12-7-2016

Predicting Ultimate Targets with Time-Dependent Predictors

Wei Fu
University of Connecticut - Storrs, fuwei001@hotmail.com

Predicting Ultimate Targets with Time-Dependent Predictors

Wei Fu, Ph.D.

University of Connecticut, 2016

ABSTRACT

The task of predicting the ultimate payment while a claim is open is critical to insurance claim reserving. It faces three main challenges: *(a)* right censoring in generalized linear models (GLM), *(b)* predicting an ultimate fixed target with time-dependent predictors, *(c)* correlated observations in the same subject. We present three methods in addressing these challenges: *(1)* a series of GLM and accelerated failure time (AFT) models, *(2)* a series of Bayesian models, *(3)* the cumulative longitudinal models (CLM) and generalized linear mixed models (GLMM). For each method, I explore the theoretical foundation, apply it on real claim data, and compare the model performance among alternatives. The advantages and limitations of each method are discussed.

My contributions include *(a)* proving the equivalence in MLE between GLM and AFT based on the gamma distribution and the log link, *(b)* empirically showing that a linear combination of insignificant predictors could be a significant predictor itself, *(c)* proposing a new way to generate the joint likelihood for nested observations in CLM. Finally, future areas of research are discussed.

Predicting Ultimate Targets with Time-Dependent Predictors

Wei Fu

M.S. Applied Statistics, Rochester Institute of Technology, 2006

M.B.A., Rochester Institute of Technology, 2000

B.Econ., Jiangxi University of Finance and Economics, 1996

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2016

APPROVAL PAGE

Doctor of Philosophy Dissertation

Predicting Ultimate Targets with Time-Dependent Predictors

Presented by

Wei Fu, B.Econ., M.B.A., M.S. App. Stat.

Major Advisor _____
Lynn Kuo

Associate Advisor _____
Naitee Ting

Associate Advisor _____
Manuel A. Nunez

University of Connecticut

2016

ACKNOWLEDGMENTS

I would like to sincerely thank Professor Lyn Kuo for her invaluable guidance, encouragement and patience throughout this incredible journey. I am deeply grateful for Professor Naitee Ting, who is as enlightening with his brillant thought as he is inspiring with his extensive experience. I greatly appreciate Professor Manuel Nunez for his detailed review of my premises and pointing out future directions.

My tremendous gratitude goes to my employer, MAPFRE Incurance, for allowing me to use their data for my research.

Last but not the least, I express my never-ending love and appreciation for my mother, Lü Shaoyu, father, Fu Zuojie, sister, Fu Jun, daughter, Fu Xinyu, and niece, Zhan Jiayi. Thank you for your unwavering support and to my parents, thank you for the gift of life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In insurance claim reserving, accurately predicting the severity, or ultimate total payment of a claim before it closes is an important task, with both over and under reserving having unfavorable implications.

At the first notice of loss (FNOL), or when a claim is first reported, the claim adjuster needs to estimate the ultimate total payout of benefits and set up the reserve. However, not all information needed for accurate estimation is necessarily available at FNOL. Some may become available later (e.g., attorney representation), while others may change over time (e.g., total number of hospital visits), making it necessary to adjust the prediction corresponding to changing information. For example, if three months after FNOL, a claim is still open and some of its characteristics have changed (e.g., more treatments, more payments have been made), a new prediction needs to be made consequently. Appendix A gives an illustration of the time-varying nature of the reserving process.

Although most claims close relatively fast, some may last for months or years,

especially for injury-related coverages such as Bodily Injury (BI), Personal Injury Protection (PIP), and Workers Compensation. For them, it is even more crucial to accurately update the reserve amount as the claim progresses.

Since there can be multiple claimants on the same claim, prediction in this research pertains to each claimant instead of the entire claim. Suppose y_{it} represents the cumulative payment for claimant i as of time t, and $\boldsymbol{x'_{it}} = (x_{kit})_{1\times k}$, $k = 0, 1, 2, \ldots, p_t$, the vector of covariates including the intercept for claim i, $i = 1, 2, \ldots, N$ at time $t, t = 1, 2, \ldots, c_i$ where c_i is the time when a claimant's case closes. So the objective in operation is to predict the ultimate payout y_{ic_i}, using information available at a given time t, including $\boldsymbol{x'_{it}}$ and y_{it} for $t, t-1, t-2, \ldots, 2, 1$.

Since, at time $t = c_i$ when claim i closes, y_{ic_i} will have been observed, which makes prediction unnecessary. So usually only predictions made at time $t < c_i$ are of interest.

Accordingly, a suitable predictive model needs to address the following challenges:

1. Right censoring: The target variable, total ultimate payment, is censored when the claim is open. Generalized Linear Model (GLM) [10, McCullagh, Nelder] is commonly used in modelling claim severity. Non-parametric methods have been developed to handle censoring in GLM [11, 1, 7] with limitations. See Wei [18] and Zhou & Li's critique [21] of the Buckley-James method. This author shows one can use an Accelerated Failure Time (AFT) model based on the gamma distribution and a log link to handle right censoring in GLM based on the gamma distribution with a log link.

2. Time-dependent predictors versus non-time-dependent target. That is, infor-

mation as of various time t is used to predict the dependent variable's value at a fixed yet unknown future time, rather than its value corresponding to each t. This is different from the typical longitudinal model [15, Skrondal, Rabe-Kesketh, p. 80], in which the target variable and the predictors correspond to the same time.

3. Nested observations: The observations, which are measurements of the same claims, are nested in groups. This has two implications. First, the group effects need to be accounted for, usually by random effects. Second, the observations in the same group are not conditionally independent given the group effects. So their joint-likelihood needs to be selected carefully. For example, in the mixed linear model, correlated observations are assumed to have the multivariate normal distribution, which will be an inappropriate assumption if it is assumed that the marginal distributions for individual observations are gamma.

Many methods have been considered including the AFT Model [18], the Proportional Hazard (PH) Model, the Generalized Linear Mixed Model (GLMM) [3], the frailty model, and certain latent variable models [15]. Many of them address some but not all of the challenges. For example, AFT does not handle time-dependent covariates. While PH handles that, it does not make inferences such as the mean of the underlying distribution. GLMM in its traditional form does not handle censoring, while the frailty model assumes conditional independence among nested observations.

