES - National Health and Nutrition Examination Survey

Chapter 1

Change-Point Models to Estimate the Limit of Detection

1.1 Introduction

In many laboratory assays, interest resides in quantifying very dilute quantities in solution. As concentrations of analytes decrease, however, the resulting measured levels from a measurement device often become less precise. At some low concentration level, a measured response cannot accurately be distinguished from background noise, the measured response from a blank sample. This low concentration point is called the limit of detection (LOD), a point that is specific to each particular measurement device (Clinical and Laboratory Standards Institute, 2004). Though the general definition given above for the limit of detection is widely accepted, the methodology used to determine the limit of detection is quite varied. In this chapter we consider the estimation of the limit of detection using repeated measurements from known analyte concentrations. This analysis is motivated by data from a study of low levels of HIV that persist despite potent therapy, in which a novel assay was developed to detect changes in low-level HIV expression after a drug intervention. The assay measuring HIV expression becomes less precise as the concentration of HIV decreases, but a limit of detection for the assay is not known. In this chapter we consider estimating the LOD for this assay based on measurements replicated on several solutions containing known quantities of HIV.

The rest of this chapter is organized as follows. In Section 1.2 we discuss past research on estimating the limit of detection, and develop the notation for the rest of the chapter. We explain how past research can incorrectly specify the error distribution for a measurement device, leading to incorrect estimation of the limit of detection. In Section 1.3, we introduce

the proposed change-point model and discuss a two-stage estimation approach for obtaining maximum likelihood estimates from the model. In Section 1.4, we examine the proposed model using a simulation study, then apply the method to data from the aforementioned HIV assay in Section 1.5. We conclude this chapter in Section 1.6 with a discussion.

1.2 Background

To distinguish between low analyte concentrations and those of a blank sample, many estimation approaches aim to quantify the distribution of measurements obtained from a blank sample. The distribution of assay measurements for a blank sample is often assumed to be Gaussian, with mean μ_{blank} and variance σ^2_{blank} (Anderson 1989). A limit of detection is then chosen to fall a "reasonable" distance outside of this blank distribution. Consequently, many definitions of LOD take the following form (Whitcomb and Schisterman 2008):

$$\text{LOD} = \mu_{blank} + K\sigma_{blank} \tag{1.1}$$

where K is a definition-specific constant, usually in the range of 2.0 to 3.0 (Browne and Whitcomb 2010, Long and Winefordner 1983, Thomsen et al. 2003).

When $K = 3$, it is expected that 99.9% of measurements from a blank sample will fall below the limit of detection. Clearly, the larger value of K that is chosen, the higher the LOD will be, and the lower the chance that a value from a blank will fall above the LOD. Using this definition, many estimation approaches are designed to accurately estimate μ_{blank} and σ_{blank}. In practice, such estimation is straightforward when many repeated measurements can be obtained from a blank sample, by taking the sample mean and standard deviation (SD) as estimates of μ_{blank} and σ_{blank}.

When blank measurements are not available, alternative estimation approaches can be utilized. One approach involves taking repeated measurements of a known low concentration of analyte and using these measurements as proxy measurements for a blank sample. In this case, the LOD definition is extremely similar to (1.1). With μ_{low} and σ_{low} representing the mean and standard deviation of the distribution of measurements at the low concentration

(again assumed to follow a Gaussian distribution), the limit of detection is defined as:

$$\text{LOD} = \mu_{low} + K\sigma_{low} \tag{1.2}$$

The previous definition of the LOD in (1.1) includes only a specification of the distribution of a blank sample. Using this definition enables direct control of the type I error, the chance of incorrectly specifying a blank sample as containing some concentration of analyte. For example, when $K = 3$ the chance of a type I error is only 0.1% for any particular blank measurement. However, this specification does not control for type II error, the chance of incorrectly specifying a sample containing analyte as coming from a blank. If the type II error is high, clearly there is still difficulty in conclusively distinguishing a concentration value near the LOD from a blank. Consequently, many definitions of the LOD take both type I and type II error into account. To ensure that measured values for concentrations at the LOD are unlikely to fall in the range of a blank sample, alternative definitions of LOD account for the distribution of measurement values at some known small concentration of analyte. The limit of detection is defined as follows (Armbruster and Pry 2008, Browne and Whitcomb 2010):

$$\text{LOD} = \mu_{blank} + 1.645\sigma_{blank} + 1.645\sigma_{low} \tag{1.3}$$

Using this definition, 95% of blank samples will fall below $\mu_{blank} + 1.645\sigma_{blank}$ (called the "limit of blank", or "limit of decision"), and 95% of measurements for concentrations at the limit of detection will fall above the limit of blank. It should be noted that definition (1.3) is only needed when it is assumed that the measurement standard deviation for a blank sample is different from the standard deviation for any other "low" concentration sample at or around the LOD (i.e. $\sigma_{blank} \neq \sigma_{low}$). Many authors (Anderson 1989, Armbruster 1994, Browne and Whitcomb 2010) assume a constant measurement error variance for any true concentrations near or below the limit of detection. In this case, the choice of K in (1.1) specifies the chance of a type I or type II error. When $K = 3.29$, the chance of either type of misclassification is 5%; when $K = 3$, the chance is 7%. Alternative (but similar) definitions to (1.3) calculate a pooled measurement standard deviation from both blank and low samples, using the pooled

estimate in place of both σ_{blank} and σ_{low} (Long and Winefordner 1982).

The LOD definitions displayed in equations (1.1) and (1.3) usually are performed under the assumption that the measurement distribution for a blank sample is Gaussian. In reality, characteristics of the measurement device often result in non-Gaussian measurement distributions for a blank. In such cases, nonparametric methods have been proposed (Linnet and Kondratovich 2004), which involve estimating quantiles from the observed blank distribution. For definition (1.3), the value of σ_{low} can still be estimated with parametric methods when it is reasonable to assume a Gaussian distribution for a low concentration sample. If the low concentration cannot be assumed Gaussian, quantile estimation can be used to estimate σ_{low} as well.

In practice, data analysts often do not have access to replicates of data from blank or "low" concentrations. This is the case for the HIV pilot study data considered in Section 1.5, in which the number of polymerase chain reaction (PCR) cycles needed to obtain a blank sample measurement is too high to be operationally feasible. In this case it is difficult to directly estimate the distribution of measurements for a blank sample, or for the distribution of any low concentration sample. Estimation often proceeds using higher analyte concentrations from which measurements are more easily obtained. A regression line is then fit to $(X_1, Y_1), ..., (X_n, Y_n)$, the n observed pairs of analyte concentrations X and measured responses Y. This fitted regression line is known as a linear calibration curve. Assuming a linear relationship between X and Y, we have the following model specification (Dunne 1995, Hubaux and Vos 1970):

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2) \tag{1.4}$$

Taking $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$, the assumptions of (1.4) specify the distribution of $Y_i|X_i, \boldsymbol{\theta} \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$. Clearly, the parameter estimates of the model can be used to directly estimate the distribution of $Y_{X_i=0}$, the response for a blank sample. When the parameter vector $\boldsymbol{\theta}$ is known, we have:

$$Y_{X_i=0}|\boldsymbol{\theta} \sim N(\beta_0, \sigma^2)$$

Using the definition of LOD in equation (1.1) with $K = 3$, the LOD under model (1.4) is:

$$\text{LOD}_Y = \beta_0 + 3\sigma$$

The above specification is conditional on the true values of the model parameters β_0, β_1, and σ^2. Let $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$ denote the Maximum Likelihood Estimates (MLE's) for β_0, β_1, and σ^2 (and denote $V[\hat{\beta}_0] = \hat{\sigma}^2_{\beta_0}$). The response distribution for a blank sample can be estimated as follows:

$$Y_{X_i=0} \sim N(\hat{\beta}_0, \hat{\sigma}^2 + \hat{\sigma}^2_{\beta_0})$$

Consequently, the limit of detection can be estimated as (Cox 2005, Hubaux and Vos 1970):

$$\widehat{\text{LOD}}_Y = \hat{\beta}_0 + 3(\hat{\sigma}^2 + \hat{\sigma}^2_{\beta_0})^{1/2} \tag{1.5}$$

In practice, the limit of detection is usually defined in terms of the concentration X instead of the measurement Y. To obtain the limit of detection for concentration, a simple linear transformation on $\widehat{\text{LOD}}_Y$ is performed (Gibbons et al., 1992), to obtain:

$$\widehat{\text{LOD}}_X = \frac{3(\hat{\sigma}^2 + \hat{\sigma}^2_{\beta_0})^{1/2}}{\hat{\beta}_1} \tag{1.6}$$

The standard analysis for estimating the LOD with a linear calibration curve assumes that the variance of measured responses is constant at all concentration values. In many practical applications this is not the case, and it is common for a measurement device to become more (or less) precise as the concentration of analyte increases. In the most basic case (or possibly under suitable transformation), the measurement standard deviation is assumed to decrease linearly with the concentration. This case was first considered by Oppenheimer et al. (1983), and specifies the error distribution from model (1.4) as follows:

$$\epsilon_i \sim N(0, (\sigma_0 + \sigma_1 X_i)^2) \tag{1.7}$$

Using this specification, the limits of detection LOD_X and LOD_Y are again estimated as in

equations (1.5) and (1.6).

Current analysis methods for estimating the limit of detection with a linear calibration curve either assume a constant standard deviation for measurement error as in (1.4), or a linear change in measurement standard deviation by concentration as in (1.7). As noted by several authors (Armbruster 1994, Hubaux and Vos 1970, Clinical and Laboratory Standards Institute 2004, Browne and Whitcomb 2010), a more realistic assumption is that the measurement standard deviation changes for "high" concentration values above the limit of detection, while remaining effectively constant for "low" concentration values. Under this assumption, the use of a constant standard deviation model like (1.4) for all concentration values can result in underestimation (if precision increases with concentration) or overestimation (if precision decreases with concentration) of the limit of detection. The use of a linear standard deviation model like (1.7) could provide the opposite effect, overestimating the LOD when precision increases with concentration and underestimating when precision decreases with concentration. To correct these potential biases in LOD estimation with a linear calibration curve, in Section 1.3 a change-point model is proposed to more accurately model the measurement error for all concentrations.

1.3 Change-Point Model

In Section 1.2 it was discussed that current analyses using a linear calibration curve usually assume either a constant measurement standard deviation for all analyte concentrations, or a measurement standard deviation that varies linearly with the analyte concentration. Because measurements below the limit of detection are indistinguishable from a blank, it follows that the measurement standard deviation should be constant for low analyte concentrations. Such a distribution can be modeled using a change-point for the measurement standard deviation. While the literature on change-point models in both regression (Bai 1997, Hawkins 2001) and mixed (Cudeck and Klebe 2002) modeling is quite rich, to our knowledge no published articles have looked at models with a change-point on the standard deviation of the error. Taking the notation for a linear calibration curve presented in equation (1.4), with σ_i representing

the measurement standard deviation for concentration X_i, we make the following assumption for the form of σ_i:

$$\sigma_i = \begin{cases} \sigma_0 & \text{if } X_i \leq \lambda \\ \sigma_0 + \sigma_1(X_i - \lambda) & \text{if } X_i > \lambda \end{cases} \tag{1.8}$$

where λ represents the change-point for measurement standard deviation. As noted in Section 1.2, a common definition for the LOD is 3 standard deviations away from the expected value of a blank sample. Given the assumption in equation (1.8), the standard deviation of a blank sample is σ_0. The expected value of the blank sample is the intercept term for the model, β_0. The LOD for both the measurement (Y) and true concentration (X) are:

$$\text{LOD}_Y = \beta_0 + 3\sigma_0 \tag{1.9}$$

$$\text{LOD}_X = \frac{3\sigma_0}{\beta_1} \tag{1.10}$$

Taking $\hat{\beta}_0$, $\hat{\sigma}_0$, $\hat{\sigma}_{\beta_0}$, and $\hat{\beta}_1$ as the MLE's for their respective parameters, the limit of detection is estimated following equations (1.5) and (1.6):

$$\begin{aligned} \widehat{\text{LOD}}_Y &= \hat{\beta}_0 + 3(\hat{\sigma}_0^2 + \hat{\sigma}_{\beta_0}^2)^{1/2} \\ \widehat{\text{LOD}}_X &= \frac{3(\hat{\sigma}_0^2 + \hat{\sigma}_{\beta_0}^2)^{1/2}}{\hat{\beta}_1} \end{aligned} \tag{1.11}$$

In the situation described above, the MLE's $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}$ have a simple closed form. However, in the HIV pilot study motivating this work the measured assay responses Y are right-censored at a constant upper limit (here denoted as γ). Accounting for this censoring, the log-likelihood for an individual observation is:

$$l_i = \begin{cases} -\log(\sigma_i) - \frac{1}{2\sigma_i^2}(Y_i - (\beta_0 + \beta_1 X_i))^2 & \text{if } Y_i \leq \gamma \\ \log\left[1 - \Phi\left\{\frac{\gamma - (\beta_0 + \beta_1 X_i)}{\sigma_i}\right\}\right] & \text{if } Y_i > \gamma \end{cases}$$

where $\Phi()$ is the cumulative distribution function of a standard normal random variable.

Again denoting $(X_1, Y_1), ..., (X_n, Y_n)$ as the n iid observations available for analysis, the log-likelihood for the model can be expressed as:

$$l(\beta_0, \beta_1, \sigma_0, \sigma_1, \lambda) = \sum_{i=1}^{n} l_i \tag{1.12}$$

In order to estimate the LOD under model 1.8, we maximize the log-likelihood (1.12) with respect to the parameter vector $(\beta_0, \beta_1, \sigma_0, \sigma_1, \lambda)$. Maximization of the log-likelihood is done under the following constraints. First, the change-point (λ) must be constrained within the range of the observed X_i. Taking $x_{(1)}...x_{(n)}$ as the order statistics for the observed X_i, this is expressed as $x_{(1)} \leq \lambda \leq x_{(n)}$. The rationale for this constraint is that the parameters σ_0 and λ become unidentifiable when $\lambda \leq x_{(1)}$, and the parameters σ_1 and λ become unidentifiable when $\lambda \geq x_{(n)}$.

The second model constraint is that the error standard deviation σ_i cannot be negative at $x_{(1)}$, and the third model constraint is that σ_i cannot be negative at $x_{(n)}$. Together, these constraints specify that σ_i is nonnegative at all points in $[x_{(1)}, x_{(n)}]$. One way to specify these constraints is to require $\sigma_0 \geq 0$ and $\sigma_0 + \sigma_1(x_{(n)} - \lambda) \geq 0$. All constraints on the model are given below:

(i) $x_{(1)} \leq \lambda \leq x_{(n)}$

(ii) $\sigma_0 \geq 0$

(iii) $\sigma_0 + \sigma_1(x_{(n)} - \lambda) \geq 0$

Constraints (i) and (ii) are both linear, so are straightforward to implement in any maximization of the resulting log-likelihood. However, constraint (iii) is not linear, as it involves the term $\sigma_1\lambda$. Therefore, maximizing (1.12) subject to (i), (ii), and (iii) is challenging since many standard optimization routines only allow for linear constraints. To get around this issue, we instead use a two-stage optimization routine (Smyth 1996). For ease of exposition, we will define $\sigma_{x_{(n)}} = \sigma_0 + \sigma_1(x_{(n)} - \lambda)$, the standard deviation at $x_{(n)}$, the maximum observed concentration value. For generic parameter ϕ, we denote $\phi^{(t)}$ as the parameter estimate at

the t-th iteration of the estimation routine. The proposed two-stage optimization routine is as follows:

1. Fix $\lambda = \lambda^{(t-1)}$. Maximize (1.12) with fixed λ, subject to the **linear** constraints:

   i) $\sigma_0 \geq 0$

   ii) $\sigma_0 + \sigma_1(x_{(n)} - \lambda) \geq 0$

2. Taking $\hat{\sigma}_0$ and $\hat{\sigma}_1$ as the estimates from step 1, fix $\sigma_{x_{(n)}} = \hat{\sigma}_0 + \hat{\sigma}_1(x_{(n)} - \lambda^{(t-1)})$. Maximize (1.12) with fixed $\sigma_{x_{(n)}}$ subject to the **linear** constraints:

   i) $x_{(1)} \leq \lambda \leq x_{(n)}$

   ii) $\sigma_0 \geq 0$

   Obtain estimates $\beta_0^{(t)}$, $\beta_1^{(t)}$, $\sigma_0^{(t)}$, $\lambda^{(t)}$, set $\sigma_1^{(t)} = (\sigma_{x_{(n)}} - \sigma_0^{(t)})/(x_{(n)} - \lambda^{(t)})$

Steps 1 and 2 in the above procedure are repeated until convergence is achieved for all parameter estimates. The convergence criterion used for parameter ϕ specifies that $\phi^{(t)} - \phi^{(t-1)} \leq K$, with K again representing a generic constant. The proposed optimization routine is relatively simple to implement, as the likelihood in (1.12) is not overly complex. For all individual model analyses presented in Sections 1.4 and 1.5, optimization in each stage was performed using R software (R Development Core Team 2008) with the `constrOptim()` function. In the following section we perform a simulation study to test the new estimation approach, and then apply our approach to data from an HIV pilot study.