My research focuses on likelihood-based methods. Method 1 was first developed in 2007 with a series of GLM and AFT models. Method 2 builds upon method

1 by utilizing modified historic priors in Bayesian update of parameter estimates. Next, I propose the Cumulative Longitudinal Model (CLM) in an attempt to address the challenges using a single model. It is an extension of GLMM with features from AFT and the first-order Markov chain.

1.1 Data

The structure of the modeling data depends on the selected modeling method. Table 1.1.1 on page 6 illustrates the data structure that is the basis for methods 1 and 2. It contains two claimants. Each observation represents a measurement of claimant i at time t. The claimants are assumed to be independent but the measurements are not.

T denotes how long a claim for the claimant has been open and is measured in months, assumed to be equal-length intervals in this example. If a claim is still open (i.e., censored), the close indicator Cls_Ind$= 0$, while Cls_Ind$= 1$ if the claim is closed. Here claimant 1's claim closes at month 6 while claimant 2's closes at month 10.

y_{11} and $y_{12}, \ldots$ show the cumulative payment for claimant 1 as of months 1 and 2 is $\$1000$ and $\$1500$ respectively, while the total payment at close is $y_{1_{n_1}} = \$6000$.

X_1 represents a non-time varying predictor, gender of the claimant. Its value may change among different claimants but not over time. $X_{11} = M$ means claimant 1 is male, while $X_{12} = F$ means claimant 2 is female.

X_2 represents a time-dependent covariate, total number of hospital visits. As we see here, the subscript for X_2 changes among claimants as well as over time t.

For claimant 1, there are 2 total hospital visits as of month 1, X_{211}. By month 2, it has increased to 4, X_{212}.

The goal is, at each month before the claim closes, to predict the ultimate payments, $\$6000$ for claimant 1 and $\$8000$ for claim 2, using all available information. For example, at month 2, that means values of cumulative payment, gender and total number of hospital visits as of month 2 and all previous months.

The data used in my research contains PIP claims that occurred between July 1, 2014 and August 8, 2016. Each observation is uniquely identified by the combination of a claim number and a claimant number. There are totally $26,285$ observations for $1,339$ distinct claimants. As of August 8, 2016, 208 or 15.53% claimants' claims were still open or censored while 1,131 or 84.47% were closed. Figure 1.1.1 on page 5 shows the vast majority of claims in this sample are closed within a year, while a small percentage of cases have lasted more than 2 years.

Figure 1.1.1: Estimated Survival Function of Claims

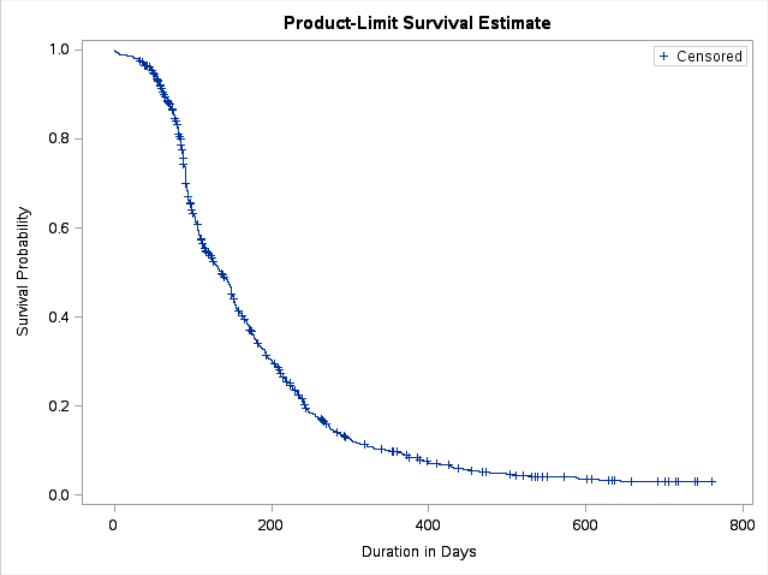


Table 1.1.1: Sample Claim Payment Data

	Target	Non-Time Varying Predictor	Time Varying Predictor	Time	Censoring Indicator
	Y	X_1	X_2	T	I
Claimant 1	y_{11}	x_{11}	x_{211}	1	0
	y_{12}	x_{11}	x_{212}	2	0
	y_{13}	x_{11}	x_{213}	3	0
	⋮	⋮	⋮	⋮	⋮
	y_{1n_1}	x_{11}	x_{21n_1}	n_1	1
Claimant 2	y_{21}	x_{12}	x_{221}	1	0
	y_{22}	x_{12}	x_{222}	2	0
	y_{23}	x_{12}	x_{223}	3	0
	⋮	⋮	⋮	⋮	⋮
	y_{2n_2}	x_{12}	x_{22n_2}	n_2	1
	⋮	⋮	⋮	⋮	⋮

	Cum. Payment	Gender	No. of Visits	Duration	Censoring Indicator
	Y	X_1	X_2	T	I
Claimant 1	0	M	1	0	0
	1000	M	2	1	0
	1500	M	4	2	0
	⋮	⋮	⋮	⋮	⋮
	6000	M	5	6	1
Claimant 2	100	F	3	0	0
	2000	F	6	1	0
	4500	F	7	3	0
	⋮	⋮	⋮	⋮	⋮
	8000	F	9	10	1
	⋮	⋮	⋮	⋮	⋮

Specifically, Table 1.1.2 on page 7 shows the Kaplan-Meier estimates of the survival probabilities. About 30% of cases are estimated to close within 3 months. This number quickly increases to $2/3$ by the end of 6 months. After 1 year, over 90% are estimated to close. These estimates are adjusted for censoring.

The ultimate total payment has a similarly skewed distribution (Figure 1.1.2(a),

Table 1.1.2: Survival Estimates By Selected Days

Duration Days	Est. Survival Probability	Est. Failure Probability	Number Left	Number Failed
91	0.70	0.30	872	384
182	0.3398	0.6602	390	814
366	0.0934	0.9066	90	1082

Note: Estimated using PROC LIFETEST.

page 8). The distributions between open or censored claimants and the closed ones are pretty similar in shape (Figure 1.1.2, page 8(b)).

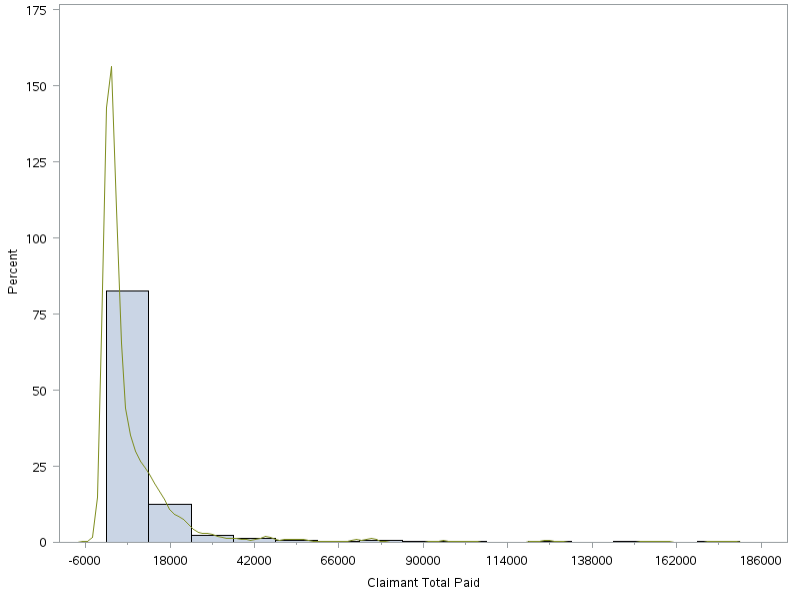
Figure 1.1.3 on page 9 depicts there is a positive correlation between duration and the log of the ultimate payment with a lot of variation. This makes sense because in general claimants with more severe injuries tend to take more time to recover, receive more treatments or more advanced treatments and end up incurring more expenses. However, there is a wide range of ultimate payments corresponding to each duration. So at the individual claim level, it is possible for a claim that has lasted for 200 days to cost more than one that has lasted 400 days.

For this study, non-time varying predictors are independent variables whose value does not change through out the claim's duration. It does not mean their values do not change forever. Here are some examples:

- Claimant characteristics: gender, occupation, age at the time of accident
- Policy characteristics: policy term (6 or 12 month policy), premium by coverage, payment plan, credit score, violations in the most recent term at the time of accident
- Accident characteristics: number of vehicles involved in the accident, vehicle

Figure 1.1.2: Distributions of Claimant Total Paid

(a) All Sample



(b) By Close Indicator

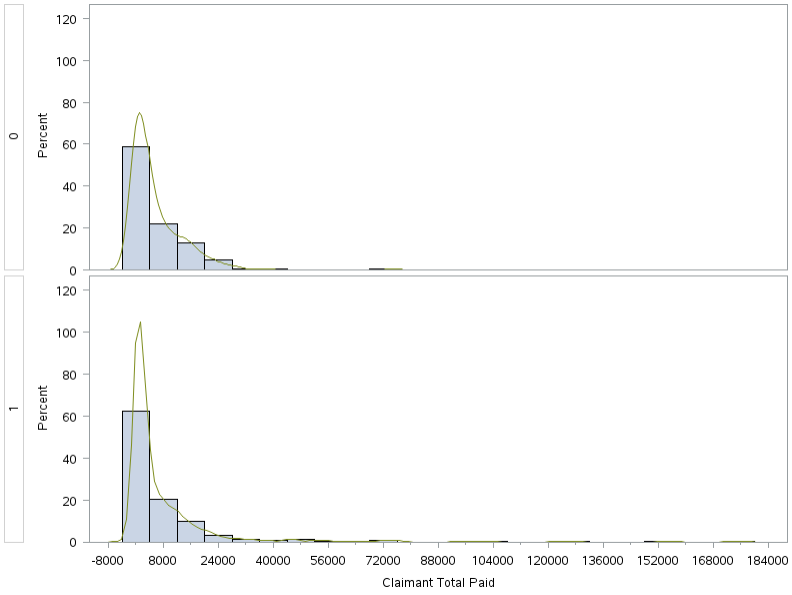
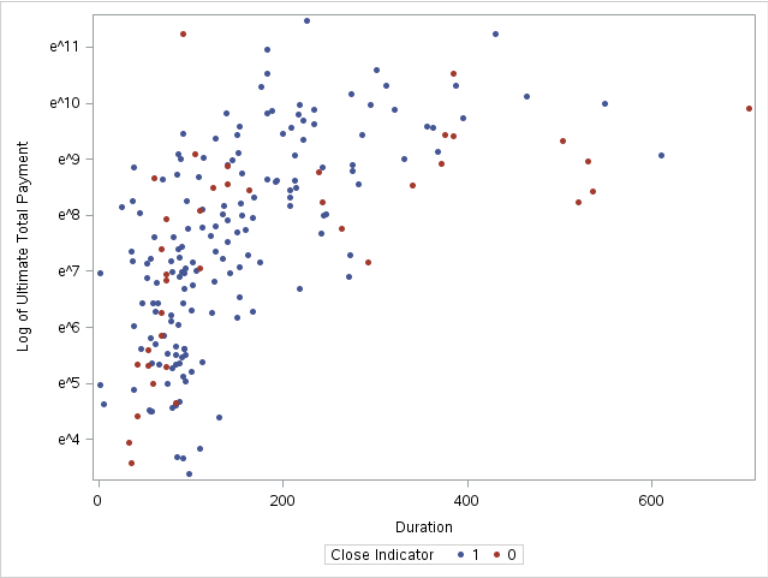


Figure 1.1.3: Scatter Plot of Log of Ultimate Total Payment versus Duration



make and model year, role of the claimant (e.g., driver, passenger, pedestrian, etc.)