1.4 Simulation Study

A simulation study was conducted to analyze the performance of the proposed change-point model. The data for this simulation study was generated using a parameter specification that mirrored that for the HIV data analysis presented in Section 1.5. The model was specified as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma_i^2)$$

Table 1.1: Parameter estimates for simulation study, comparing change-point model to linear standard deviation model and constant standard deviation model

		N = 80						N = 150						N = 300					
		CP*		Linear		Constant		CP		Linear		Constant		CP		Linear		Constant	
Parameter	True	Bias	SD†	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Bias	SD
Model 1: $\lambda = 1.5$																			
β_0	45	-0.10	0.195	-0.08	0.204	-0.14	0.213	-0.10	0.142	-0.08	0.149	-0.14	0.156	-0.10	0.100	-0.08	0.105	-0.14	0.110
β_1	-3.7	0.02	0.043	0.02	0.045	0.04	0.051	0.02	0.031	0.02	0.033	0.04	0.038	0.02	0.022	0.02	0.023	0.04	0.027
σ_0	1.1	-0.19	0.123	0.22	0.167	-0.40	0.069	-0.18	0.091	0.22	0.120	-0.40	0.051	-0.18	0.063	0.22	0.085	-0.40	0.036
σ_1	-0.25	-0.01	0.060	0.03	0.037	-	-	-0.01	0.041	0.03	0.027	-	-	0.00	0.028	0.03	0.019	-	-
λ	1.5	0.74	0.640	-	-	-	-	0.74	0.479	-	-	-	-	0.74	0.336	-	-	-	-
LOD_X	0.89	-0.13	0.103	0.20	0.138	-0.30	0.062	-0.14	0.076	0.19	0.099	-0.31	0.044	-0.14	0.053	0.19	0.070	-0.31	0.031
AIC Best Fit		60.3%		39.7%		0.0%		78.8%		21.2%		0.0%		94.5%		5.5%		0.0%	
Model 2: $\lambda = 2.5$																			
β_0	45	-0.11	0.215	-0.08	0.235	-0.14	0.224	-0.11	0.156	-0.08	0.171	-0.14	0.164	-0.11	0.110	-0.08	0.121	-0.14	0.116
β_1	-3.7	0.02	0.047	0.02	0.051	0.04	0.054	0.02	0.034	0.02	0.037	0.04	0.040	0.02	0.024	0.02	0.026	0.04	0.028
σ_0	1.1	-0.12	0.121	0.45	0.210	-0.31	0.077	-0.11	0.089	0.46	0.151	-0.30	0.056	-0.11	0.064	0.46	0.107	-0.30	0.040
σ_1	-0.34	-0.02	0.214	0.08	0.048	-	-	-0.01	0.066	0.08	0.034	-	-	-0.01	0.045	0.08	0.024	-	-
λ	2.5	0.28	0.581	-	-	-	-	0.31	0.435	-	-	-	-	0.33	0.320	-	-	-	-
LOD_X	0.89	-0.07	0.102	0.39	0.173	-0.22	0.068	-0.08	0.075	0.39	0.124	-0.23	0.049	-0.08	0.054	0.38	0.088	-0.23	0.034
AIC Best Fit		89.3%		10.7%		0.0%		98.4%		1.6%		0.0%		100%		0.0%		0.0%	
Model 3: $\lambda = 3.5$																			
β_0	45	-0.12	0.222	-0.08	0.259	-0.14	0.226	-0.12	0.162	-0.08	0.190	-0.14	0.166	-0.12	0.114	-0.08	0.134	-0.14	0.117
β_1	-3.7	0.02	0.048	0.02	0.055	0.04	0.056	0.02	0.035	0.02	0.040	0.04	0.041	0.02	0.025	0.02	0.028	0.04	0.029
σ_0	1.1	-0.08	0.112	0.71	0.303	-0.23	0.081	-0.08	0.081	0.72	0.211	-0.22	0.059	-0.08	0.057	0.73	0.147	-0.22	0.042
σ_1	-0.57	-0.30	16.08	0.27	0.073	-	-	-0.22	6.941	0.27	0.050	-	-	-0.13	3.297	0.26	0.034	-	-
λ	3.5	0.03	0.495	-	-	-	-	0.09	0.366	-	-	-	-	0.12	0.267	-	-	-	-
LOD_X	0.89	-0.04	0.095	0.60	0.249	-0.15	0.071	-0.05	0.068	0.60	0.173	-0.16	0.051	-0.05	0.048	0.60	0.120	-0.17	0.036
AIC Best Fit		98.7%		1.2%		0.1%		100%		0.0%		0.0%		100%		0.0%		0.0%	
Model 4: $\lambda = 4.5$																			
β_0	45	-0.12	0.227	-0.09	0.269	-0.14	0.229	-0.12	0.165	-0.09	0.193	-0.14	0.167	-0.12	0.115	-0.09	0.133	-0.14	0.117
β_1	-3.7	0.02	0.05	0.02	0.058	0.04	0.059	0.02	0.036	0.02	0.041	0.03	0.044	0.02	0.025	0.02	0.028	0.04	0.03
σ_0	1.1	-0.06	0.11	0.65	0.513	-0.17	0.085	-0.05	0.073	0.69	0.393	-0.16	0.061	-0.05	0.051	0.72	0.29	-0.16	0.044
σ_1	-1.7	-0.08	21.8	1.44	0.144	-	-	-0.43	13.693	1.43	0.11	-	-	-2.05	39.995	1.43	0.081	-	-
λ	4.5	-0.62	0.67	-	-	-	-	-0.41	0.385	-	-	-	-	-0.3	0.31	-	-	-	-
LOD_X	0.89	-0.02	0.094	0.56	0.417	-0.1	0.075	-0.03	0.062	0.58	0.32	-0.11	0.053	-0.03	0.043	0.59	0.236	-0.12	0.038
AIC Best Fit		96.0%		0.6%		3.4%		99.5%		0.0%		0.5%		100%		0.0%		0.0%	

*CP = Change-point model (1.8), Linear = Linear standard deviation without a change-point (1.7), Constant = Constant standard deviation (1.4)
†SD = Standard deviation

with σ_i having the change-point specification given by (1.8).

Following the HIV data, only 5 different values of concentration X_i were used at 1, 2, 3, 4, and 5. For each of the concentration values used, repeated measurements (Y_i) were generated. The number of Y_i generated for each of the 5 concentration values was equal, a balanced allocation. For all simulations, the parameter values were specified as follows: $\beta_0 = 45, \beta_1 = -3.7, \sigma_0 = 1.1$. Four different sets of simulations were run using a different value for the change-point, with λ taking values 1.5, 2.5, 3.5, and 4.5 to span the range of X_i. The value of $\sigma_{X_{(n)}}$, the standard deviation at the maximum concentration value, was kept constant at 0.25 for all simulations (again mirroring results from the HIV data analysis). The specified values of σ_0, λ, and $\sigma_{X_{(n)}}$ determined the parameter value of σ_1 for each simulated data set. Following the real-life data set in Section 1.5, all values of Y_i falling above 42 were set as right-censored.

For each simulation scenario, 10,000 data sets of size 80, 150, and 300 were generated. The proposed change-point model was then fit to the data, using the two-stage estimation approach to obtain maximum likelihood estimates of all the model parameters. For comparison, model (1.7) assuming a linear change in standard deviation with no change-point and model (1.4) assuming constant standard deviation were also fit to the simulated data sets.

Table 1.1 presents the mean bias and standard deviation (SD) of the 10,000 estimates for every parameter in the model. The change-point model exhibited less bias in estimating the LOD than both the linear standard deviation and constant standard deviation models, for every simulation considered. The change-point model also produced LOD estimates with a smaller standard deviation than the linear standard deviation model for all simulations considered. The change-point model tended to slightly underestimate the limit of detection, particularly when the change-point was small relative to the range of the observed concentration values. This bias decreased as the change-point increased, a similar pattern as was observed with the constant standard deviation model. This characteristic was reversed for the linear standard deviation model, as the bias increased for larger values of the change-point. Increased sample size did not seem to affect the bias in any of the models considered, though the standard deviations of the LOD estimates decreased.

In addition to parameter estimates, the Akaike Information Criterion (AIC, Akaike 1974) was also calculated for each of the three models fit to every simulated data set. For each set of 10,000 simulated data sets, the model with the lowest AIC was selected as the best fit for the current data set. Table 1.1 displays the proportion of simulated data sets that resulted in a particular model having the best fit. For example, with N = 80 and $\lambda = 4.5$, the change-point model had the best model fit in 96.0% of the simulated data sets, compared to 0.6% for the linear standard deviation model and 3.4% for the constant standard deviation model. The results displayed in table 1.1 show that the change-point model produces the best fit to the data a much higher proportion of the time than either the linear standard deviation or constant standard deviation models, for all simulation scenarios. This "relative fit" of the change-point model increased with increasing change point, and also with sample size, from 60.3% in the N = 80, $\lambda = 1.5$ simulation to 100% in the N = 300, $\lambda = 4.5$ simulation.

1.5 HIV Data

Data for this analysis comes from an HIV pilot study analyzing the effects of a drug on HIV transcription. Resting cells from HIV infected patients are treated with the drug, with interest in the degree to which HIV transcription is increased. HIV RNA in general is too unstable and must be reverse transcribed into the more stable form, cDNA. The concentration of HIV RNA in patient samples is much too low to be directly measured, and following conversion to cDNA subsequent amplification by quantitative PCR is necessary (Nolan et al. 2006, Palmer et al. 2003). The region of the HIV genome amplified in this assay codes for a highly conservative region known as *gag* which is measured with primers and probes as described by Agarwal et al. (2007). RNA from patient samples are quantified using a standard curve with a known concentration of HIV cDNA. The PCR machine measures unknown quantities through fluorescence that is proportional to sample concentration and amplifies over many cycles. A cycle-threshold is defined as the PCR cycle that results in the highest increase in fluorescence. By comparing the cycle-threshold value for a given unknown concentration of RNA to a linear calibration curve for different known HIV concentrations, the unknown

concentration can be estimated.

For each patient in the pilot study, a linear calibration curve is created by measuring the cycle-threshold value for different **known** dilutions of HIV. Data for the study consists of calibration curve data for six experiments (one for each patient), with each experiment consisting of 20 measurements for each of 4-5 known concentrations of HIV. The goal of the analysis is to estimate the limit of detection for the concentration of HIV individually for each experiment. Complicating the analysis is the restriction that each sample was run for a maximum of 42 cycles of PCR amplification; HIV concentrations resulting in more than 42 cycles are right-censored. A plot of the raw data for all six experiments is presented in Figure 1.1.

It is important to note here that the concentration of HIV (X) is inversely related to the cycle-threshold value (Y) in the analyzed data. A lower concentration of HIV will take more PCR cycles to fluoresce, resulting in a higher cycle-threshold value. This relationship is the opposite of what is usually observed when relating known concentrations to measured values, where measurement (Y) usually increases with analyte concentration (X). Because of the inverse relationship between Y and X in the current data, the LOD estimates will be slightly altered from (1.13), taking the form:

$$\begin{aligned} \widehat{\text{LOD}}_Y &= \hat{\beta}_0 - 3(\hat{\sigma}_0^2 + \hat{\sigma}_{\beta_0}^2)^{1/2} \\ \widehat{\text{LOD}}_X &= \frac{-3(\hat{\sigma}_0^2 + \hat{\sigma}_{\beta_0}^2)^{1/2}}{\hat{\beta}_1} \end{aligned} \tag{1.13}$$

Analysis of the data was performed in two ways. First, the change-point model proposed in Section 1.3 was fitted separately for each individual experiment, generating experiment-specific LOD estimates. Additionally, a mixed-model approach was also considered. The mixed model allows for simultaneous estimation of the LOD for all experiments. The model specification is given as follows:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{i0} + b_{i1} X_{ij} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma_{ij}^2) \tag{1.14}$$

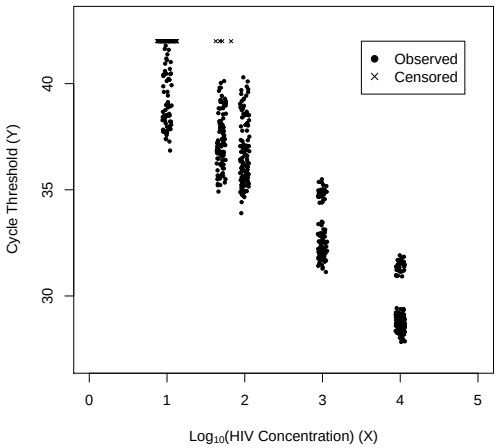


Figure 1.1: Plot of raw data from six experiments in the HIV study

$$\sigma_{ij} = \begin{cases} \sigma_0 & \text{if } X_{ij} \leq \lambda_i \\ \sigma_0 + \sigma_1(X_{ij} - \lambda_i) & \text{if } X_{ij} > \lambda_i \end{cases}$$

$$(b_{i0}, b_{i1}) \sim MVN\left(\mathbf{0}, \begin{bmatrix} \sigma_{b_0}^2 & \rho\sigma_{b_0}\sigma_{b_1} \\ \rho\sigma_{b_0}\sigma_{b_1} & \sigma_{b_1}^2 \end{bmatrix}\right)$$

where Y_{ij} is the cycle-threshold value and X_{ij} is the $\log_{10}$ concentration of HIV for experiment i and measurement j. The abbreviation MVN denotes a multivariate normal distribution. Maximum likelihood estimation for this model was performed using `PROC NLMIXED` in SAS software version 9.3. As with the simulation study, both linear standard deviation and constant standard deviation models were included for comparison. The model fit was again analyzed using the AIC.

Parameter estimates for both the regression and mixed model approaches are given in Table 1.2, and a plot of the model fit for experiments 1 and 3 is given in Figure 1.2. The dashed lines about the predicted regression line in Figure 1.2 represent a 95% prediction interval for the data, with the vertical and horizontal dashed lines representing the estimated LOD. Estimates of experiment-specific LODs (denoted $\widehat{LOD}_X$ in table 1.2) using the change-point model range from 0.468 to 1.195, which correspond to LOD estimates on the untransformed

HIV concentrations of 2.94 to 15.68 copies of *gag*. LOD estimates from the change-point model were lower than those from the linear standard deviation model, and were higher than estimates from the constant standard deviation model, for all experiments. Interestingly, the AIC for the change-point model was lower than the AIC for the linear standard deviation model in only one of the six experiments tested, suggesting that the linear standard deviation model generally provided a better fit to the data when the regression model was utilized. In experiments 1, 2, 5, and 6, the change-point estimates equal 1.0, the lowest observed concentration value. This makes the likelihood for the model identical to the linear standard deviation model (notice the identical parameter estimates for β_0 and β_1), only with more parameters estimated in the change-point model. This results in the higher AIC value for the change-point model.

The mixed model results also give LOD estimates for the change-point model that are higher than the constant standard deviation model, and lower than the linear standard deviation model. The un-logged LOD estimate of 15.49 is in the range of LOD estimates for the regression change-point models on each experiment, as expected. The AIC results indicate that the change-point model provides a better fit to the available data than does the linear standard deviation or constant standard deviation mixed models.

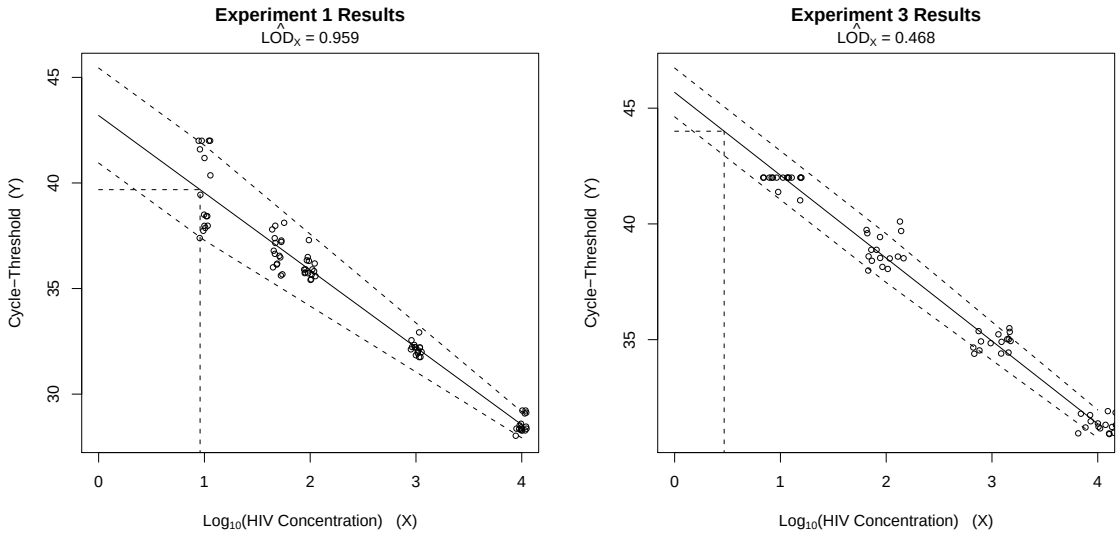


Figure 1.2: Change-point model results for experiments 1 and 3

Table 1.2: Parameter estimates for HIV study, using regression models and mixed models

				Regression Model				
Experiment	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{\lambda}$	$\widehat{LOD}_X$	$10^{\widehat{LOD}_X}$	AIC
Change-Point Model								
1	43.19	-3.661	1.147	-0.278	1.000	0.959	9.10	181.68
2	45.20	-4.104	1.612	-0.452	1.000	1.195	15.68	235.64
3	45.68	-3.579	0.537	-0.115	2.000	0.468	2.94	85.41
4	45.35	-3.472	0.747	-0.255	1.699	0.661	4.58	118.93
5	42.42	-3.395	1.186	-0.262	1.000	1.063	11.57	291.30
6	43.17	-3.684	1.262	-0.310	1.000	1.041	11.00	286.98
Linear Standard Deviation Model								
1	43.19	-3.661	1.425	-0.278	-	1.183	15.23	179.68
2	45.20	-4.104	2.062	-0.452	-	1.522	33.25	233.64
3	45.74	-3.593	0.676	-0.085	-	0.580	3.80	86.88
4	45.43	-3.492	1.118	-0.239	-	0.971	9.36	118.10
5	42.42	-3.395	1.448	-0.262	-	1.292	19.60	289.30
6	43.17	-3.684	1.572	-0.310	-	1.291	19.55	284.98
Constant Standard Deviation Model								
1	43.26	-3.688	0.920	-	-	0.776	5.96	219.63
2	44.95	-4.004	1.199	-	-	0.927	8.46	294.05
3	45.66	-3.563	0.464	-	-	0.409	2.56	88.08
4	45.27	-3.424	0.622	-	-	0.565	3.67	156.98
5	42.44	-3.406	0.920	-	-	0.830	6.76	326.45
6	43.53	-3.825	0.972	-	-	0.781	6.04	339.77
				Mixed Model				
Experiment	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{\lambda}_i$	$\widehat{LOD}_X$	$10^{\widehat{LOD}_X}$	AIC
Change-Point Model								
All	44.37	-3.698	1.342	-0.313	*	1.190	15.49	2955.9
Linear Standard Deviation Model								
All	44.38	-3.703	1.459	-0.250	-	1.273	18.75	2994.0
Constant Standard Deviation Model								
All	44.43	-3.725	0.973	-	-	0.910	8.13	3240.4

*λ is experiment-specific in this model

1.6 Discussion

In this chapter we have developed a change-point model to estimate the limit of detection with a linear calibration curve. In certain settings, the proposed approach may provide a more realistic modeling of the underlying distribution of measurement errors in a linear calibration curve. Estimation is performed via a two-stage estimation technique, such that the nonlinear constraints on the model parameters are satisfied. We have demonstrated application of the proposed model using both an individual regression model and a mixed model.