Example of time-dependent variables include:

- Number of treatments received by type as of a particular time. For example, acupuncture, ambulance, anaesthesia, chiropractice, CT scan, electric stimulation, ER, manual therapy, radiology, ultra sound, and x-ray.
- Number of medical bills received
- Cumulative payments made as of a particular time.

There are over 200 variables initially. Many of them are dropped because of quality issues, for example, the missing percentage is too high, or privacy issues.

Chapter 2

Modelling Strategy

2.1 GLM With Right Censoring

The first challenge to be addressed is GLM with right censoring. Assuming we have a sample with independent observations from the same time, let $\mathbf{x}_i' = (1, x_{i1}, x_{i2}, \cdots, x_{ip})$, $i = 1, 2, \cdots n$ represent the independent variables, and y_i the dependent variable, the total payment for each claim.

In GLM, $g(E(y_i))$ is modelled by a linear predictor $\mathbf{x}_i'\boldsymbol{\beta}$, where $g(\cdot)$ is a link function, $E(\cdot)$ denotes the mean, the y_i is a random variable from the exponential family of distributions [10, McCullagh, Nelder].

Because the ultimate total payment has a skewed distribution (Figure 1.1.2), belongs to $\mathbb{R}^+$ and has a roughly linear relationship with some predictors on the log scale, the GLM based on the gamma distribution with a log link is a common

model for claim severity. Assuming $y_i \sim \Gamma(\alpha, \theta_i)$, y_i's p.d.f. is

$$\frac{1}{\Gamma(\alpha)\theta_i^{\alpha}} y_i^{\alpha-1} e^{-y_i/\theta_i} \tag{2.1.1}$$

In the context of GLM, it is convenient to re-parametrize equation (2.1.1) as follows:

$$\frac{1}{\Gamma(k)y_i} \left(\frac{y_i k}{\mu_i}\right)^k \exp\left(-\frac{y_i k}{\mu_i}\right) \tag{2.1.2}$$

where

- $k = \alpha$
- $\mu_i = k\theta_i$
- $\log\left(E\left(y_i\right)\right) = \mathbf{x}_i'\boldsymbol{\beta} \Rightarrow \mu_i = e^{\mathbf{x}_i'\boldsymbol{\beta}}$

2.1.1 Nonparametric Methods

Miller and Halpern [11] compared four non-parametric methods of handling censored data in regression. Among them, the Buckley-James method [1] has been used widely in regular linear regression and accelerated failure time (AFT) model. This method replaces a right-censored observation with its conditional expectation given the censoring, estimated by using the Kaplan-Meier estimator, which is essentially a version of the EM algorithm. Then model parameters can be estimated by solving the ordinary least square (OLS) normal equations iteratively. The Buckley-James method has shown good results in Monte Carlo simulations [7]. However, Miller and Halpern [11], L.J. Wei [18], and Zhou and Li [21] pointed out this method's lack of theoretical foundation.

Lai and Ying [7] introduced a modification to the Buckley-James estimator and showed under certain regularity conditions the modified estimator is consistent and asymptotically normal. If the distribution of the underlying error is normal, it is also asymptotically efficient. To address the difficulty in variance estimation of the Buckley-James estimator, Zhou and Li [21] proposed a testing and confidence interval procedure using the empirical likelihood method. Jin, Lin and Yin [6] developed a re-sampling procedure to estimate the limiting covariance matrix of this estimator in AFT.

Wang et al. [17] studied asymptotic properties for censored generalized linear model (GLM) with an unknown link function. Their procedure also utilized the Kaplan-Meier estimator.

Yu et al [19] proposed a quasi-likelihood based method to estimate parameters in GLM with right-censored data. Their method also estimates the conditional expectation of an censored observation using the Kaplan-Meier estimator just as the Buckley-James method does. The difference is that for parameter estimation, Yu's method solves the quasi score equation while the other solves the usual least square normal equations. Yu et al. [19] shows that the quasi likelihood method outperforms the Buckley-James method for both the extreme value and the normal error distributions.

All the methods mentioned above are nonparametric or semi-parametric methods, which are not limited by the distribution of the error or the dependent variable. Therefore they are versatile.

However, there are situations where parametric methods are desired, for example, in Bayesian analysis. Further, if the underlying distribution is known, a likelihood-based method may have more power and provide theoretic basis for in-

ference. The next section demonstrates in detail the parametric method of dealing with censoring in AFT, which uses the survival function for the likelihood of right-censored observations, can be applied to GLM based on the gamma distribution and the log link (see Section 2.1).

2.1.2 Survival Function Method

In AFT, $z_i = \mathbf{x}_i'\boldsymbol{\beta}^* + \sigma\epsilon_i$, where $z_i = \log y_i$ and ϵ_i is a random disturbance term.

Assuming $y_i \sim \Gamma(\alpha, \theta_i)$, the logarithm of the scale parameter θ_i is modelled by the linear predictor $\mathbf{x}_i'\boldsymbol{\beta}^*$. This leads to the following re-parametrization of y_i's pdf:

$$\begin{aligned} f(y_i) &= \frac{\exp\left(k\sigma\epsilon_i - e^{\sigma\epsilon_i}\right)}{y_i\Gamma(k)} \\ &= \frac{1}{y_i\Gamma(k)} \exp\left(k\sigma\frac{\log(y_i) - \mathbf{x}_i'\boldsymbol{\beta}^*}{\sigma} - \exp\left(\sigma\frac{\log(y_i) - \mathbf{x}_i'\boldsymbol{\beta}^*}{\sigma}\right)\right) \\ &= \frac{1}{\Gamma(k)e^{\mathbf{x}_i'\boldsymbol{\beta}^*\cdot k}} y_i^{k-1} \exp\left(-y_i/e^{\mathbf{x}_i'\boldsymbol{\beta}^*}\right) \end{aligned} \tag{2.1.3}$$

where

- $\epsilon_i = \frac{\log(y_i) - \mathbf{x}_i'\boldsymbol{\beta}^*}{\sigma}$
- $\theta_i = e^{\mathbf{x}_i'\boldsymbol{\beta}^*} \Rightarrow \log(\theta_i) = \mathbf{x}_i'\boldsymbol{\beta}^*$
- $k = \alpha$

Note this is still the PDF for y_i, not that for ϵ_i.

Model Parameter Comparison

In GLM, with parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$, we have $\log E(y_i) = \log(\alpha \cdot \theta_i) = \mathbf{x}_i'\boldsymbol{\beta}$. In AFT, with $\boldsymbol{\beta}^* = \left(\beta_0^*, \beta_1^*, \cdots, \beta_p^*\right)'$, we have $\log(\theta_i) = \mathbf{x}_i'\boldsymbol{\beta}^*$. So

$$\log\alpha + \mathbf{x}_i'\boldsymbol{\beta}^* = \mathbf{x}_i'\boldsymbol{\beta}$$

$$\mathbf{x}_i'(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \log\alpha$$

Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)'$ and $\mathbf{1} = (1, 1, \cdots, 1)'$, so

$$\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \log\alpha \cdot \mathbf{1}$$

$$\begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} (\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \log\alpha \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{2.1.4}$$

Assuming $\mathbf{X}$ has full column rank, we have following results.

Theorem 1. When $\alpha = 1$, $(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \mathbf{0}$ is the only solution to (2.1.4) .

Proof. When $\alpha = 1$, $\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \mathbf{0}$. Because $\mathbf{X}$ has full column rank, all its columns are LIN. Thus the only solution to (2.1.4) is $(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \mathbf{0}$ □

This means, when Exponential distribution is the underlying distribution, parameters in GLM and AFT are identical.

Theorem 2. $(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = (\log\alpha, 0, \cdots, 0)'$ is the only solution to (2.1.4).

Proof. suppose $(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = (a_0, a_1, \cdots, a_p)'$ is a solution to (2.1.4). So

$$\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} a_0 + \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} a_1 + \cdots + \begin{pmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{pmatrix} a_p = \log\alpha \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} (a_0 - \log\alpha) + \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} a_1 + \cdots + \begin{pmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{pmatrix} a_p = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{2.1.5}$$

Because $\mathbf{X}$ has full column rank, the only solution to (2.1.4) is

$$\begin{aligned} (a_0 - \log\alpha, a_1, \cdots, a_p) &= (0, 0, \cdots, 0) \\ (a_0, a_1, \cdots, a_p) &= (\log\alpha, 0, \cdots, 0) \\ (\boldsymbol{\beta} - \boldsymbol{\beta}^*) &= (\log\alpha, 0, \cdots, 0)' \end{aligned} \tag{2.1.6}$$

□

Therefore when the underlying gamma distribution is not Exponential distribution (i.e., $\alpha \neq 1$), the intercept under GLM differs from that under AFT by a constant (i.e., $\log\alpha$), while parameters for all other predictors are the same. This relationship is the foundation for adopting the survival function method for handling right censoring from AFT to GLM in this special case.

The next section shows how parameters are estimated in both GLM and AFT based on the survival function method.

Estimation for GLM

The log-likelihood function based on (2.1.2) for uncensored observations is

$$\begin{aligned} l(\mathbf{y}_{(0)}) &= \sum_{i:I_i=0} \left\{ -\log(y_i\Gamma(k)) - \frac{y_i k}{\mu_i} + k\log y_i k - k\log\mu_i \right\} \\ &= \sum_{i:I_i=0} \left\{ -\log(y_i\Gamma(k)) - \frac{y_i k}{e^{\mathbf{x}_i'\boldsymbol{\beta}}} + k\log y_i k - k\mathbf{x}_i'\boldsymbol{\beta} \right\} \end{aligned} \tag{2.1.7}$$

where I_i is the censoring indicator with 0 representing no censoring and 1 censoring, and $\mathbf{y}_{(0)} = \{y_i : I_i = 0\}$. For maximum-likelihood estimation, we first find

$$\begin{aligned} \frac{\partial}{\partial\beta_j} l(\mathbf{y}_{(0)}) &= \sum_{i:I_i=0} \frac{\partial}{\partial\beta_j} \left(-\frac{y_i k}{e^{\mathbf{x}_i'\boldsymbol{\beta}}} - k\mathbf{x}_i'\boldsymbol{\beta} \right) \\ &= \sum_{i:I_i=0} k x_{ij} \left(\frac{y_i}{e^{\mathbf{x}_i'\boldsymbol{\beta}}} - 1 \right) \end{aligned} \tag{2.1.8}$$

When there is no censoring, we solve $\frac{\partial}{\partial\beta_j} l(\mathbf{y}_{(0)}) = 0$. That is,

$$\sum_{i:I_i=0} x_{ij} \left(\frac{y_i}{e^{\mathbf{x}_i'\boldsymbol{\beta}}} - 1 \right) = 0 \tag{2.1.9}$$

which is free of k. This means when there is no censoring in GLM, $\boldsymbol{\beta}$ is estimated without involving the estimate for the common shape parameter.

For right-censored observations, the survival function replaces the p.d.f. in the

likelihood

$$\begin{aligned} P(Y_i > y_i) &= 1 - P(Y_i \leq y_i) \\ &= 1 - \frac{1}{\Gamma(k)} \int_0^{y_i} \frac{1}{y_i} \left(\frac{y_i k}{\mu_i}\right)^k \exp\left(-\frac{y_i k}{\mu_i}\right) dy_i \\ &= 1 - \frac{1}{\Gamma(k)} \int_0^{\frac{y_i k}{\mu_i}} W^{k-1} e^{-W} dW \\ &= 1 - \frac{1}{\Gamma(k)} \gamma\left(k, y_i k / e^{\mathbf{x}_i' \boldsymbol{\beta}}\right) \end{aligned} \tag{2.1.10}$$

where $\gamma(k, y_i k / e^{\mathbf{x}_i' \boldsymbol{\beta}^*})$ is the lower incomplete gamma function. The log-likelihood function for right-censored observations is

$$l(\mathbf{y}_{(1)}) = \sum_{i:I_i=1} \log\left[1 - \frac{1}{\Gamma(k)} \Gamma\left(k, \frac{y_i k}{e^{\mathbf{x}_i' \boldsymbol{\beta}}}\right)\right] \tag{2.1.11}$$