The simulation results presented in Table 1.1 demonstrate that the proposed change-point model can dramatically improve estimation of the limit of detection when compared to both the linear standard deviation and constant standard deviation models. When measurement error is constant for low concentrations of analyte, the linear standard deviation model tends to overestimate the measurement error for a blank sample, and consequently tends to overestimate the limit of detection. This is shown quite dramatically in Table 1.1, where estimates using the change-point model exhibit smaller bias than the linear standard deviation model, particularly when more of the observed data falls below the true change-point. The constant standard deviation model was shown to underestimate the LOD for all simulations considered, with a significantly larger bias than the change-point model. When AIC fit statistics were analyzed, the change-point model was correctly identified as the model providing the best fit to the data, for all simulations considered.

The key assumption of the proposed change-point model is that the measurement error standard deviation is constant below some low concentration value. If this assumption does not hold (the standard deviation instead continues to increase or decrease with concentration), the change-point model would be expected to exhibit a greater bias than the linear standard deviation model. In this case, when the measurement error increases with concentration, the change-point model would tend to overestimate the limit of detection. When the measurement error decreases with concentration, the change-point model would tend to underestimate the LOD.

The proposed linear regression change-point model is quite straightforward to implement,

and convergence of the parameter estimates was achieved very quickly in both the simulation and HIV analyses. The mixed model approach in Section 1.5 also converged very quickly, making the proposed approaches quite feasible. Both methods produced similar estimates of the limit of detection, suggesting that either would be appropriate for analysis.

Chapter 2

Maximum Likelihood Estimation in Generalized Linear Models With Multiple Covariates Subject to Detection Limits

2.1 Introduction

While the previous chapter concerned estimation of the limit of detection itself, the current topic considers analysis of data subject to detection limits with a predetermined limit of detection. Specifically, we are interested in estimation with generalized linear models (GLM's) in which multiple covariates are subject to a limit of detection. While the proposed methodology in this chapter can be applied to both right- and left-censored covariate data, the real and simulated examples presented here consider only left-censored data, as is most common in real-life studies with detection limits. To motivate these methods, we consider a study in cancer incidence conducted within the National Health and Nutrition Examination Survey (NHANES). As part of this study, levels of urinary heavy metals were recorded, along with presence of any form of cancer. Recorded urinary heavy metals included cadmium, uranium, tungsten, and dimethylarsonic acid. The measurement device used to examine levels of each urinary heavy metal can only be calibrated down to a specific limit of detection (i.e. only above 1.7 ug/L for dimethylarsonic acid). As a result, 24.1% of the 1350 patients had at least one covariate value that fell below the limit of detection for the measurement device. Study subjects were also surveyed as to past cancer status, the response variable for this study.

Past research on data subject to detection limits has considered models where either the response or covariates alone are subject to detection limits. The simplest and most straightforward method for dealing with such data is to remove or delete all observations

falling below the limit of detection. This is known as complete-case analysis. Complete-case analysis is generally discouraged because of the loss of useful information in the data. Though complete-case analysis can give unbiased parameter estimates in regression models (Rigobon and Stoker 2007, D'Angelo and Weissfeld 2008, Nie et al. 2010), the standard errors of those estimates will be inflated due to the decreased sample size. This deficiency is particularly significant for studies where a large proportion of data falls below the limit of detection. Additionally, background parameter estimates for the covariate distribution of interest will be biased (Helsel, 2005). Another very common approach is to use ad-hoc substitution methods. These often include substituting some fraction of the limit of detection for all observations falling below the limit of detection, such as the limit of detection itself (LOD), LOD/2, LOD/$\sqrt{2}$, or zero. Such methods are commonly employed because they are simple both to understand and implement. However, numerous authors have concluded that such methods are statistically inappropriate (see Helsel, 2006 and Lubin, 2004 for a discussion with censored responses, and Lynn, 2001 for censored covariates). Helsel (2005) provides a review of several of these substitution procedures, concluding that the substitution method leads to highly biased estimates and has no theoretical basis. Singh and Nocerino (2002) analyzed the substitution method on censored response values in environmental studies, concluding that highly biased estimates result even in cases with a small percent of censored values and only a single detection limit. The bias increases as more detection limits are introduced. For regression with a censored outcome, Thompson and Nelson (2003) found that substitution of half the detection limit led to biased parameter estimates and artificially small standard error estimates. These results have provided strong evidence against using ad-hoc substitution techniques.

In a linear regression setting, further substitution methods have been proposed for cases when a single covariate is subject to a limit of detection. Richardson and Ciampi (2003) proposed substituting the conditional expected value of each censored covariate, given all observed covariates. This method relies on a specification of the underlying covariate distribution, which often is not known with certainty. When the covariate distribution is unknown, Schisterman (2006) proposed substituting the average of all *observed* covariates in the model,

which was shown to achieve unbiased results. Another common method is maximum likelihood (ML) estimation, which also requires knowledge of the underlying covariate distribution. These methods were compared with the previously discussed ad-hoc substitution methods in Nie et. al (2010) when only one covariate is subject to a limit of detection. It concluded that maximum likelihood performed best when the covariate distribution is known, as ML estimation is unbiased and results in small standard errors. These results were echoed by Lynn (2001), who compared substitution methods to multiple imputation and maximum likelihood estimation. Both papers noted that maximum likelihood estimation should not be attempted when the underlying covariate distribution is not known. In this case, Nie et al. (2010) suggests using complete-case analysis.

The preference for maximum likelihood approaches has also been seen in studies using logistic regression with a single covariate subject to a limit of detection. Cole et al. (2009) compared ad-hoc substitution methods to complete-case analysis and maximum likelihood estimation, concluding that maximum likelihood resulted in relatively unbiased estimates with smaller standard errors than either complete-case or substitution methods, especially when the proportion of censored values was large (50% or more).

Methods have also been proposed for Cox Regression models with up to two covariates subject to a lower limit of detection. D'Angelo and Weissfeld (2008) presents an index-based EM Algorithm-type method for this problem. The E-step for this method involves substituting the conditional expectation of each censored covariate, while the M-step uses standard Cox regression. It found that the index-based approach provided improvements over complete-case analysis in terms of variance estimates, but that a small bias existed in the index approach compared to the unbiased complete-case analysis. The approach was not shown to provide much improvement over the biased LOD/2 and LOD/$\sqrt{2}$ substitution approaches, however.

When the response variable is subject to a limit of detection, two common methods of estimation include Tobit Regression (Tobin, 1958) and multiple imputation. Generally, Tobit Regression is used when interest resides primarily on the regression parameters. When interest is on estimating a "complete" dataset, however, multiple imputation is often used to impute the missing values. Lubin et al. (2004) developed a multiple imputation approach based

on bootstrapping, and compared the results to substitution methods and Tobit regression. It found that both the proposed multiple imputation approach and Tobit Regression have reduced biases with respect to other ad-hoc substitution methods.

All the methods previously mentioned here concern models with either a censored response and fully-observed covariates, or a fully-observed response and *at most* 2 censored covariates. To the authors knowledge, no general likelihood-based approach has been developed to account for a large number of left-censored covariates in a generalized linear model. In this chapter, we investigate maximum likelihood methods for fitting models with covariates subject to a limit of detection. We show that this maximum likelihood estimation can be carried out directly via an EM algorithm called the *EM by the Method of Weights* (Ibrahim, 1990). For covariates subject to a limit of detection, we specify the covariate distribution via a sequence of one dimensional conditional distributions. We discuss the missing data mechanism for censored data and explain how the notion of missingness differs from that of regular missing data problems.

In this chapter, we propose a method for estimating parameters in generalized linear models with censored covariates and an effectively ignorable missing data mechanism. We consider the case of continuous covariates only in this work, because censored categorical covariates are unlikely to occur in real-world applications. Following Lipsitz and Ibrahim (1996), the joint covariate distribution is modeled via a sequence of one dimensional conditional distributions. Modeling the joint covariate distribution in this fashion facilitates a more straightforward specification of the distribution. The response variable is assumed to be completely observed, though our method can be easily extended to the case where the response is subject to a limit of detection. We derive the E and M steps of the EM algorithm with effectively ignorable missing covariate data. For continuous covariates, we use a Monte Carlo version of the EM algorithm to obtain the maximum likelihood estimates via the Gibbs sampler. We derive the E-step for the Monte Carlo version of EM. In addition, we show that the relevant conditional distributions needed for the E-step are log-concave, so that the Gibbs sampler is straightforward to implement when the covariates are continuous. The work presented in this chapter is an extension of the methods proposed for missing data in Ibrahim, Lipsitz, and

Chen (1999). The proposed methods are computationally feasible and can be implemented in a straightforward fashion.

The rest of this chapter is organized as follows. In Section 2.2, we given some general notation for generalized linear models. In Section 2.3, we discuss the proposed methods of estimation and give a detailed discussion of the various models used. In Section 2.4, we demonstrate the methodology with a simulation study involving a linear regression model. In Section 2.5, we demonstrate the methodology with an example involving real data. We conclude the chapter with a discussion section.

2.2 Notation for GLM's

In this chapter, we will take $(x_1, y_1), ..., (x_n, y_n)$ as a set of n independent observations, with y_i representing the response variable and x_i representing a p x 1 vector of covariates. The joint distribution of (y_i, x_i) is written as a sequence of one-dimensional conditional distributions $[y_i|x_i]$ and $[x_i]$, representing the conditional distribution of y_i given x_i and the marginal distribution of x_i. The notation $p(y_i|x_i)$ is used throughout the chapter to denote the conditional density of y_i given x_i.

The conditional distribution $[y_i|x_i]$ is specified by a $k \times 1$ parameter vector θ, with the conditional density being represented as $p(y_i|x_i, \theta)$. For the class of generalized linear models, the parameter vector θ is usually specified as $\theta = (\beta, \tau)$, with β representing the regression model coefficients and τ representing the dispersion parameter. The logistic, poisson, and exponential models have a τ value exactly equal to one; in these cases, β and θ are equal. For nonlinear models with a normal errors, we write the parameter vector as $\theta = (\theta^*, \sigma^2)$, with θ^* representing the expectation parameters and σ^2 representing the variance of the errors.

The marginal density for x_i is taken as $p(x_i|\alpha)$, with α representing the parameters for the marginal distribution of x_i. The joint density for (y_i, x_i) can then be represented by the following sequence of conditional densities for subject i.

$$p(y_i|x_i) = p(y_i|x_i, \theta)p(x_i|\alpha) \tag{2.1}$$

Combining this formula for all subjects leads to the complete-data log-likelihood:

$$\begin{aligned} l(x, y|\gamma) &= \sum_{i=1}^{n} l(x_i, y_i|\gamma) \\ &= \sum_{i=1}^{n} \log\left[p(y_i|x_i, \theta)\right] + \log\left[p(x_i|\alpha)\right]. \end{aligned} \tag{2.2}$$

Here $l(x_i, y_i|\gamma)$ represents the log-likelihood contribution for subject i, and $\gamma = (\theta, \alpha)$. In the present analysis, our primary interest is in estimating θ; here, α is considered a nuisance parameter.

Extending this notation to censored covariate data, we write $x_i = (x_{cens,i}, x_{obs,i})$, where $x_{obs,i}$ are the fully observed covariates, and $x_{cens,i}$ is a $q_i \times 1$ vector of censored covariates. For individual censored covariate values, we use the notation $x_{cens,i} = (x^*_{i1}, ..., x^*_{iq_i})$. We allow a different censoring interval for each covariate and subject, taking (c_{lij}, c_{uij}) as the censoring interval for subject i and covariate j. We note here that the censoring intervals are considered to be fully known here. In some applications, limits of detection are not known explicitly, and must be estimated. We also note that in most cases the censoring intervals will not vary across subjects, this is included for generality. This notation is easily generalized to right or left-censoring. For left-censored covariates, take $c_{lij} = -\infty$. For right-censored covariates, take $c_{uij} = \infty$. We use the shorthand notation $(c_l < x_{cens,i} < c_u)$ to denote that each element of $x_{cens,i}$ takes a value within its respective censoring interval. That is:

$$(c_l < x_{cens,i} < c_u) \equiv \bigcap_{x_{ij} \in x_{cens,i}} (c_{lij} < x_{ij} < c_{uij})$$

2.3 Covariate Data Subject to a Limit of Detection

We now propose maximum likelihood methods for covariate data subject to a limit of detection. We will allow left, right, or interval censoring on each covariate, and for ease of exposition will assume that $\tau = 1$. For clarity, we develop the methodology here for the class of generalized linear models.

Suppose $y_1, ..., y_n$ are independent and

$$p(y_i|x_i,\beta) = \exp\left\{y_i\theta(x_i'\beta) - b(\theta(x_i'\beta))\right\}$$

for $i = 1, ..., n$. In general, the EM-algorithm maximizes the expected value of the complete data log-likelihood of (y_i, x_i), given the observed data, i.e.,

$$\begin{aligned} Q(\gamma|\gamma^{(t)}) = \sum_{i=1}^{n} E\Big[&\log[p(y_i|x_i,\beta)] + \\ &\log[p(x_i|\alpha)]|\text{observed}_i, \gamma^{(t)}\Big] \end{aligned} \tag{2.3}$$

Unlike the usual missing covariate problem in which the 'observed data' for subject i is $(y_i, x_{obs,i})$, in the censored covariate problem the 'observed data' are $(y_i, x_{obs,i})$ and $(c_l < x_{cens,i} < c_u)$. In the usual missing covariate problem with $x_{mis,i}$ completely missing, the 'weights' in the EM by the Method of Weights are the conditional probabilities $p(x_{mis,i}|x_{obs,i}, y_i, \gamma)$. Now, with the additional information that $(c_l < x_{cens,i} < c_u)$ in the censored covariate problem, the weights are the conditional probabilities $p[x_{cens,i}|x_{obs,i}, (c_l < x_{cens,i} < c_u), y_i, \gamma]$.

If the censored covariates are all continuous (the most common case), then the E-step of the EM algorithm consists of an integral, which typically does not have a closed form for GLM's. We can write the E-step for the i^{th} observation as

$$\begin{aligned} Q_i(\gamma|\gamma^{(t)}) = &\\ &\int \log[p(y_i|x_i,\beta)]p(x_{cens,i}|x_{obs,i}, y_i, \gamma^{(t)}) \\ &\qquad \times I(c_l < x_{cens,i} < c_u)dx_{cens,i} \\ &+ \int \log[p(x_i|\alpha)]p(x_{cens,i}|x_{obs,i}, y_i, \gamma^{(t)}) \\ &\qquad \times I(c_l < x_{cens,i} < c_u)dx_{cens,i} \\ &= Q_i^{(1)}(\beta|\gamma^{(t)}) + Q_i^{(2)}(\alpha|\gamma^{(t)}). \end{aligned} \tag{2.4}$$

We note here that in the above equation, $x_{cens,i}$ is a vector consisting of all covariates in observation i that fall within their respective censoring intervals. In cases where $x_{cens,i}$ contains more than a single censored covariate, equation (2.4) consists of multiple integrations, one over each censored covariate, integrating over the range of the censoring interval. For example, with 3 censored covariates ($x_{cens,i} = (x^*_{i1}, x^*_{i2}, x^*_{i3})$), we have:

$$\begin{aligned} p(x_{cens,i}|x_{obs,i}, y_i, \gamma^{(t)}) &= p(x^*_{i1}|x^*_{i2}, x^*_{i3}, ...) \\ &\times p(x^*_{i2}|x^*_{(i3)}, ...) \times p(x^*_{i3}|...) \end{aligned}$$

and

$$\begin{aligned} Q_i(\gamma|\gamma^{(t)}) = & \\ & \int\int\int \log[p(y_i|x_i, \beta)] p(x_{cens,i}|x_{obs,i}, y_i, \gamma^{(t)}) \\ & \qquad \times I(c_l < x_{cens,i} < c_u) dx^*_{i1} dx^*_{i2} dx^*_{i3} \\ & + \int\int\int \log[p(x_i|\alpha)] p(x_{cens,i}|x_{obs,i}, y_i, \gamma^{(t)}) \\ & \qquad \times I(c_l < x_{cens,i} < c_u) dx^*_{i1} dx^*_{i2} dx^*_{i3} \end{aligned}$$

From this, it should be clear that closed-form solutions to equation (2.4), even if available (i.e. for a small number of censored covariates), are complicated, and the maximization can be very difficult. We now propose a general approach to evaluating equation (2.4), regardless of the number of censored covariates.

To evaluate (2.4) at the $(t+1)^{st}$ iteration of EM, we use the Monte Carlo version of the EM algorithm given by Wei and Tanner (1990). To do this, we first need to generate a sample from the truncated distribution $[x_{cens,i}|x_{obs,i}, y_i, \gamma^{(t)}]I(c_l < x_{cens,i} < c_u)$. This truncated distribution is log-concave in each component of $x_{cens,i}$ for most link functions. Thus we can use the Gibbs sampler along with the adaptive rejection metropolis algorithm (ARMS) of Gilks, Best, and Tan (1995) to successively sample from the truncated distribution $[x_{cens,ij}|x_{cens,ik}, k \neq j, x_{obs,i}, y_i, \gamma^{(t)}]I(c_l < x_{cens,i} < c_u)$, where $x_{cens,ij}$ denotes the j^{th}

component of $x_{cens,i}$.

The ARMS algorithm is an extension of the Adaptive Rejection Sampling (ARS) algorithm of Gilks and Wild (1992), and can sample values from complex likelihood functions which are not required to be log-concave. ARMS works by constructing an envelope function around the desired log-density. It performs rejection sampling on the envelope function, shrinking the envelope around the desired log-density with each successive sample. For log-densities that are not concave, the ARMS algorithm performs an additional Metropolis step on each potential sampled value (Metropolis, 1953). The shrinking envelope function provides an efficient means of sampling from a complicated log-density, without having to evaluate each point of the density directly. ARMS also allows for straightforward sampling from truncated distributions, as all potential points falling outside the censoring interval are immediately rejected.

Use of the EM algorithm requires complete sampled data for each of the n observations in the dataset. For observation i, a new sample must be obtained for each of the q_i censored covariate within $x_{cens,i}$. This is done by successively sampling from the distribution of $x_{cens,ij}, j = 1, ..., q_i$ until a new sample vector z_i is obtained for the censored vector $x_{cens,i}$. The sampled vector z_i contains q_i sampled values, one for each of censored covariates in $x_{cens,i}$. Now, suppose for the i^{th} observation, we take a sample of size $m_i, z_{i1}, ..., z_{im_i}$, from the truncated distribution $[x_{cens,i}|x_{obs,i}, y_i, \gamma^{(t)}]I(c_l < x_{cens,i} < c_u)$ via the Gibbs sampler in conjunction with the adaptive rejection algorithm. We note here that each z_{ik} is a $q_i \times 1$ vector for each $k = 1, ..., m_i$, with q_i representing the length of $x_{cens,i}$. The E-step for the i^{th} observation at the $(t+1)^{st}$ iteration for the GLM can be written as

$$\begin{aligned} Q_i(\gamma|\gamma^{(t)}) &= \frac{1}{m_i}\sum_{k=1}^{m_i} l(z_{ik}, x_{obs,i}, y_i, \gamma). \\ &= Q_i^{(1)}(\beta|\gamma^{(t)}) + Q_i^{(2)}(\alpha|\gamma^{(t)}). \end{aligned} \tag{2.5}$$

We notice that this E-step is the EM by the Method of Weights with each $x_{cens,i}$ being filled in by a set of m_i values each contributing a weight $1/m_i$. The M-step then maximizes

equation 2.3, which can be expressed as

$$Q(\gamma|\gamma^{(t)}) = \sum_{i=1}^{n} Q_i(\gamma|\gamma^{(t)})$$

The maximization can be performed first by taking

$$\dot{Q}(\gamma|\gamma^{(t)}) = (\dot{Q}^{(1)}(\beta|\gamma^{(t)}), \dot{Q}^{(2)}(\alpha|\gamma^{(t)}))'$$

as the $q \times 1$ gradient vector of $Q(\gamma|\gamma^{(t)})$. This can be calculated by taking

$$\begin{aligned} \dot{Q}(\gamma|\gamma^{(t)}) &\equiv \sum_{i=1}^{n} \dot{Q}_i(\gamma|\gamma^{(t)}) \\ &= \sum_{i=1}^{n} \frac{1}{m_i} \sum_{k=1}^{m_i} \frac{\partial}{\partial \gamma} l\{z_{ik}, x_{obs,i}, y_i, \gamma\} \end{aligned} \tag{2.6}$$

Using this procedure, the EM algorithm can then be run until convergence. In practical application, the maximization of the weighted log-likelihood (with respect to the model parameters) can often be performed by standard software.

Also here we let $\ddot{Q}(\gamma|\gamma^{(t)})$ denote the $q \times q$ matrix of the second derivatives of $Q(\gamma|\gamma^{(t)})$. Let $\hat{\gamma}$ denote the estimate of γ at convergence. The asymptotic covariance matrix can then by calculated by the method of Louis (1982). The estimated observed information matrix of γ based on the observed data is taken as

$$\begin{aligned} I(\hat{\gamma}) = &- \ddot{Q}(\hat{\gamma}|\hat{\gamma}) \\ &- \left\{ \sum_{i=1}^{n} \sum_{x_{cens,i,j}} \frac{1}{m_i} S_i(\hat{\gamma}; x_i, y_i) S_i(\hat{\gamma}; x_i, y_i)' \right. \\ &\left. - \sum_{i=1}^{n} \dot{Q}_i(\hat{\gamma}|\hat{\gamma}) \dot{Q}_i(\hat{\gamma}|\hat{\gamma})' \right\} \end{aligned}$$

where

$$S_i(\hat{\gamma}; x_i, y_i) = \left[\frac{\partial l(\gamma; x_i, y_i)}{\partial \gamma}\right]_{\gamma=\hat{\gamma}}$$

The estimate of the asymptotic covariance matrix is then calculated as $I(\hat{\gamma})^{-1}$.

We note here that the E-step for censored data is different from the standard missing data notation. Specifically, the censored data E-step in equation (2.4) omits the $\int log[p(r_i|y_i, x_i, \phi)]...dx_{cens,i}$ section used in missing data problems, where r_i represents an indicator for missingness. This is because the notion of ignorability is fundamentally different in detection limit problems when compared to missing data problems. In detection limit problems it is generally assumed that the detection limits are known values. With detection limits known, the probability of censoring ("missingness" in the missing data case) clearly depends on the true value of the covariate (x_i), suggesting a non-ignorable mechanism. However, in the detection limits case the true value of x_i *explicitly* determines whether or not the value is censored. The value of $p(r_i|x_i)$ is either 0 or 1, for all values of x_i. It follows that the non-ignorable component of the E-step equation for missing data is omitted in the detection limit case.

It should be noted that having a continuous outcome variable also subject to a limit of detection only marginally complicates the situation at hand. In this case, the E-step requires an additional integration over the possible values of the censored outcome. Equation (2.4) then becomes:

$$\begin{aligned} Q_i(\gamma|\gamma^{(t)}) = &\int\int \log[p(y_i, |x_i, \beta)] \ldots dx_{cens,i} dy_{cens,i} \\ &+ \int\int \log[p(x_i|\alpha)] \ldots dx_{cens,i} dy_{cens,i} \end{aligned}$$

This situation is further simplified when sampling from the distribution of an outcome value below the detection limit, however, because we are dealing with the class of generalized linear models. The distribution of the outcome given the covariates and parameters is assumed to come from an exponential family. Therefore, the distribution of an outcome value below the limit of detection is just a truncated form of a well-known distribution, be it normal, gamma, etc. Such sampling is straightforward.

In this chapter, we will investigate the maximum likelihood estimation with censored

covariates as outlined above. We will study the EM algorithm for this problem and consider GLM's with covariates subject to a detection limit. Examples analyzed include both linear and logistic regression.

2.4 Simulation Study

Here we consider a simple linear model involving six covariates:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma_y^2)$. The response y_i is fully observed, as are the first three covariates $x_1, ..., x_3$. The last three covariates $x_4, ..., x_6$ are subject to a prespecified detection limit. Detection limits are specified according to a desired overall censoring percentage. In this case, detection limits were chosen such that 30% and 50% of observations had at least one covariate that fell below the limit of detection. The covariate distribution was specified as multivariate normal, with arbitrary prespecified parameter values and correlated observations with $0.3 \leq |\rho| \leq 0.7$ for all covariate pairs. Using this specification, datasets of size 200 were then generated; covariate values for $x_4 - x_6$ falling below the detection limit were set as missing.

Each simulation presented in this chapter was performed on 1000 datasets created as described above, each from identical background parameter distributions and detection limits. The *EM by Method of Weights* was then applied to each dataset. Initial parameter estimates for the model and covariate distribution were taken from a complete-case analysis of the data. These were passed to the ARMS algorithm as parameters in the initial iteration of the EM algorithm. For each observation with at least one covariate falling below the limit of detection, ARMS was used to generate $m_i = 250$ samples of complete covariate data. For observations with a single covariate falling below the limit of detection, these samples were taken from the distribution of $x_{cens,i}|x_{obs,i}, y_i, \gamma^{(t)}$ truncated over the censoring interval. For each observation with multiple covariates falling below the limit of detection, ARMS was used

Table 2.1: Parameter estimates and standard errors, comparing EM algorithm approach to complete-case analysis and substitution of LOD/$\sqrt{2}$. 1000 Datasets were used for each analysis, with 250 samples taken for each observation below the limit of detection.

30% Below Limit of Detection												
		Maximum Likelihood Estimation (MLE)				Complete Case (CC)			Substitution LOD/$\sqrt{2}$			Significant?
Parameter	True	Estimate	SE	Boot SE*	P-Value	Estimate	SE	P-Value	Estimate	SE	P-Value	MLE/CC
β_0	1.00	1.0105	0.3161	0.3275	0.0014	1.0029	0.3771	0.0078	0.9242	0.1996	<.0001	Yes/Yes
β_1	-0.75	-0.7517	0.0782	0.0794	<.0001	-0.7504	0.0910	<.0001	-0.6875	0.0842	<.0001	Yes/Yes
β_2	0.26	0.2576	0.1089	0.1090	0.0180	0.2604	0.1241	0.0359	0.0684	0.1118	0.5405	Yes/No
β_3	-0.17	-0.1696	0.0881	0.0853	0.0541	-0.1677	0.1003	0.0944	-0.3128	0.0830	0.0002	Yes/No
β_4	3.00	3.0042	0.1191	0.1191	<.0001	3.0001	0.1416	<.0001	2.9359	0.1175	<.0001	Yes/Yes
β_5	0.20	0.2001	0.0841	0.0821	0.0173	0.1990	0.1024	0.0520	0.1661	0.1020	0.1036	Yes/No
β_6	-0.60	-0.6000	0.0853	0.0831	<.0001	-0.5988	0.1040	<.0001	-0.8589	0.0637	<.0001	Yes/Yes
σ_y^2	0.50	0.4795	0.0519	0.0499	<.0001	0.4958	0.0598	<.0001	0.4255	0.0512	<.0001	Yes/Yes
50% Below Limit of Detection												
		EM Algorithm				Complete Case			Substitution LOD/$\sqrt{2}$			Significant?
Parameter	True	Estimate	SE	Boot SE*	P-Value	Estimate	SE	P-Value	Estimate	SE	P-Value	MLE/CC
β_0	1.00	1.0215	0.3360	0.3440	0.0024	1.0097	0.4788	0.0350	0.3729	0.1068	0.0005	Yes/Yes
β_1	-0.75	-0.7531	0.0848	0.0842	<.0001	-0.7473	0.1123	<.0001	-0.5715	0.1005	<.0001	Yes/Yes
β_2	0.30	0.2928	0.1153	0.1144	0.0111	0.2982	0.1502	0.0471	0.0721	0.1505	0.6318	Yes/No
β_3	-0.19	-0.1966	0.0933	0.0896	0.0350	-0.1937	0.1165	0.0964	-0.3495	0.1051	0.0009	Yes/No
β_4	3.00	3.0100	0.1283	0.1270	<.0001	3.0053	0.1712	<.0001	2.7380	0.1381	<.0001	Yes/Yes
β_5	0.20	0.1999	0.0925	0.0888	0.0307	0.1929	0.1343	0.1511	0.2046	0.1410	0.1469	Yes/No
β_6	-0.60	-0.6062	0.0921	0.0893	<.0001	-0.6001	0.1297	<.0001	-0.9548	0.0889	<.0001	Yes/Yes
σ_y^2	0.50	0.4786	0.0551	0.0538	<.0001	0.4994	0.0720	<.0001	0.3352	0.0512	<.0001	Yes/Yes

*SE = standard error.

sequentially to sample from the distribution for each censored covariate until a new complete sample of covariate values was produced. The $m_i = 250$ samples from each censored observation were then combined, creating an augmented dataset of fully observed observations along with sampled values. The M-step of the EM algorithm was then performed via a weighted maximum likelihood estimation. Weights of 1 were used for each fully observed observation, and 1/250 was used for each sampled observation. This weighted maximum likelihood procedure produced new estimates for β in the model, along with updated parameter estimates for the covariate distribution. The updated covariate parameter estimates were then passed back to ARMS as the estimates for the following E-step, and the procedure was run iteratively until convergence.

Convergence of this algorithm was checked by calculating the average β estimate for the previous 10 iterations. This average was compared to the β average for the 10 iterations prior. In other words, at iteration t the mean beta values from t:(t-9) are compared to values from (t-10):(t-19). A difference of $\leq 10^{-3}$ was used for convergence. After convergence was reached for all parameters, final β estimates were taken as the average of the previous 10 estimates of β in the chain.

Bootstrap standard errors were calculated for each parameter in the dataset, for comparison to the standard error of the estimates obtained. For each of the 1000 datasets in a simulation, 25 bootstrapped datasets of size n = 200 were generated. The proposed EM algorithm was then run on each bootstrapped dataset, and final β estimates were obtained. The standard error for each population of 25 β estimates was then calculated for each parameter in the model. The mean of these standard errors were then taken as the final bootstrap standard error estimate for the model, and are used for comparison with the normal β standard error (SE) from the proposed maximum likelihood approach.

Table 2.1 displays results from analysis on all 1000 datasets. Final estimates and variances for each parameter are calculated as the mean and variance of final beta estimates for all 1000 datasets. The true prespecified parameter values are given, along with variance estimates calculated using the bootstrap procedure described above. Results are also presented for an ad-hoc substitution of LOD/$\sqrt{2}$ for each covariate falling below the limit of detection,

along with a complete-case analysis. As expected, both the maximum-likelihood approach and complete-case analysis appear to be largely unbiased, while the substitution approach produced very biased estimates. Maximum likelihood resulted in standard errors for the parameter estimates that were lower than those obtained with complete case analysis, and similar standard errors to the substitution approach. In addition, all calculated standard error estimates for maximum likelihood are close to the asymptotic bootstrapped estimates. The reduction in standard error seen with the maximum likelihood approach was large enough to result in a change in statistical significance (here taken at the $\alpha = 0.05$ level) for several parameters in the model when compared with the complete-case analysis. These conclusions hold for both 30% and 50% censored observations, suggesting that the benefit seen is robust to the degree of censoring observed. The EM-algorithm was also observed to converge rather quickly using the described criterion. Only 28 EM iterations were needed on average for all model parameters to converge.

2.5 NHANES Data

Here we consider data from the National Health and Nutrition Examination Survey (NHANES) concerning the effect of urinary heavy metal levels on cancer status. The survey years considered here are 2005-2006. The outcome variable in this study is cancer status, a binary variable recorded via questionnaire to the question "Have you ever been told by a doctor or other health professional that you had cancer or malignancy of any kind?". Urinary heavy metals were recorded via physical examination. The measurement device for each urinary heavy metal in the study can only be calibrated down to a specific limit of detection (LOD), leading to many left-censored observations. The degree of censoring varied by each covariate. The urinary heavy metals analyzed in this study include Dimethylarsonic Acid (13.7% below LOD), Cadmium (5.3% below LOD), Tungsten (10.7% below LOD), and Uranium (9.6% below LOD). In total, 24.1% of the 1350 patients in the study had at least one urinary heavy metal value that fell below a limit of detection. A logistic regression model was chosen for analysis, to predict the binary outcome measure of cancer status. Other covariates

included in the model are gender, race (dichotomized to white/nonwhite), physical activity (dichotomized survey response for any physical activity during an average day), and current nicotine use (yes/no). A log transformation was performed on each of the urinary heavy metals variables prior to modeling, and a multivariate normal prior distribution was assumed for these continuous covariates. An independent bernoulli prior was assumed for the binary covariates gender, race, and smoking status.