where $\mathbf{y}_{(1)} = \{y_i : I_i = 1\}$. Let $R = 1 - \frac{1}{\Gamma(k)} \Gamma\left(k, y_i k / e^{\mathbf{x}_i' \boldsymbol{\beta}}\right)$. So

$$\begin{aligned} \frac{\partial}{\partial \beta_j} l(\mathbf{y}_{(1)}) &= \frac{\partial}{\partial \beta_j} \sum_{i:I_i=1} \log(R) \\ &= \sum_{i:I_i=1} \frac{1}{R} \frac{\partial}{\partial \beta_j} R \\ &= \sum_{i:I_i=1} \frac{-1}{R\Gamma(k)} \frac{\partial}{\partial \beta_j} \int_0^{y_i k / e^{\mathbf{x}_i' \boldsymbol{\beta}}} W^{k-1} e^{-W} dW \\ &= \sum_{i:I_i=1} \frac{-1}{R\Gamma(k)} \left[\int_0^{y_i k / e^{\mathbf{x}_i' \boldsymbol{\beta}}} \frac{\partial}{\partial \beta_j} W^{k-1} e^{-W} dW + \left(\frac{y_i k}{e^{\mathbf{x}_i' \boldsymbol{\beta}}}\right)^{k-1} e^{-\frac{y_i k}{e^{\mathbf{x}_i' \boldsymbol{\beta}}}} \frac{\partial}{\partial \beta_j} \left(\frac{y_i k}{e^{\mathbf{x}_i' \boldsymbol{\beta}}}\right)\right] \\ &= \sum_{i:I_i=1} \frac{-1(-x_{ij})}{\Gamma(k) - \Gamma\left(k, y_i k / e^{\mathbf{x}_i' \boldsymbol{\beta}}\right)} \left(\frac{y_i k}{e^{\mathbf{x}_i' \boldsymbol{\beta}}}\right)^k e^{-\frac{y_i k}{e^{\mathbf{x}_i' \boldsymbol{\beta}}}} \\ &= \sum_{i:I_i=1} \frac{x_{ij}}{\Gamma\left(k, y_i k / e^{\mathbf{x}_i' \boldsymbol{\beta}}\right)} \left(\frac{y_i k}{e^{\mathbf{x}_i' \boldsymbol{\beta}}}\right)^k e^{-\frac{y_i k}{e^{\mathbf{x}_i' \boldsymbol{\beta}}}} \end{aligned} \tag{2.1.12}$$

where $\Gamma\left(k, y_i k / e^{\mathbf{x}_i' \boldsymbol{\beta}}\right))$ is the upper incomplete gamma distribution.

The log-likelihood of all observations is thus $l(\mathbf{y}) = l(\mathbf{y}_{(0)}) + l(\mathbf{y}_{(1)})$. The Newton-Raphson algorithm may be used to find parameter estimates that maximize $l(\mathbf{y})$. When there is no censoring or the common shape parameter k is known, the following equation is updated at each iteration $r+1$.

$$\boldsymbol{\beta}_{r+1} = \boldsymbol{\beta}_r - \mathbf{H}_1^{-1} \mathbf{S}_1 \tag{2.1.13}$$

where

- $\mathbf{H}_1 = \left[\frac{\partial^2}{\partial \beta_i \partial \beta_j} l(\mathbf{y})\right]_{ij}$
- $\mathbf{S}_1 = \left[\frac{\partial}{\partial \beta_j} l(\mathbf{y})\right]_j$
- $\log\left(E\left(y_i\right)\right) = \mathbf{x}_i' \boldsymbol{\beta} \Rightarrow \mu_i = e^{\mathbf{x}_i' \boldsymbol{\beta}}$

When there is right-censoring or the shape parameter is unknown, the following equation is updated.

$$\begin{bmatrix} \boldsymbol{\beta} \\ k \end{bmatrix}_{r+1} = \begin{bmatrix} \boldsymbol{\beta} \\ k \end{bmatrix}_r - \begin{bmatrix} \mathbf{H}_1 & \frac{\partial^2}{\partial \beta_j \partial k} l(\mathbf{y}) \\ \frac{\partial^2}{\partial k \partial \beta_j} l(\mathbf{y}) & \frac{\partial^2}{\partial k^2} l(\mathbf{y}) \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_1 \\ \frac{\partial}{\partial k} l(\mathbf{y}) \end{bmatrix} \tag{2.1.14}$$

Estimation for AFT

Based on equation (2.1.3) the log-likelihood function for uncensored observations is

$$l(\mathbf{y}_{(0)}) = -\sum_{i:I_i=0} \log(y_i \Gamma(k)) + k\sigma \sum_{i:I_i=0} \epsilon_i - \sum_{i:I_i=0} e^{\sigma \epsilon_i} \tag{2.1.15}$$

where I_i is the censoring indicator with 0 representing no censoring and 1 the opposite, and $\mathbf{y}_{(0)} = \{y_i : I_i = 0\}$. So

$$\begin{aligned}\frac{\partial}{\partial\beta_j} l(\mathbf{y}_{(0)}) &= \sum_{i:I_i=0} \left(k\sigma \frac{\partial \epsilon_i}{\partial \beta_j} - \frac{\partial}{\partial \beta_j} e^{\sigma\epsilon_i} \right) \\ &= \sum_{i:I_i=0} \left(k\sigma \frac{-x_{ij}}{\sigma} - \frac{-y_i x_{ij}}{e^{\mathbf{x}_i'\boldsymbol{\beta}^*}} \right) \\ &= \sum_{i:I_i=0} x_{ij} \left(\frac{y_i}{e^{\mathbf{x}_i'\boldsymbol{\beta}^*}} - k \right) \end{aligned} \tag{2.1.16}$$

Unless k is known, AFT always has to estimate both the shape parameter and $\boldsymbol{\beta}^*$ together. This is in contrast to GLM.