Initial parameter estimates for the model were taken from a complete-case analysis. Every observation with a urinary heavy metal covariate value falling below the LOD was then sampled $m_i = 250$ times using the ARMS algorithm. For observations with multiple covariate values below the LOD, each missing covariate value was consecutively sampled until a complete sampled observation was obtained. In such cases, 250 complete sampled observations were recorded. A weighted logistic regression model was then fit to the data, and MLE estimates and standard errors were obtained. Parameter estimates for the prior distributions were updated, and the procedure was run iteratively until convergence of the logistic model parameter estimates. The convergence criterion used here was identical to the procedure detailed in Section 2.4. Upon convergence, final beta estimates and standard errors were taken as the average estimates of the previous 10 iterations.

Table 2.2 summarizes the results of this study again comparing the maximum likelihood approach to both a complete-case analysis and ad-hoc substitution of $\text{LOD}/\sqrt{2}$. The substitution of $\text{LOD}/\sqrt{2}$ is particularly relevant in this case, as urinary heavy metals falling below the limit of detection are actually reported by the NHANES researchers as $\text{LOD}/\sqrt{2}$ in the available public data releases. As can be seen, the maximum likelihood approach results in significantly smaller standard errors for the parameter estimates when compared to complete-case analysis, and slightly smaller than those obtained via substitution with $\text{LOD}/\sqrt{2}$. This leads to a change in statistical significance (at the $\alpha = 0.05$ level) for the effect of Tungsten on cancer status. In this simulation, 30 EM-iterations were needed for convergence of all model parameters. It should be noted here that the ML standard errors reported in table 2.2 are based on only one simulation, and are calculated via a straightforward fitting of the weighted logistic regression model at convergence. Standard errors for the simulation study reported in

table 2.1 were calculated as the standard error of the population of 1000 final beta estimates, one for each simulated dataset. These estimation procedures are not equivalent, and it is important to note this difference.

It should also be noted that the fitted model used here does not include age as a covariate in the prediction of cancer status. A logistic model including the age covariate was also fit to this data, and age was found to be highly significant. The current model (without an age covariate) has been included here to more clearly display the potential benefits of the proposed methodology.

2.6 Discussion

In this chapter, we have proposed a method of maximum-likelihood estimation in generalized linear models with an unlimited number of covariates subject to a limit of detection. We have proposed models for the joint covariate distribution, which is based on a sequence of one-dimensional conditional distributions. The methodology presented here can be easily extended to cases where both the response and the covariates are subject to a limit of detection. The maximum likelihood approach presented here is much simpler computationally than a direct computation by way of the observed-data likelihood, especially for cases with multiple covariates subject to a LOD. When only a single covariate (or just the response) is subject to a LOD, closed-form solutions can often be used.

For the example considered in Section 2.4, the variance estimates for β are significantly improved over the complete case analysis. This result was echoed in our real-life analysis of NHANES data. This improvement can clearly lead to higher statistical power in studies that include data subject to detection limits.

A consistent drawback to maximum likelihood estimation in GLMs with data subject to detection limits is that a new algorithm needs to be created for each individual analysis that is performed. For sampling within ARMS, the log-likelihood function for the model of interest needs to be explicitly specified. In cases where the covariates are considered to follow a multivariate normal distribution, for example, the log-likelihood function is consistent

Table 2.2: Logistic regression model summary for NHANES data, comparing maximum likelihood approach to complete case analysis and ad-hoc substitution of LOD/$\sqrt{2}$

Parameter	Method	Estimate	SE	P-Value	Significant
Intercept	Complete Case	-0.6047	0.7954	0.4471	No
	Substitution	-0.6562	0.7009	0.3492	No
	ML	-1.5459	0.6765	0.0221	Yes
Gender	Complete Case	-0.0035	0.2305	0.9879	No
	Substitution	-0.0934	0.2041	0.6472	No
	ML	0.2730	0.1990	0.1703	No
Race	Complete Case	-1.6468	0.2789	<.0001	Yes
	Substitution	-1.6022	0.2516	<.0001	Yes
	ML	-1.2572	0.2288	<.0001	Yes
Physical Activity	Complete Case	-0.2641	0.2379	0.2671	No
	Substitution	-0.1984	0.2121	0.3494	No
	ML	-0.3139	0.2048	0.1250	No
Nicotine	Complete Case	-1.1471	0.3221	0.0004	Yes
	Substitution	-1.1103	0.2798	0.0001	Yes
	ML	-1.1710	0.2849	<.0001	Yes
Dimethylarsonic Acid	Complete Case	-0.2309	0.1856	0.2133	No
13.7% below LOD	Substitution	-0.1969	0.1515	0.1936	No
	ML	-0.0443	0.1389	0.7449	No
Cadmium	Complete Case	0.6172	0.1465	<.0001	Yes
5.3% below LOD	Substitution	0.7236	0.1225	<.0001	Yes
	ML	0.4812	0.1159	<.0001	Yes
Tungsten	Complete Case	-0.2689	0.1493	0.0717	No
10.7% below LOD	Substitution	-0.1958	0.1234	0.1126	No
	ML	-0.2400	0.1157	0.0389	Yes
Uranium	Complete Case	0.0769	0.1396	0.5817	No
9.6% below LOD	Substitution	0.0297	0.1249	0.8120	No
	ML	0.0016	0.1205	0.9980	No

and straightforward. However, more complicated covariate distributions will require a less standard computation of the log-likelihood, which can take significant additional time and can lead to error.

For both the simulation study and real-data analysis presented here, $m_i = 250$ samples were taken for each observation with covariates below a limit of detection. Based on the authors experience and other extensive simulations performed with this type of data, we feel that a sample size of at least $m_i = 100$ is necessary for accurate inference.

The computing time required to achieve EM convergence here clearly depends on the number of covariates in a model, the degree of censoring that is observed, and the number of samples that are taken for each censored observation. The simulation presented in Section 2.4 tended to converge quickly, with only an average of 28 iterations performed per dataset. This simulation of 1000 datasets took about 16 hours to complete on a Lenovo laptop with a dual-core Pentium processor, making this approach very computationally feasible.

While the analyses presented here discuss applications to generalized linear models, much interest exists in studies of longitudinal and survival data where covariates are subject to a limit of detection. Future research should be performed to extend the methodology presented here to such models.

Chapter 3

Joint Modeling of Longitudinal and Survival Data with Missing and Left-Censored Time-Varying Covariates

3.1 Introduction

In many longitudinal studies, time to event data is recorded in addition to the longitudinal and baseline covariates. In such studies, interest often lies in understanding the relationships between the longitudinal history of a process and it's effect on the risk of an event. For analysis of this type of data, a class of models called *joint models* has been developed, which *jointly* model both components simultaneously. Joint modeling has been used most extensively in studies of subjects with Human Immunodeficiency Virus (HIV, Wulfsohn and Tsiatis 1997, DeGruttola and Tu 1994, etc.). As with any large dataset, and particularly in the case of longitudinal data, it is often the case that a high degree of covariate and response data is missing. Additionally, in an HIV positive individual the measurement of viral load (the amount of virus in the blood) is only accurate down to a particular limit of detection (LOD). Values below the limit of detection cannot be reliably quantified or distinguished from a "blank" blood sample with no virus. In many cases (Wulfsohn and Tsiatis 1997, etc.) any missing covariate data is usually omitted from the analysis, and estimation proceeds on the complete data. However, in cases where a high degree of covariate data is missing, a great deal of information is lost when a model is only fit on complete data. Additionally, analysis using only complete data often requires the strong assumption that missing observations are missing completely at random (MCAR, Little 1995) for these imprecise inferences to be valid.

The goal of the analysis presented in this chapter is to develop a joint modeling strategy that accounts for both missing and left-censored time-varying covariates. This analysis

is motivated by data from the Multicenter AIDS Cohort Study (MACS, Kaslow 1987), a prospective study of disease progression in participants infected with, or at risk for infection with, HIV. The subset of MACS participants who seroconvert with HIV while under observation are followed from the date of HIV seroconversion, with many variables including CD4 cell counts and viral load measured at planned study visits every 6 months. Interest lies in the progression of CD4 cell counts and viral load from seroconversion with HIV, and their impact on survival. Of the available viral load data, 27% is missing and 17% falls below a limit of detection. Using a Bayesian analysis, we model the progression of CD4 cell counts over time, while accounting for the missingness and left-censoring on the available viral load data. We assume that the intermittent missingness is missing at random (MAR, Little 1995).

Although a great deal of attention has been paid to developing joint models in recent years, the literature on censored and/or missing covariate data within a joint model is sparse. A paper by Wu et al. (2008) investigated the joint modeling in an AIDS clinical trial with informative dropout. This paper incorporated a missing data mechanism into the joint model likelihood, to model missingness in the model covariates due to subject dropout. Estimation was performed using an EM algorithm. To the authors knowledge, no paper has investigated joint modeling with intermittent missing covariate data, or with covariate data subject to a limit of detection. Using data from the Multicenter AIDS Cohort Study, the goal of the analysis presented in this chapter is to jointly model the longitudinal progression of disease in study participants while accounting for both intermittent missingness and a limit of detection on a single covariate. Estimation will be undertaken using a Bayesian framework, which contrasts with the EM approach taken in the Wu et al. (2008) paper.

The rest of this chapter is organized as follows. In Section 3.2 we give a review of joint models, and develop notation. In Section 3.3 we develop a Bayesian approach to this problem, and apply this approach to the MACS data. We compare results obtained with those from ad-hoc estimation approaches. We conclude the chapter in Section 3.4 with a discussion.

3.2 Preliminaries

3.2.1 The Longitudinal Model

Of the two submodels included in a joint model, the longitudinal component is the lessor complicated with a model formulation very similar (if not identical) to that of a model fit for the longitudinal data alone. The dataset consists of data for $i = 1, ..., N$ subjects, with $j = 1, ..., n_i$ measurements recorded for subject i. The response y_{ij}, main-effect covariate vector $\mathbf{x}_{ij} = (x_{1ij}, ..., x_{pij})'$, and random-effect covariate vector $\mathbf{z}_{ij} = (z_{1ij}, ..., z_{qij})'$ are recorded at times t_{ij}. The longitudinal model is usually specified as a linear mixed effects model (Laird and Ware, 1982):

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b_i} + \epsilon_{ij} = \psi_{ij}(\boldsymbol{\beta}, \mathbf{b}_i) + \epsilon_{ij} \tag{3.1}$$

where $\boldsymbol{\beta}$ is the $p \times 1$ main-effect parameter vector, and $\mathbf{b}_i$ is the $q \times 1$ vector of random effects for subject i, with $\mathbf{b}_i$ specified as having a multivariate normal distribution, $\mathbf{b}_i \sim N_q(\boldsymbol{\mu_b}, \boldsymbol{\Sigma}_b)$. We emphasize here that the random effects $\mathbf{b}_i$ have mean μ_b, unlike the usual linear mixed effects model where the random effects have mean zero. This specification is important in longitudinal models that do not contain a main-effect intercept, which is often used to avoid issues of nonidentifiability with the baseline hazard function in the survival component of the joint model. The error vector $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{in_i})'$ is usually specified to have a multivariate normal distribution, $\epsilon_i \sim N_{n_i}(0, \xi^{-1}\mathbf{I}_{n_i})$, where $\mathbf{I}_{n_i}$ represents the identity matrix of dimension n_i. The trajectory function for the model is defined as $\psi_{ij}(\boldsymbol{\beta}, \mathbf{b}_i) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b_i}$. More generally, (3.1) can be written in terms of $y_i(t)$, the response at any time t. Taking $\mathbf{x}_i(t) = (x_{i1}(t), ..., x_{ip}(t))'$ and $\mathbf{z}_i(t) = (z_{i1}(t), ..., z_{iq}(t))$ to represent the main-effect and random-effect covariate vectors at time t respectively, the model can be rewritten as:

$$y_i(t) = \mathbf{x}'_i(t)\boldsymbol{\beta} + \mathbf{z}'_i(t)\mathbf{b}_i + \epsilon_i(t) = \psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t) + \epsilon_i(t) \tag{3.2}$$

where the error term $\epsilon_i(t) \sim N(0, \xi)$, and the trajectory function $\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t) = \mathbf{x}'_i(t)\boldsymbol{\beta} + \mathbf{z}'_i(t)\mathbf{b}_i$. In many AIDS studies using joint models, the longitudinal component uses random effects with functions of time only (Tsiatis and Davidian, 2004). The form of the random effect

covariate vector $\mathbf{z}_i(t)$ is usually simple, including only a random slope and time effect, or at most a quadratic term for time. In this case, the trajectory can be specified at generic time t, as follows:

$$\psi_i(\mathbf{b}_i, t) = \mathbf{z}_i'(t)\mathbf{b}_i \tag{3.3}$$

It should be noted that many authors have considered a more complex version of (3.3), involving an additional mean-zero stochastic process that does not depend on $\mathbf{z}_i(t)$ or $\mathbf{b}_i$. This form allows within-subject autocorrelation that accounts for fluctuations from the hypothesized "smooth" trajectory function included in the model. This extended form is not considered in the proposed modeling approach presented in Section 3.3.

3.2.2 The Survival Model

The second submodel in a joint model is the survival model. This is usually taken as a Cox proportional hazards model (Cox 1972), which predicts the hazard function $\lambda_i(t)$ for subject i at time t. The survival component of the joint model includes a link to the longitudinal submodel, the unique characteristic that makes the model "joint". The link in this case is the inclusion of a portion (or all) of the longitudinal trajectory $\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t)$ as a covariate within the survival model. The survival component is expressed as:

$$\lambda_i(t) = \lambda_0(t) \exp\left\{h(\boldsymbol{\beta}, \mathbf{b}_i)\theta + \mathbf{x}_{si}'(t)\boldsymbol{\beta}_s\right\} \tag{3.4}$$

Here $h(\boldsymbol{\beta}, \mathbf{b}_i)$ is a function of the main effects and random effects in the longitudinal model, with θ as the scalar parameter that links the two submodels. The survival covariate vector $\mathbf{x}_{si}(t) = (x_{si1}, ..., x_{sir})'$ usually includes baseline covariates for subject i, with $\boldsymbol{\beta}_s$ representing the $r \times 1$ parameter vector for these baseline covariates. The baseline hazard function is given as $\lambda_0(t)$. The form that $h(\boldsymbol{\beta}, \mathbf{b}_i)$ takes determines the type of joint model that is fit. In a **selection** model, we have $h(\boldsymbol{\beta}, \mathbf{b}_i) = \psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t) = \mathbf{x}_i'(t)\boldsymbol{\beta} + \mathbf{z}_i'(t)\mathbf{b}_i$, such that the full longitudinal trajectory is included in the survival component. In a **shared parameter** model, only individual parameters from the longitudinal model are included instead of the full trajectory. One example is to take $h(\boldsymbol{\beta}, \mathbf{b}_i) = \mathbf{z}_i'(t)\mathbf{b}_i$, such that only the random effects

are included in the survival component. The parameter $\boldsymbol{\beta}_s$ in (3.4) is a parameter vector for covariates unique to the survival submodel. These additional covariates $\mathbf{x_{si}}$ are usually baseline covariates. In most formulations, the baseline covariates are only included in the survival component of the model, as inclusion in the longitudinal component can lead to problems of identifiability when fitting the full joint model.

3.2.3 Likelihood for Joint Model

With both the longitudinal and survival submodels specified, we now combine the two to form the likelihood for the full joint model. In this case we will specify the joint model using a selection model in the survival component, such that the full longitudinal trajectory is included as a survival model covariate. We take T_i to represent the potential failure time for subject i, and C_i to represent the potential censoring time for subject i. We define $S_i = \min(T_i, C_i)$ as the observed failure/censoring time for subject i, with δ_i taken as an indicator for observing failure, with $\delta_i = 1$ when $T_i < C_i$, and $\delta_i = 0$ otherwise. We define $\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t)$ as the value of the longitudinal trajectory for subject i at time t (with $\psi_{ij}(\boldsymbol{\beta}, \mathbf{b}_i)$ as the longitudinal trajectory for subject i at visit j). With $f(\cdot)$ representing a generic density function, the likelihood for the ith subject in the joint model can be written as:

$$
\begin{aligned}
L_i \propto & f_i(\text{Survival} \mid \text{Longitudinal}) \times f_i(\text{Longitudinal}) \\
= & f(S_i|\theta, \delta_i, \boldsymbol{\beta}_s, \psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t), \mathbf{x}_{si}) \times f(\mathbf{y_i}|\mathbf{X_i}, \mathbf{Z_i}, \boldsymbol{\beta}, \mathbf{b}_i) f(\mathbf{b}_i) \\
= & \left[\left\{ \lambda_0(S_i) \exp\left(\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, S_i)\theta + \mathbf{x}'_{si}(S_i)\boldsymbol{\beta}_s\right) \right\}^{\delta_i} \right. \\
& \qquad \left. \times \exp\left\{ - \int_0^{S_i} \lambda_0(u) \exp(\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, u)\theta + \mathbf{x}'_{si}(u)\boldsymbol{\beta}_s) du \right\} \right] \\
& \times \left[\left\{ \frac{\xi}{(2\pi)} \right\}^{n_i/2} \exp\left\{ -\frac{\xi}{2} \sum_{j=1}^{n_i} (y_{ij} - \psi_{ij}(\boldsymbol{\beta}, \mathbf{b}_i))^2 \right\} P(\mathbf{b}_i) \right]
\end{aligned}
\tag{3.5}
$$

and the likelihood for all subjects is $L = \prod_{i=1}^{N} L_i$. In Section 3.3.2, the likelihood will be appended to account for a covariate subject to left truncation and missingness.

3.2.4 Fitting the Model

Estimation of a joint model may be performed in at least two ways. The first estimation approach is to use the EM algorithm. This approach has been used often in past analysis of AIDS data (DeGruttola and Tu 1994, Wulfsohn and Tsiatis 1997). The R package `JM` (Rizopoulos, 2010) was recently released and fits shared parameter models using the EM algorithm. A second approach to estimation uses a Bayesian framework, fitting the model with Markov Chain Monte Carlo (MCMC) methods. This approach is discussed in detail in Ibrahim, Chen, and Sinha (2001, chap.7), and has been used by many authors (Xu and Zeger 2001, Wang and Taylor 2001, Brown and Ibrahim 2003, etc.). Guo and Carlin (2004) provide WinBUGS software for fitting joint models using a Bayesian framework. The shared parameter joint model that is fit is based on the models proposed by Henderson et al. (2000), in which random effects are used in both the survival and longitudinal submodels. The submodels are linked using common random effects, not the full longitudinal trajectory as in our curent analysis. For the analysis presented in this chapter, a Bayesian framework is used based on Ibrahim, Chen, and Sinha (2001), with all computation performed using R software (R Development Core Team 2008).

3.3 MACS Data Analysis

3.3.1 Background

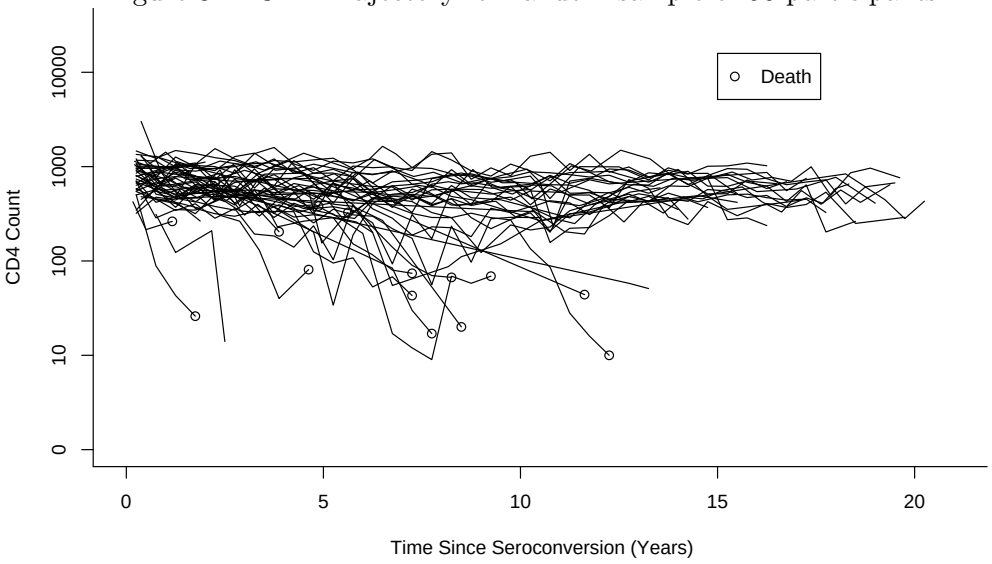
The motivating data for the analysis presented in this chapter comes from the Multicenter AIDS Cohort study (MACS), a prospective study of disease progression in participants infected with, or at risk for infection with, HIV. The data collected by the MACS study is of particular interest because participants are followed from the time of seroconversion, when they first develop antibodies to HIV (as a response to contracting the virus). The study population includes participants who contracted HIV during the study follow-up (1986-2005). Participants in the study were seen at semiannual visits, where demographic information was recorded along with laboratory measurements including viral load and CD4 cell counts. Survival data for each participant was also recorded, specifically for deaths attributable to AIDS.

Of the 470 subjects in the study who seroconverted with HIV during follow-up, 443 were observed at 3 or more visit times, and were included in the study analysis. This 3-visit minimum was required in order to have identifiable random effect parameters in the longitudinal component of the joint model. Of the 443 subjects, 165 (37.2%) died due to AIDS during the study period and had the time of death recorded.

In studies of HIV progression, interest lies in the relationship between CD4 cell count and viral load measurements over time. CD4 cell count is a measure of immune system strength, while viral load is a measure of the amount of circulating virus. These two biomarkers are inversely correlated, as high levels of virus (viral load) indicate low immune system strength (CD4 count). A complication that often arises in HIV studies is that viral load values are subject to a lower limit of detection. Values of viral load falling below this limit are unable to be detected by laboratory tests. In long-term longitudinal studies such as MACS, it is common for limit of detection to change over time, as newer technology is able to detect even lower levels of viral load. The available MACS data contains viral load values subject to two known limits of detection. Viral load data from earlier study periods is subject to a detection limit of 400 copies/mL, while viral load data from later study periods is subject to a detection limit of only 50 copies/mL. In total, 16.9% of the available viral load data fell below these detection limits. Additionally, 27.1% of the viral load data is missing intermittently in the dataset (MAR, Little 1995). A trajectory plot for CD4 since seroconversion for a random sample of 50 participants in the study is given in figure 3.1.

One additional limitation to the public-use MACS data was that time of visit for each participant was only available at the year level, no month or day dates are supplied. For a participant with multiple visits in the same year, the available data only lists the year of visit and the chronological ordering of multiple visits in the same year (i.e. that one visit precedes another). To account for this limitation, exact visit dates were imputed for each subject in the dataset. For a subject with two visits in year X, the time of the first visit was imputed at X + 0.25, with visit 2 at X + 0.75. For a subject with 3 visits time was imputed as X + 0.17, X + 0.5, and X + 0.83. The time of HIV seroconversion was imputed as the midpoint between times of the first visit where HIV antibodies were detected and the visit immediately

Figure 3.1: CD4 Trajectory for random sample of 50 participants



preceding this visit. For a particular subjects data to be included in the analysis, baseline covariates for race and age at seroconversion needed to be recorded. Additionally, at least one of CD4 cell count or viral load needed to be recorded.

3.3.2 Joint Model

To account for both the longitudinal trajectory of CD4 and survival, a joint model was specified for analysis. The longitudinal component of the model is specified as a mixed-effects model, with a random slope and intercept for each subject. In this model, we again have $i = 1, ..., N$ for each subject, and $j = 1, ..., n_i$ for each visit. Both CD4 and viral load were $\log_{10}$-transformed, with $CD4_{ij}$ and VL_{ij} representing the $\log_{10}$ transformed values of CD4 and viral load for subject i and visit j, occurring at time t_{ij}. Additionally, a covariate was included to account for the indirect effect of Highly Active Antiretroviral Treatment therapy (HAART), an HIV treatment that consists of several antiretroviral drugs being taken concurrently. HAART treatment has had a dramatic positive effect on the survival of HIV (Hammer et al. 1997, Cameron et al. 1998). Though records for HAART treatment are available in the MACS public-use data, we instead use HAART calendar period as an instrumental variable (Angrist, 1996) for HAART. This approach is similar to those in past HIV studies (Detels et al. 1998,

Tarwater et al. 2001), allowing us to circumvent potential bias in results due to residual confounding by indication that could occur if we used the direct HAART variable. We define HAART calendar period as all visit times occurring after January 1, 1998. We define covariate PD_{ij} as an indicator for the HAART calendar period, such that $PD_{ij} = 1$ if $t_{ij} > 1/1/98$, and $PD_{ij} = 0$ otherwise. The final longitudinal model is specified as follows:

$$CD4_{ij} = VL_{ij}\beta_1 + PD_{ij}\beta_2 + b_{0i} + t_{ij}b_{1i} + \epsilon_{ij} \tag{3.6}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ is the vector of parameters for the main effect covariates. Following standard estimation approaches for a linear mixed-effects model, the joint distribution of the random effects $\mathbf{b}_i = (b_{i0}, b_{i1})'$ was again assumed bivariate normal, with mean $\boldsymbol{\mu}_b$ and covariance matrix $\boldsymbol{\Sigma}_b$. The error term ϵ_{ij} is assumed to have a normal distribution, with $\epsilon_{ij} \sim N(0, 1/\xi)$. It should be noted that equation (3.6) does not include a main-effect intercept term. The intercept was excluded here to avoid issues of identifiability that arise when fitting both a main-effect intercept in the longitudinal component and the baseline hazard function in the survival component of a joint model.

A selection model was chosen for analysis, such that the full longitudinal trajectory was included as a covariate in the survival model. This trajectory is specified as $\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t) = VL_{it}\beta_1 + PD_{it}\beta_2 + b_{0i} + tb_{1i}$. Here VL_{it} and PD_{it} represent their respective values at the most recent observed visit to time t. We also denote $\psi_{ij}(\boldsymbol{\beta}, \mathbf{b}_i) = VL_{ij}\beta_1 + PD_{ij}\beta_2 + b_{0i} + t_{ij}b_{1i}$, the value of the longitudinal trajectory for subject i at visit j. Other baseline covariates of interest included the age at which a subject contracted HIV (AGE_i), and race ($RACE_i$), with $RACE_i = 1$ if subject i is white, and $RACE_i = 0$ otherwise. We again define θ as the parameter linking the longitudinal and survival submodels, with $\boldsymbol{\beta}_s = (\beta_{s1}, \beta_{s2})'$ as the parameters corresponding to the baseline covariates. The Cox proportional hazards submodel is specified below, with $\lambda(t)$ and $\lambda_0(t)$ representing the hazard and baseline hazard functions at time t, respectively.

$$\lambda(t|\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t)) = \lambda_0(t)\exp(\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t)\theta + RACE_i\beta_{s1} + AGE_i\beta_{s2}) \tag{3.7}$$

The key innovation in this analysis is that we are able to account for the missingness (27.1%) and left-censoring (16.9%) occurring in the viral load data. To do this, a normal prior distribution was specified for viral load, such that $VL_{ij} \sim N(\mu_v, 1/\eta_v)$. For a viral load observation VL_{ij} falling below a limit of detection LD_{ij}, the prior distribution is truncated at LD_{ij}, taking nonzero density only below LD_{ij}. For a viral load observations that are missing, no such truncation is needed. The missing viral load values are assumed to be missing at random, as parameters involving viral load are distinct from others in the model. Because of this assumption, the complete-data likelihood that now accounts for missing and left-censored viral load will be appended from (3.5) to include the viral load prior distribution. We will again denote $\mathbf{S} = (S_1, ..., S_N)$ as the vector of observed failure/censoring times for each subject, with $\boldsymbol{\delta} = (\delta_1, ..., \delta_N)$ taken as the vector of indicators for observing failure (with $\delta_i = 1$ if observed failure and 0 otherwise). The complete-data likelihood for the joint model can be expressed as follows:

$$\begin{aligned} L &= f(\text{Survival}|\text{Longitudinal}) \times f(\text{Longitudinal}) \\ &= f(\mathbf{S}, \boldsymbol{\delta}|\theta, \boldsymbol{\beta_s}, \psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t)) \times f(\mathbf{CD4}, \mathbf{b}, \mathbf{VL}|\boldsymbol{\beta}) \\ &= f(\mathbf{S}, \boldsymbol{\delta}|\theta, \boldsymbol{\beta_s}, \psi_i(\boldsymbol{\beta}, \mathbf{b}_i, t)) \times [f(\mathbf{CD4}|\boldsymbol{\beta}, \mathbf{b}, \xi) f(\mathbf{b}|\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b) f(\mathbf{VL}|\mu_v, \eta_v)] \end{aligned} \tag{3.8}$$

Expanding this out into a full formula for the complete-data likelihood for subject i, we have:

$$
\begin{aligned}
L_i \propto & \left[\left\{ \lambda_0(S_i) \exp\big(\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, S_i)\theta + RACE_i\beta_{s1} + AGE_i\beta_{s2}\big) \right\}^{\delta_i} \right. \\
& \left. \times \exp\left\{ -\int_0^{S_i} \lambda_0(u) \exp\big(\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, u)\theta + RACE_i\beta_{s1} + AGE_i\beta_{s2}\big) du \right\} \right] \\
& \times \left[\xi^{n_i/2} \exp\left\{ -\frac{\xi}{2} \sum_{j=1}^{n_i} \big(CD4_{ij} - \psi_{ij}(\boldsymbol{\beta}, \mathbf{b}_i)\big)^2 \right\} \right. \\
& \quad \times \ |\boldsymbol{\Sigma_b^{-1}}|^{1/2} \exp\left\{ -\frac{1}{2} (\mathbf{b}_i - \boldsymbol{\mu_b})' \boldsymbol{\Sigma_b^{-1}} (\mathbf{b}_i - \boldsymbol{\mu_b}) \right\} \\
& \quad \left. \times \ \eta_v^{n_i/2} \exp\left\{ -\frac{\eta_v}{2} \sum_{j=1}^{n_i} (VL_{ij} - \mu_v)^2 \right\} \right]
\end{aligned} \tag{3.9}
$$

For the fully Bayesian estimation approach, noninformative and improper prior distributions were placed on each model parameter. Following the examples in Ibrahim, Chen, and Sinha (2001, Section 7.3), independent uniform improper priors were taken for $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_s$, with $\pi(\boldsymbol{\beta}) \propto \mathbf{1}$, $\pi(\boldsymbol{\beta}_s) \propto \mathbf{1}$. Additional priors are specified as follows: $\xi \sim Gamma(10^{-3}, 10^{-3})$, $\mu_v \sim N(0, 10^5)$, $\eta_v \sim Gamma(10^{-3}, 10^{-3})$, $\mu_b \sim N(0, 10^6)$, $\boldsymbol{\Sigma}_b^{-1} \sim \text{Wishart}(3, 10^6\mathbf{I})$. The baseline hazard function $\lambda_0(t)$ was specified as having the form of a piecewise constant hazard, taking the constant value λ_k for each of the $k = 1, ..., 10$ time intervals $(s_k, s_{k+1}]$ that span the range of the observed times t_{ij}. Computation of $\exp\left\{ -\int_0^{S_i} \lambda_0(u) \exp(...) du \right\}$ was then performed using the approximation given in Ibrahim, Chen, and Sinha (2001, p. 277-278). It can be shown that with this choice of priors, the joint posterior distribution is proper.

Gibbs sampling was performed by sampling from the full conditional distribution for each model parameter. A derivation of the conditionals is given in the Appendix. For parameters with a closed-form conditional distribution, sampling is straightforward. For parameters with no closed-form conditional distribution, sampling was performed using the Adaptive Rejection Metropolis Sampling of Gilks, Best, and Tan (1995). Estimation was performed using Gibbs samples from 10,000 iterations, with a burn-in of 1000 iterations. For comparison, several simpler models were also applied to the MACS data. First, a two-stage model was fit, in which each of the two submodels was fit separately. In the first stage, the longitudinal submodel in

equation (3.6) was fit independently of the survival component. The fitted trajectory from the longitudinal component was then fixed, and was included as a covariate in the survival model in equation (3.7). The second stage fitted this survival model, giving parameter estimates for the survival component only. Such a model is computationally simpler because the likelihood functions for each model are separate, and are not combined as in equation (3.9). Additionally, a joint model was also fit to only the 56% of total observations with fully observed values of viral load (complete-case analysis). A joint model was also fit in which substituted values of viral load were used for all left-censored viral load values. For a viral load measurement falling below limit of detection LOD, the common substitution of $LOD/\sqrt{2}$ was used as the "true" viral load value at the specified visit. This substitution analysis was then performed on 72.9% (56% observed + 16.9% substituted) of the total observations. In addition to the full joint model given by (3.6) and (3.7), a joint model incorporating a quadratic random time effect in the longitudinal component was also fit to the data. To achieve parameter convergence for this model, we restricted the dataset to include only subjects with observed data at 5 visit times (compared to the minimum of 3 visits for all other models). In addition, the effects of time and age at seroconversion were centered for this model only. The simulation results from the two-stage, complete-case, substitution, and full joint models are given Table 3.1. Posterior estimates are taken from the 9000 sampled values. Figure 3.2 provides trace plots and probability density histograms (with overlaid kernal smoothed density functions) for parameters of interest from the full joint model.