For right-censored observations, the likelihood is

$$\begin{aligned} P(Y_i > y_i) &= 1 - P(Y_i \leq y_i) \\ &= 1 - \frac{1}{\Gamma(\alpha)} \gamma\left(\alpha, y_i/\theta_i\right) \\ &= 1 - \frac{1}{\Gamma(k)} \gamma\left(k, y_i/e^{\mathbf{x}_i'\boldsymbol{\beta}^*}\right) \end{aligned} \tag{2.1.17}$$

where $\gamma(k, y_i/e^{\mathbf{x}_i'\boldsymbol{\beta}^*})$ is the lower incomplete gamma function. The log-likelihood function for right-censored observations is

$$l(\mathbf{y}_{(1)}) = \sum_{i:I_i=1} \log \left[1 - \frac{1}{\Gamma(k)} \Gamma\left(k, y_i/e^{\mathbf{x}_i'\boldsymbol{\beta}^*}\right) \right] \tag{2.1.18}$$

where $\mathbf{y}_{(1)} = \{y_i : I_i = 1\}$. Let $Q = 1 - \frac{1}{\Gamma(k)}\Gamma\left(k, y_i/e^{\mathbf{x}_i'\boldsymbol{\beta}^*}\right)$. Then

$$\begin{aligned}
\frac{\partial}{\partial\beta_j} l(\mathbf{y}_{(1)}) &= \frac{\partial}{\partial\beta_j} \sum_{i:I_i=1} \log(Q) \\
&= \sum_{i:I_i=1} \frac{1}{Q}\frac{\partial}{\partial\beta_j} Q \\
&= \sum_{i:I_i=1} \frac{1}{Q}\frac{-1}{\Gamma(k)}\frac{\partial}{\partial\beta_j}\Gamma\left(k, y_i/e^{\mathbf{x}_i'\boldsymbol{\beta}^*}\right) \\
&= \sum_{i:I_i=1} \frac{-1}{Q\Gamma(k)}\frac{\partial}{\partial\beta_j}\int_0^{y_i/e^{\mathbf{x}_i'\boldsymbol{\beta}^*}} W^{k-1}e^{-W}dW \\
&= \sum_{i:I_i=1} \frac{-1}{Q\Gamma(k)}\left[\int_0^{y_i/e^{\mathbf{x}_i'\boldsymbol{\beta}^*}} \frac{\partial}{\partial\beta_j} W^{k-1}e^{-W}dW + \left(\frac{y_i}{e^{\mathbf{x}_i'\boldsymbol{\beta}^*}}\right)^{k-1} e^{-\frac{y_i}{e^{\mathbf{x}_i'\boldsymbol{\beta}^*}}} \frac{\partial}{\partial\beta_j}\left(\frac{y_i}{e^{\mathbf{x}_i'\boldsymbol{\beta}^*}}\right)\right] \\
&= \sum_{i:I_i=1} \frac{-1(-x_{ij})}{\Gamma(k) - \Gamma\left(k, y_i/e^{\mathbf{x}_i'\boldsymbol{\beta}^*}\right)} \left(\frac{y_i}{e^{\mathbf{x}_i'\boldsymbol{\beta}^*}}\right)^k e^{-\frac{y_i}{e^{\mathbf{x}_i'\boldsymbol{\beta}^*}}} \\
&= \sum_{i:I_i=1} \frac{x_{ij}}{\Gamma\left(k, y_i/e^{\mathbf{x}_i'\boldsymbol{\beta}^*}\right)} \left(\frac{y_i}{e^{\mathbf{x}_i'\boldsymbol{\beta}^*}}\right)^k e^{-\frac{y_i}{e^{\mathbf{x}_i'\boldsymbol{\beta}^*}}} \qquad (2.1.19)
\end{aligned}$$

where $\Gamma\left(k, e^{\mathbf{x}_i'\boldsymbol{\beta}^*}\right)$ is the upper incomplete gamma distribution. When shape parameter $k = 1$, (2.1.16) and (2.1.19) are identical to their counterparts in GLM, (2.1.8) and (2.1.12), which means the same maximum likelihood estimates for $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}$. When $k \neq 1$ but is known, theorem 2 applies.

The log-likelihood of all observations is again $l(\mathbf{y}) = l(\mathbf{y}_{(0)}) + l(\mathbf{y}_{(1)})$ and the Newton-Raphson algorithm can be used to find MLEs for $\boldsymbol{\beta}^*$ and k. Regardless of censoring, the shape parameter k always needs to be estimated, unless it is known.

The findings above have an important application. In SAS, PROC GENMOD, the main procedure for fitting GLM, does not handle censoring; but PROC LIFEREG, the counterpart for AFT, does. Therefore, when the underlying distribution is gamma and the link function is log, one can use PROC LIFEREG to obtain pa-

rameter estimates related to those in GLM, achieving much convenience.

Considerations for Generalization

Although the survival function method is not limited to GLM based on the gamma distribution, Theorems 1 and 2 do not hold in general. This is because GLM and AFT focus on modeling different aspects of an underlying distribution. The GLM links the mean of a random variable to a linear predictor while the AFT links the scale parameter to the linear predictor. For a distribution to be applicable in both GLM and AFT, it needs to be a member of the Exponential family with both a finite mean and a scale parameter, for example, gamma, normal, log-normal, and inverse Gaussian distributions. Distributions that are not members of the Exponential family such as Weibull and Student's t do not work in GLM, while distributions that work in GLM but are not a scale-family member, for example, Poisson, do not work in AFT.

Even when a distribution is appropriate for both GLM and AFT, the relationship between the estimated parameters in GLM and AFT depends on the specific distribution and the link function in GLM. Theoretically, theorems 1 and 2 hold for other Exponential family distributions used in GLM with a log link as long as they have a scale parameter and the mean is a multiple of the scale parameter. However, many common distributions do not have this property. For instance, the normal distribution has both a finite mean and a scale parameter and thus can be used in both GLM and AFT. But since the mean and the scale are independent, the linear predictor's parameters in GLM have no relationship to their counterparts in AFT.

More general likelihood-based approaches to right-censoring in GLM remain a subject of further research.

2.2 Nested Observations

2.2.1 Cross-sectioning the Data

Although the measurements of the same claimants are correlated, the claimants are assumed to be independent. So if a modeling sample includes only 1 observation of each claimant, the observations will be independent of each other. Therefore, the first approach is to eliminate the correlation among nested observations by cross-sectioning the overall sample by time period.