3.3.3 Results

The results in Table 3.1 show that decreasing CD4 cell count is associated with an increased risk of death, as expected. Specifically, the full joint model predicts that each 10% decrease in CD4 cell count results in a 15.9% increase in the risk of death. This estimate ranges from 13.6% in the two-stage model to 20.5% in the substitution model. Additionally, CD4 cell count and viral load levels are shown to be inversely related, with each 10% increase in viral load resulting in a predicted 0.48% decrease of CD4. This estimated decrease ranges from

Table 3.1: Parameter estimates for MACS data analysis in all models

	Independent Model[1]				Complete-Case Joint[2]				Full Joint[3]			
Parameter	Mean	SD	Lower*	Upper*	Mean	SD	Lower	Upper	Mean	SD	Lower	Upper
Longitudinal Component												
$\beta_1(VL)$	-0.0446	0.0029	-0.0493	-0.0400	-0.0778	0.0051	-0.0871	-0.0696	-0.0501	0.0029	-0.0548	-0.0452
$\beta_2(PD)$	0.1725	0.0116	0.1531	0.1915	0.1447	0.0142	0.1215	0.1679	0.1589	0.0109	0.1411	0.1771
μ_{b_0}	3.0707	0.0139	3.0483	3.0940	3.2299	0.0232	3.1940	3.2709	3.0920	0.0135	3.0694	3.1142
μ_{b_1}	-0.1052	0.0031	-0.1104	-0.1101	-0.1027	0.0026	-0.1071	-0.0985	-0.1018	0.0021	-0.1053	-0.0983
$\mathbf{\Sigma}_{b11}$	0.0915	0.0081	0.0790	0.1054	0.0368	0.0039	0.0308	0.0437	0.0501	0.0043	0.0434	0.0574
$\mathbf{\Sigma}_{b12}$	-0.0289	0.0035	-0.0350	-0.0234	-0.0083	0.0014	-0.0108	-0.0061	-0.0123	0.0016	-0.0150	-0.0098
$\mathbf{\Sigma}_{b22}$	0.0252	0.0026	0.0212	0.0296	0.0091	0.0010	0.0076	0.0108	0.0107	0.0010	0.0092	0.0123
μ_v	3.5006	0.0212	3.4665	3.5356	-	-	-	-	3.543	0.0215	3.5083	3.5785
η_v	0.4309	0.0086	0.4169	0.4453	-	-	-	-	0.4251	0.0087	0.4110	0.4395
ξ	26.529	0.4986	25.7104	27.3515	26.4948	0.6618	25.4061	27.5989	26.3931	0.4951	25.591	27.1984
Survival Component												
$\theta(CD4)$	-2.7774	0.1197	-2.9803	-2.5843	-3.9867	0.2463	-4.3930	-3.5784	-3.2338	0.1645	-3.5033	-2.9561
$\beta_{s1}(Race)$	0.1832	0.2242	-0.1778	0.5610	0.2771	0.2920	-0.1977	0.7683	0.2792	0.2437	-0.1168	0.6825
$\beta_{s2}(Age)$	0.0078	0.0092	-0.0071	0.0231	0.0156	0.0111	-0.0024	0.0337	0.0076	0.0102	-0.0096	0.0240

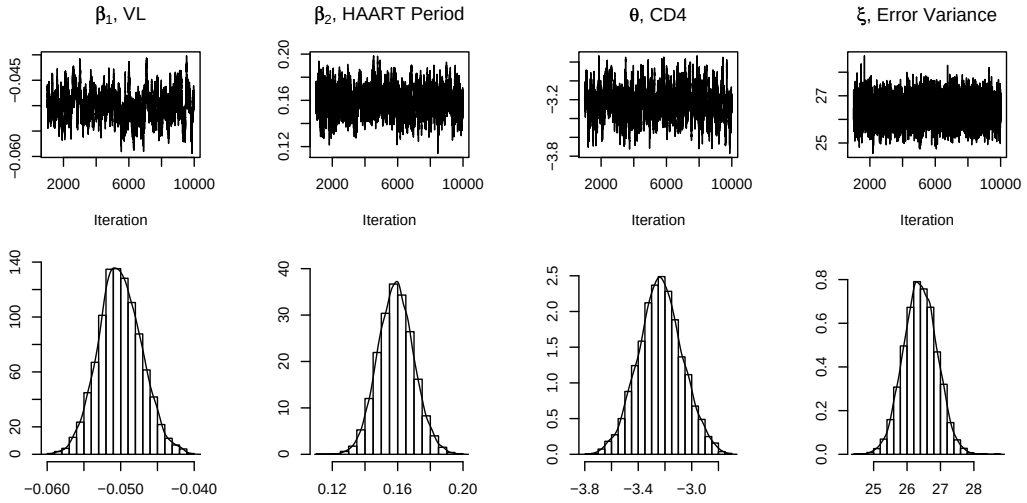
*Lower and upper 95% bounds from parameter distribution.

[1]Model fitting first longitudinal model, then using longitudinal trajectory as covariate in independent survival model.

[2]Joint model on 56% of observations with observed viral load.

[3]Joint model on 100% of observations, sampling both missing and left-censored viral load values.

Figure 3.2: Trace Plots and Sampled Densities of Selected Parameters from Full Joint Model



0.42% in the two-stage model to 0.74% in the complete-case model. Neither race nor the age at seroconversion were found to be significantly correlated with risk of death. The calendar period associated with HAART treatment was shown to result in an increase in CD4 cell count values in all models. For the full joint model, a participant during the HAART calendar period is expected to have a CD4 cell count value that is 44.2% higher than a participant in the pre-HAART period. Combining this estimate with θ (the survival model estimate for CD4 cell count) indicates that the HAART calendar period is associated with a 40% decrease in the risk of death for any particular participant. The results from the two-stage, substitution, and complete-case models show a predicted decrease of 38%, 41%, and 44%, respectively. The quadratic model results (not shown) indicated that a participant during the HAART calendar period expected to have a CD4 cell count that is 32.8% higher than a participant in the pre-HAART period, with HAART being associated with a 32% decrease in the risk of death. It is important to note that the quadratic model was fit on a subset of the data used for the other models, so results are not directly comparable.

3.4 Discussion

We have proposed a joint model for the analysis of longitudinal and survival data that accounts for both missingness and left-censoring in the longitudinal covariates. The proposed model allows use of a much greater proportion of available data when longitudinal covariates are missing or left-censored. In many infectious disease studies, measures of biomarkers are subject to a lower limit of detection, resulting in many left-censored cases. Previous analyses on only complete-case data then are not able to capture the information contained when subjects have very low levels of left-censored biomarkers. The proposed methodology accounts for this left-censoring, and also intermittent missingness that can be considered MAR.

The analysis of the MACS data presented in Table 3.1 shows that posterior estimates obtained from a joint model can be strongly influenced by the inclusions of observations with covariates that are missing or left-censored. In the available data, only 56% of viral load measurements were observed, with 27.1% missing and 16.9% falling below the limit of detection. Consequently, a complete-case analysis could only be performed on roughly half of the available data points. Including all cases in the proposed joint model is clearly more desirable, and as shown can produce results that vary from the complete-case results. However, in the worked example we did not see a difference in the estimated relative hazard for the calendar period associated with HAART. Yet, the precision was notably better for the proposed method compared to a complete-case analysis.

The computing time necessary to fit the proposed model can vary widely depending on the software that is used. The code for the simulations presented here was fit using R software (R Development Core Team 2008), which was able to run approximately 1500 iterations in 24 hours. This relatively long computing time could likely be lessened by use of alternative programming languages, such as C or WinBUGS.

While the proposed modeling approach can improve estimation with missing data in joint models, the assumptions still specify that the intermittently missing covariates are MAR. In many analyses, this assumption about the missing data mechanism may not be correct, as missing data can often arise from a more complicated mechanism. Future research is needed

to develop joint models for more complex missing data mechanisms.

Appendix

This appendix displays the full conditional distributions for the full joint model with likelihood given by (3.9). We use the same data notation given in Section 3.3.2, with slight modifications as detailed here. We define $\mathbf{CD4}_i = (CD4_{i1}, ..., CD4_{in_i})'$, $\mathbf{VL}_i = (VL_{i1}, ..., VL_{in_i})'$, $\mathbf{PD}_i = (PD_{i1}, ..., PD_{in_i})'$, and $\mathbf{t}_i = (t_{i1}, ..., t_{in_i})'$ to denote the covariate vectors of CD4, viral load, treatment period, and time for subject i. For ease of exposition, we will define $\mathbf{z}_i = (RACE_i, AGE_i)'$ and $\boldsymbol{\beta}_s = (\beta_{s1}, \beta_{s2})'$, such that $\mathbf{z}_i'\boldsymbol{\beta}_s = RACE_i\beta_{s1} + AGE_i\beta_{s2}$. The baseline hazard function λ_0 is specified as piecewise constant, taking the the value λ_k fore each of the $k = 1, ..., K$ time intervals, with $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_K)'$. For time interval k, we use d_k to denote the number of failures that occur within that interval. Computation of $\exp\left\{-\int_0^{S_i} \lambda_0(u)\exp(...)du\right\}$ was performed using the approximation given in Ibrahim, Chen, and Sinha (2001, p. 277-278). The notation for this approximation is as follows:

$$\exp\left[-\sum_{k=1}^{K}\lambda_k B_{ik}\right] \approx \exp\left[-\int_0^{S_i}\lambda_0\exp(\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, u)\theta + \mathbf{z}_i'\boldsymbol{\beta}_s)du\right]$$

To simplify notation when writing the full conditionals, we will take $\Omega = (\boldsymbol{\lambda}, \theta, \boldsymbol{\beta}_s, \xi, \boldsymbol{\Sigma}_b, \boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\mu}_b, \eta_v, \mu_v)$ to denote the set of all parameters in the model. We will use the notation $\Omega_{(-\boldsymbol{\beta})}$ to denote the set Ω without the parameter $\boldsymbol{\beta}$ (and similar notation when excluding other parameters). We will use the notation $\mathbf{D}_i$ to denote the set of complete data for subject i, such that $\mathbf{D}_i = (\mathbf{CD4}_i, \mathbf{VL}_i, \mathbf{PD}_i, \mathbf{t}_i, S_i, \delta_i, \mathbf{z}_i)$. We use the shorthand notation $\mathbf{D}_{i(-\mathbf{VL}_i)}$ to denote the set of complete data $\mathbf{D}_i$ not including $\mathbf{VL}_i$. The full set of complete data is denoted $\mathbf{D} = (\mathbf{D}_1, ..., \mathbf{D}_N)$ (with $\mathbf{D}_{(-\mathbf{VL}_i)}$ denoting the set of complete data excluding $\mathbf{VL}_i$).

1. $\boldsymbol{\beta} : \pi(\boldsymbol{\beta}) \sim \mathbf{1}$

$$
\begin{aligned}
P(\boldsymbol{\beta}|\Omega_{-\boldsymbol{\beta}},\mathbf{D}) \propto & \\
& \exp\left[-\frac{\xi}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}(CD4_{ij}-\psi_{ij}(\boldsymbol{\beta},\mathbf{b}_i))^2\right] \\
& \times \exp\left\{\sum_{i=1}^{N}\delta_i\left[(\psi_i(\boldsymbol{\beta},\mathbf{b}_i,S_i)\theta+\mathbf{z}_i'\boldsymbol{\beta}_s\right]\right\} \\
& \times \exp\left[-\sum_{i=1}^{N}\sum_{k=1}^{K}\lambda_k B_{ik}\right] \\
& = \text{No closed form}
\end{aligned}
$$

2. $\mathbf{b}_i : P(\mathbf{b}_i) \sim N_2(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$

$$
\begin{aligned}
P(\mathbf{b}_i|\Omega_{(-\mathbf{b}_i)},\mathbf{D}) \propto & \\
& \exp\left[-\frac{\xi}{2}\sum_{j=1}^{n_i}(CD4_{ij}-\psi_{ij}(\boldsymbol{\beta},\mathbf{b}_i))^2\right] \\
& \times \exp\left(-\frac{1}{2}(\mathbf{b}_i-\boldsymbol{\mu}_b)'\boldsymbol{\Sigma}_{\boldsymbol{b}}^{-\mathbf{1}}(\mathbf{b}_i-\boldsymbol{\mu}_b)\right) \\
& \times \exp\left\{\delta_i\left[(\psi_i(\boldsymbol{\beta},\mathbf{b}_i,S_i)\theta+\mathbf{z}_i'\boldsymbol{\beta}_s\right]\right\} \\
& \times \exp\left[-\sum_{k=1}^{K}\lambda_k B_{ik}\right] \\
& = \text{No closed form}
\end{aligned}
$$

3. $\xi : P(\xi) \sim \text{Gamma}(\text{Shape} = a_\xi, \text{Rate} = b_\xi)$

$$
\begin{aligned}
P(\xi|\Omega_{(-\xi)}, \mathbf{D}) \propto & \\
& \prod_{i=1}^{N} \xi^{n_i/2} \exp\left[-\frac{\xi}{2}\sum_{j=1}^{n_i}\left(CD4_{ij} - \psi_{ij}(\boldsymbol{\beta}, \mathbf{b}_i)\right)^2\right] \times \xi^{a_\xi - 1}\exp\left[-b_\xi \xi\right] \\
& \propto \xi^{\frac{1}{2}\sum_{i=1}^{n} n_i + a_\xi - 1} \exp\left(-\xi\left[\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(CD4_{ij} - \psi_{ij}(\boldsymbol{\beta}, \mathbf{b}_i)\right)^2 + b_\xi\right]\right) \\
& \sim \text{Gamma}\Bigg(\text{Shape} = \frac{1}{2}\sum_{i=1}^{n} n_i + a_\xi, \\
& \qquad\qquad \text{Rate} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(CD4_{ij} - \psi_{ij}(\boldsymbol{\beta}, \mathbf{b}_i)\right)^2 + b_\xi\Bigg)
\end{aligned}
$$

4. $\mu_v : P(\mu_v) \sim N(\mu_{\mu_v}, \eta_{\mu_v}^{-1})$

$$
\begin{aligned}
P(\mu_v|\Omega_{(-\mu_v)}, \mathbf{D}) \propto & \\
& \exp\left\{-\frac{\eta_v}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}(VL_{ij} - \mu_v)^2\right\} \times \exp\left\{-\frac{\eta_{\mu_v}}{2}(\mu_v - \mu_{\mu_v})^2\right\} \\
& \qquad \text{Note: } N_t = \sum_{i=1}^{N} n_i, \quad \bar{v} = \frac{1}{N_t}\sum_{i=1}^{N}\sum_{j=1}^{n_i} VL_{ij} \\
& \propto \exp\left\{-\frac{1}{2}\left(N_t\eta_v + \eta_{\mu_v}\right)\mu_v^2 + \left(N_t\eta_v\bar{v} + \eta_{\mu_v}\mu_{\mu_v}\right)\mu_v\right\} \\
& \propto \exp\left\{-\frac{A}{2}(\mu_v - C)^2\right\} \\
& \text{Where} \quad A = N_t\eta_v + \eta_{\mu_v}, \qquad C = \frac{N_t\eta_v\bar{v} + \eta_{\mu_v}\mu_{\mu_v}}{N_t\eta_v + \eta_{\mu_v}} \\
& \sim \text{Normal}\left(\frac{N_t\eta_v\bar{v} + \eta_{\mu_v}\mu_{\mu_v}}{N_t\eta_v + \eta_{\mu_v}}, \frac{1}{N_t\eta_v + \eta_{\mu_v}}\right)
\end{aligned}
$$

5. $\eta_v : P(\eta_v) \sim \text{Gamma}(\text{Shape} = a_\eta, \text{Rate} = b_\eta)$

$$
\begin{aligned}
P(\eta_v|\Omega_{(-\eta_v)}, \mathbf{D}) \propto & \\
& \prod_{i=1}^{N} \eta_v^{n_i/2} \exp\left[-\frac{\eta_v}{2}\sum_{j=1}^{n_i}(VL_{ij}-\mu_v)^2\right] \times \eta_v^{a_\eta - 1}\exp\left[-b_\eta\eta_v\right] \\
& \propto \eta_v^{\frac{1}{2}\sum_{i=1}^{n} n_i + a_\eta - 1} \exp\left(-\eta_v\left[\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}(VL_{ij}-\mu_v)^2 + b_\eta\right]\right) \\
& \sim \text{Gamma}\left(\text{Shape} = \frac{1}{2}\sum_{i=1}^{N} n_i + a_\eta, \text{Rate} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}(VL_{ij}-\mu_v)^2 + b_\eta\right)
\end{aligned}
$$

6. $\mathbf{\Sigma}_b^{-1} : P(\mathbf{\Sigma}_b^{-1}) \sim \text{Wishart}(n_0, c_0\mathbf{I})$, ($\mathbf{I}$ = Identity matrix)

$$
\begin{aligned}
P(\mathbf{\Sigma}_b^{-1}|\Omega_{(-\mathbf{\Sigma}_b)}, \mathbf{D}) \propto & \\
& |\mathbf{\Sigma_b^{-1}}|^{N/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{b_i}-\boldsymbol{\mu}_b)'\mathbf{\Sigma_b^{-1}}(\mathbf{b_i}-\boldsymbol{\mu}_b)\right) \\
& \quad \times |\mathbf{\Sigma_b^{-1}}|^{\frac{1}{2}(n_0-p-1)}\exp\left[-\frac{1}{2}\text{tr}\left((c_0\mathbf{I})^{-1}\mathbf{\Sigma_b^{-1}}\right)\right] \\
& \propto |\mathbf{\Sigma_b^{-1}}|^{\frac{1}{2}(N+n_0-p-1)}\exp\left(-\frac{1}{2}\left[\sum_{i=1}^{N}(\mathbf{b_i}-\boldsymbol{\mu}_b)'\mathbf{\Sigma_b^{-1}}(\mathbf{b_i}-\boldsymbol{\mu}_b)' + \text{tr}\left[(c_0\mathbf{I})^{-1}\mathbf{\Sigma_b^{-1}}\right]\right]\right) \\
& \propto |\mathbf{\Sigma_b^{-1}}|^{\frac{1}{2}(N+n_0-p-1)}\exp\left(-\frac{1}{2}\text{tr}\left[\left(\sum_{i=1}^{N}(\mathbf{b_i}-\boldsymbol{\mu}_b)(\mathbf{b_i}-\boldsymbol{\mu}_b)' + (c_0\mathbf{I})^{-1}\right)\mathbf{\Sigma_b^{-1}}\right]\right) \\
& \sim \text{Wishart}\left(N+n_0, \left(\sum_{i=1}^{N}(\mathbf{b}_i-\boldsymbol{\mu}_b)(\mathbf{b}_i-\boldsymbol{\mu}_b)' + (c_0\mathbf{I})^{-1}\right)^{-1}\right)
\end{aligned}
$$

7. $\theta, \boldsymbol{\beta}_s : \pi(\theta), \pi(\boldsymbol{\beta}_s) \sim \mathbf{1}$

$$
\begin{aligned}
P(\theta|\Omega_{-\theta}, \mathbf{D}), P(\boldsymbol{\beta}_s|\Omega_{-\boldsymbol{\beta}_s}, \mathbf{D}) \propto & \\
& \exp\left\{\sum_{i=1}^{N} \delta_i \left[(\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, S_i)\theta + \mathbf{z}_i'\boldsymbol{\beta}_s\right]\right\} \\
& \times \exp\left[-\sum_{i=1}^{N}\sum_{k=1}^{K} \lambda_k B_{ik}\right] \\
& = \text{No closed form}
\end{aligned}
$$

8. $VL_{ij} : P(VL_{ij}) \sim N(\mu_v, \eta_v) I(0 \leq VL_{ij} \leq c_{ij})$, where $c_{ij} = \infty$ when VL_{ij} is missing, and $c_{ij} = L_{ij}$ when VL_{ij} is left-censored at limit of detection L_{ij}. I() denotes the indicator function.

$$
\begin{aligned}
P(VL_{ij}|\Omega, \mathbf{D}_{-VL_{ij}}, 0 \leq VL_{ij} \leq c_u) \propto & \\
& \exp\left\{-\frac{\xi}{2}(CD4_{ij} - \psi_{ij}(\boldsymbol{\beta}, \mathbf{b}_i))^2\right\} \\
& \times \exp\left\{-\frac{\eta_v}{2}(VL_{ij} - \mu_v)^2\right\} \\
& \times \exp\left\{\delta_i \left[(\psi_i(\boldsymbol{\beta}, \mathbf{b}_i, S_i)\theta + \mathbf{z}_i'\boldsymbol{\beta}_s\right]\right\} \\
& \times \exp\left[-\sum_{k=1}^{K} \lambda_k B_{ik}\right] \\
& \times I(0 \leq VL_{ij} < c_{ij}) \\
& = \text{No closed form}
\end{aligned}
$$

9. $\lambda_k : P(\lambda_k) \sim \text{Gamma}(\text{Shape} = a_\lambda, \text{Rate} = b_\lambda)$

$$
\begin{aligned}
P(\lambda_k|\Omega_{(-\lambda_k)}, \mathbf{D}) \propto & \\
& \lambda_k^{d_k} \times \exp\left[-\sum_{i=1}^{N} \lambda_k B_{ik}\right] \\
& \quad \times \lambda_k^{a_\lambda - 1} \exp\left[-b_\lambda \lambda_k\right] \\
& \sim \text{Gamma}\left(\text{Shape} = d_k + a_\lambda, \text{Rate} = \sum_{i=1}^{N} B_{ik} + b_\lambda\right)
\end{aligned}
$$

10. $\boldsymbol{\mu}_b : P(\boldsymbol{\mu}_b) \sim N_2(\mathbf{0}, \Sigma_{\boldsymbol{\mu}_b})$

$$
\begin{aligned}
P(\boldsymbol{\mu}_b|\Omega_{(-\boldsymbol{\mu}_b)}, \mathbf{D}) &\propto \frac{1}{|\Sigma_{\boldsymbol{\mu}_b}^{-1}|^{1/2}} \exp\left[\boldsymbol{\mu}_b'(\Sigma_{\boldsymbol{\mu}_b})^{-1}(\boldsymbol{\mu}_b)\right] \\
&\quad \times \exp\left[-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{b}_i - \boldsymbol{\mu}_b)'\boldsymbol{\Sigma}_b^{-1}(\mathbf{b}_i - \boldsymbol{\mu}_b)\right] \\
&\propto \exp\left[-\frac{1}{2}\boldsymbol{\mu}_b'(\Sigma_{\boldsymbol{\mu}_b})^{-1}\boldsymbol{\mu}_b - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{b}_i - \boldsymbol{\mu}_b)'\boldsymbol{\Sigma}_b^{-1}(\mathbf{b}_i - \boldsymbol{\mu}_b)\right] \\
&\quad \text{Note} : \bar{\mathbf{b}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{b}_i \\
&\sim N_2\left(\left(\Sigma_{\mu_b} + N\Sigma_b^{-1}\right)^{-1}\left(N\Sigma_b^{-1}\bar{\mathbf{b}}\right), \left(\Sigma_{\mu_b} + N\Sigma_b^{-1}\right)^{-1}\right)
\end{aligned}
$$

Bibliography

Akaike, H. (1974) "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19:6, 716-723.

Agarwal, A., Sankaran, S., Vajpayee, M., Sreenivas, V., Seth, P., and Dandekar, S. (2007) "Correlation of Immune Activation With HIV-1 RNA Levels Assayed by Real-Time RT-PCR in HIV-1 Subtype C Infected Patients in Northern India," *Journal of Clinical Virology*, 40:4, 301-306.

Anderson, DJ. (1989) "Determination of the Lower Limit of Detection [letter]," *Clinical Chemistry*, 35, 2152-3.

Angrist, J.D., Imbens, G.W., and Rubin, D.B. (1996). "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-455.

Armbruster, D.A., and Pry, T. (2008) "Limit of Blank, Limit of Detection and Limit of Quantitation," *The Clinical Biochemist Reviews*, 29, S49-S52.

Armbruster, D.A., Tillman, M.D., and Hubbs, L.M. (1994) "Limit of Detection (LOD)/Limit of Quantitation (LOQ): Comparison of the Empirical and the Statistical Methods Exemplified with GC-MS Assays of Abused Drugs," *Clinical Chemistry*, 40:7, 1233-1238.

Bai, J. (1997) "Estimation of a Change Point in Multiple Regression Models," *Review of Economics and Statistics*, 79:4, 551-563.

Brown, E.R., and Ibrahim, J.G. (2003a). "A Bayesian Semiparametric Joint Hierarchical Model for Longitudinal and Survival Data," *Biometrics*, 59, 221-228.

Browne, R.W., and Whitcomb, B.W. (2010) "Procedures for Determination of Detection Limits," *Epidemiology*, 21:4, S4-S9.

Cameron, D.W., Heath-Chiozzi, M., Danner, S., Cohen, C., Kravcik, S., Maurath, C., Sun, E., Henry, D., Rode, R, Potthoff, A., and Leonard, J. (1998). "Randomised Placebo-Controlled Trial of Ritonavir in Advanced HIV-1 Disease," *Lancet*, 351:9102, 543-549.

Centers for Disease Control and Prevention (CDC). National Health and Nutrition Examination Survey Data. 2005-2006.

Clinical and Laboratory Standards Institute (2004) "Protocols for Determination of Limits of Detection and Limits of Quantitation, Approved Guideline. CLSI document EP17," Wayne, PA USA: CLSI.

Cole, S.R., Chu, H., Nie, L. Schisterman, E.F. (2009) "Estimating the Odds Ratio When Exposure Has a Limit of Detection," *International Journal of Epidemiology*, 38, 1674-1680.

Cox, C. (2005) "Limits of Quantitation for Laboratory Assays," *Applied Statistics*, 54, 63-76.

Cox, D.R. (1972). "Regression Models and Life Tables (with Discussion)," *Journal of the Royal Statistical Society Series B*, 34, 187-200.

Cudeck, R., and Klebe, K. (2002) "Multiphase Mixed-Effects Models for Repeated Measures Data," *Psychological Methods*, 7:1, 41-63.

D'Angelo, G., and Weissfeld, L. (2008) "An Index Approach for the Cox Model with Left Censored Covariates," *Statistics in Medicine*, 27, 4502-4514.

DeGruttola, V., and Tu, X.M. (1994). "Modeling Progression of CD-4 Lymphocyte Count and its Relationship to Survival Time," *Biometrics*, 50, 1003-1014.

Detels, R., Munoz, A., McFarlane, G., Kingsley, L.A., Margolick, J.B., Giorgi, J., Schrager, L.K., and Phair, J.P. (1998). "Effectiveness of Potent Antiretroviral Therapy on Time to AIDS and Death in Men With Known HIV Infection Duration," *Journal of the American Medical Association*, 280:17, 1497-1503.

Dunne, A. (1995) "Decision and Detection Limits for Linear Homoscedastic Assays," *Statistics in Medicine*, 14, 1949-1959.

Gibbons, R.D., Grams, N.E., Jarke, F.H., and Stoub, K.P. (1992) "Practical Quantitation Limits," *Chemometrics and Intelligent Laboratory Systems*, 12, 225-235.

Gilks, W.R., Best, N.G. and Tan, K.K.C. (1995) "Adaptive Rejection Metropolis Sampling," *Applied Statistics*, 44, 455-472.

Gilks, W.R. and Wild, P. (1992) "Adaptive Rejection Sampling for Gibbs Sampling," Applied Statistics, 41, 337-348.

Guo, X., and Carlin, B.P. (2004). "Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages," *The American Statistician*, 58:1, 16-24.

Guo, Y., Harel, O., and Little, R.J. (2010) "How Well Quantified Is the Limit of Quantification?," *Epidemiology*, 21:4, S10-S16.

Hammer, S.M., Squires, K.E., Hughes, M.D., Grimes, J.M., Demeter, L.M., Currier, J.S., Eron, J.J., Feinberg, J.E., Balfour, H.H., Dayton, L.R., Chodakewitz, J.A., and Fischl, M.A. (1997). "A Controlled Trial of Two Nucleoside Analogues Plus Indinavir in Persons with Human Immunodeficiency Virus Infection and CD4 Cell Counts of 200 per Cubic Millimeter or Less," *New England Journal of Medicine*, 337:11, 725-733.

Hawkins, D.M. (2001) "Fitting Multiple Change-Point Models to Data," *Computational*

Statistics & Data Analysis, 37, 323-341.

Helsel, D.R. (2005) *Nondetects and Data Analysis: Statistics for Censored Environmental Data.* New York: Wiley.

Helsel, D.R. (2006) "Fabricating Data: How Substituting Values for Nondetects Can Ruin Results, and What Can Be Done About It." *Chemosphere* 65, 2434-2439.

Henderson, R., Diggle, P., and Dobson, A. (2000). "Joint Modelling of Longitudinal Measurements and Event Time Data," *Biostatistics*, 1:4, 465-480.

Hubaux, A., and Vos, G. (1967) "Decision and Detection Limits for Linear Calibration Curves," *Analytical Chemistry*, 42:8, 849-855.

Ibrahim, J. G. (1990), "Incomplete Data in Generalized Linear Models," *Journal of the American Statistical Association*, 85, 765-769.

Ibrahim, J. G., Lipsitz, S.R., and Chen, M. (1999) "Missing Covariates in Generalized Linear Models When the Missing Data Mechanism is Non-Ignorable," *Journal of the Royal Statistical Society, Series B*, 61, 173-190.

Ibrahim, J.G., Chen, M., and Sinha, D. (2001). "Bayesian Survival Analysis," New York: Springer.

Kaslow, R.A., Ostrow, D.G., Detels, R., Phair, J.P., Polk, B.F., and Rinaldo, C.R. (1987). "The Multicenter AIDS Cohort Study: Rationale, Organization, and Selected Characteristics of the Participants," *American Journal of Epidemiology*, 126:2, 310-318.

Laird, N.M., and Ware, J.H. (1982). "Random Effects Models for Longitudinal Data," *Biometrics*, 38, 963-974.

Linnet, K., and Kondratovich, M. (2004) "Partly Nonparametric Approach for Determining the Limit of Detection," *Clinical Chemistry*, 50:4, 732-740.

Lipsitz, S. R., and Ibrahim, J. G. (1996), "A Conditional Model For Incomplete Covariates in Parametric Regression Models," *Biometrika*, 72, 916-922.

Little, R.J.A. (1995). "Modeling the Drop-Out Mechanism in Repeated Measures Studies," *Journal of the American Statistical Association*, 90, 1112-1121.

Long, G.L., and Winefordner, J.D. (1983) "Limit of Detection: a Closer Look at the IUPAC Definition," *Analytical Chemistry*, 55, 712A-724A.

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

Lubin, J.H., Colt, J.S., Camann, D., Davis, S., Cerhan, J.R., Severson, R.K., Bernstein, L., and Hartge, P. (2004), "Epidemiologic Evaluation of Measurement Data in the Presence of Detection Limits," *Environmental Health Perspectives*, 112:17, 1691-1696.

Lynn, H. (2001) "Maximum Likelihood Inference For Left-Censored HIV RNA Data," *Statistics in Medicine*, 20:33-45.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) "Equations of State Calculations by Fast Computing Machines," *J. Chem. Phys.*, 21, 1087-1092.

Needleman, S.B., and Romberg, R.W. (1990) "Limits of Linearity and Detection for Some Drugs of Abuse," *Journal of Analytical Toxicology*, 14, 34-38.

Nolan, T., Hands, R.E., and Bustin, S.A. (2006) "Quantification of mRNA Using Real-Time RT-PCR," *Nature Protocols*, 1:3, 1559-1582.

Nie, L., Chu, H., Liu, C., Cole, S.R., Vexler, A., Schisterman, E.F. (2010) "Linear Regression With an Independent Variable Subject to a Detection Limit," *Epidemiology*, 21:4,S1-S8.

Nix, A.B.J., and Wilson, D.W. (1990) "Assay Detection Limits: Concept, Definition, and Estimation," *European Journal of Clinical Pharmacology*, 39, 203-206.

Oppenheimer, L., Capizzi, T.P., Weppelman, R.M., and Mehta, H. (1983) "Determining the Lowest Limit of Reliable Assay Measurement," *Analytical Chemistry*, 55, 638-643.

Palmer, S., Wiegand, A.P., Maldarelli, F., Bazmi, H., Mican, J.M., Polis, M., Dewar, R.L., Planta, A., Liu, S.Y., Metcalf, J.A., Mellors, J.W., and Coffin, J.M. (2003) "New Real-Time Reverse Transcriptase-Initiated PCR Assay with Single-Copy Sensitivity for Human Immunodeficiency Virus Type 1 RNA in Plasma," *Journal of Clinical Microbiology*, 41:10, 4531-4536.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Richardson, D.B., and Ciampi, A. (2003) "Effects of Exposure Measurement Error When an Exposure Variable Is Constrained by a Lower Limit," *American Journal of Epidemiology*, 157:4, 355-363.

Rigobon, R., and Stoker, T.M. (2007) "Estimation With Censored Regressors: Basic Issues," *International Economic Review*, 48:4, 1441-1467.

Rizopoulos, D. (2010). "JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data." *Journal of Statistical Software*, 35(9), 1-33.

Schisterman, E., Vexler, A., Whitcomb, B., and Liu, A. (2006) "The Limitations due to Exposure Detection Limits for Regression Models," *American Journal of Epidemiology*, 163:4, 374-383.

Singh, A., and Nocerino, J. (2002) "Robust Estimation of Mean and Variance Using Environmental Data Sets with Below Detection Limit Observations," *Chemometrics and Intelligent Laboratory Systems*, 60, 69-86.

Smyth, G.K. (1996) "Partitioned Algorithms for Maximum Likelihood and Other Non-Linear Estimation," *Statistics and Computing*, 6, 201-216.

Tarwater, P.M., Mellors, J., Gore, M.E., Margolick, J.B., Phair, J., Detels, R., Munoz, A. (2001). "Methods to Assess Population Effectiveness of Therapies in Human Immunodeficiency Virus Incident and Prevalent Cohorts," *American Journal of Epidemiology*, 154:7, 675-681.

Thompson, M., and Nelson, K. P. (2003) "Linear Regression With Type I Interval- and Left-Censored Response Data," *Environmental and Ecological Statistics*, 10, 221-230.

Thomsen, V., Schatzlein, D., and Mercuro, D. (2003) "Limits of Detection in Spectroscopy," *Spectroscopy*, 18:12, 112-114.

Tobin, J. (1958) "Estimation of Relationships for Limited Dependent Variables." *Econometrica* 26:24-36.

Tsiatis, A.A., and Davidian, M. (2004). "Joint Modeling of Longitudinal and Time-To-Event Data: An Overview,", *Statistica Sinica*, 14, 809-834.

Wang, Y, and Taylor, J.M.G. (2001). "Jointly Modeling Longitudinal and Event Time Data with Application to Acquired Immunodeficiency Syndrome," *Journal of the American Statistical Association*, 96, 895-905.

Wei, G.C., and Tanner, M. A. (1990), "A Monte Carlo Implementation of the EM algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American Statistical Association.* 85, 699-704.

Whitcomb, B.W., and Schisterman, E.F. (2008) "Assays with Lower Detection Limits: Implications for Epidemiological Investigations," *Paediatric and Perinatal Epidemiology*, 22, 597-602.

Wu, L., Hu, X.J., and Wu, H. (2008). "Joint Inference for Nonlinear Mixed-Effects Models and Time to Event at the Presence of Missing Data," *Biostatistics*, 9:2, 308-320.

Wulfsohn, M., and Tsiatis, A. (1997). "A Joint Model for Survival and Longitudinal Data Measured with Error," *Biometrics*, 53, 330-339.

Xu, J., and Zeger, S.L. (2001). "Joint Analysis of Longitudinal Data Comprising Repeated Measures and Times to Events," *Applied Statistics*, 50, 375-387.