Table 2.2.1: Sample by Time Period

Overall Sample

	Cum. Payment	Gender	No. of Visits	Duration	Censoring Indicator
	Y	X_1	X_2	T	I
Claimant 1	0	M	1	0	0
	1000	M	2	1	0
	1500	M	4	2	0
	1800	M	5	3	0
	⋮	⋮	⋮	⋮	⋮
	6000	M	5	6	1
Claimant 2	100	F	3	0	0
	2000	F	6	1	0
	4500	F	7	3	0
	4900	F	8	6	0
	⋮	⋮	⋮	⋮	⋮
	8000	F	9	10	1
⋮	⋮	⋮	⋮	⋮	⋮

3-Month Sample

	Cum. Payment	Gender	No. of Visits	Duration	Censoring Indicator
	Y	X_1	X_2	T	I
Claimant 1	1800	M	5	**3**	0
Claimant 2	4500	F	7	**3**	0
⋮	⋮	⋮	⋮	⋮	⋮

6-Month Sample

	Cum. Payment	Gender	No. of Visits	Duration	Censoring Indicator
	Y	X_1	X_2	T	I
Claimant 1	6000	M	5	**6**	1
Claimant 2	4900	F	8	**6**	0
⋮	⋮	⋮	⋮	⋮	⋮

Note: Observations in the 3-month and 6-month sub-samples are independent respectively.

As Table 2.2.1 on page 22 illustrates, after being open for 3 months, for claimant

1 there have been 5 hospital visits and the cumulative payment has reached $\$1800$, while for claimant 2 there have been 7 visits and $\$4500$ in cumulative payment. And these two observations are independent. So if we compose a sub-sample of only the 3-month measurements of each claimant, it will consist of independent observations. The same goes for 6-month, 9-month sub-samples and so on.

The cross-sectioning approach is the basis for methods 1 and 2, which are detailed in Chapters 3 and 4.

2.2.2 Modeling the Covariance

Approach 2 incorporates random effects to model the covariance structure of the nested observations in the overall sample. In linear mixed-effect models [15, Chp. 3, p. 56] [13, Chp. 10, p. 396],

$$\mathbf{y} = \mathbf{X}\tau + \mathbf{Z}\gamma + \epsilon \tag{2.2.1}$$

where

- $\mathbf{y}_{N\times 1}$ is the response vector; $\mathbf{X}_{N\times p}$ and $\mathbf{Z}_{N\times q}$ are covariate matrices; $\tau_{p\times 1}$ and $\gamma_{q\times 1}$ represent fixed effects and random effects respectively, and $\epsilon_{N\times 1}$ is the random error vector.
- Assuming $\gamma \sim N(\mathbf{0}, \boldsymbol{D_\gamma})$, $\epsilon \sim N(\mathbf{0}, \boldsymbol{R})$, $\mathbf{Cov}(\gamma, \epsilon) = \mathbf{0}$

Based on assumptions above, $\mathbf{y} \sim N(\mathbf{X}\tau, \mathbf{Z}\boldsymbol{D_\gamma}\mathbf{Z}' + \boldsymbol{R})$. Evidently, this would not apply in my case because the marginal distribution of any individual random variable in vector $\mathbf{y}$ would be normal instead of the gamma distribution.

Generalized Linear Mixed Model (GLMM) combines features from both GLM and the linear mixed model. For GLM based on the gamma distribution and the log link 2.1.2, its GLMM counterpart is

$$\log(\mathbf{E}(\mathbf{y})) = \mathbf{X}\tau + \mathbf{Z}\gamma \tag{2.2.2}$$

where the marginal distribution of each y_i in $\mathbf{y}$ is assumed to be gamma while other assumptions are the same as those for equation (2.2.1).

Next, we need to find the joint-likelihood of $\mathbf{y}$, which is no longer multivariate normal. A typical assumption is that the nested observations are conditionally independent given the group effect. That would assume the measurements of the same claimant over time are independent of each other conditional upon the effect of the claimant, in which case the joint-likelihood of $\mathbf{y}$ would simply be the product of the individual likelihood contributions of y_i. However, this is problematic because if the dependent variable is cumulative payment, it is hard to assume later values are independent of earlier ones as it is non-decreasing over time.

In Chapter 5, I propose a method to generate the likelihood for $\mathbf{y}$ taking into account concerns above and right censoring, and a way to transform the data so that the assumption of conditional independence would be more acceptable.

2.3 Time-Dependent Predictors vs. Fixed Target

The approaches to handle this challenge is closely related to those for nested observations. As methods 1 and 2 eliminate the correlation among measurements of the same claimants, the issue of time-varying predictors versus the fixed target

also vanishes. This is because in the sub-samples (Table 2.2.1), all observations are a snapshot of different claimants at the same point in the claim's life cycle, for example, 3 months after a claim has been open. There is no longitudinal information in each sub-sample.

When the model is based on the overall sample, longitudinal information is present and must be dealt with appropriately. Chapter 5 details approaches to handle this challenge.

Chapter 3

Method 1: Series of GLM and AFT models

3.1 Methodology

While working on worker's compensation claims around 2007, I started developing this method, in which a series of models are built at selected time points. For example, 9 sets of GLM and AFT models are built at months 3, 6, 9, 12, 15, 18 21, 24, and 36 after a claim opens for a total of 18 models.

At each time t, the modelling sample consists of claims that have lasted beyond that point. For instance, if a claim is closed after 8 months, then the measurement of it as of month 3 is included in the sample for month-3 models. Its month-6 but not its month-3 measurement is used in building month-6 models. None of its measurements are used for month 9 or any later models because it has closed by then. This means the sample size for each t is non-increasing. See Table 2.2.1 on

page 22 for an illustration.

Regardless of time points, the target variable remains the same for each claim, which is, y_{ic_i}, the ultimate payment for the claim [1].

Let d_i be the duration of a claim, or how long a claim lasts until it closes. Empirically, the longer a claim lasts, the higher its ultimate payment tends to be, all else being equal. So it is reasonable to attempt to use duration, d_i, as a predictor of claim severity y_{ic_i}. However, in operation, when we try to predict the ultimate payment for claims still open, say, after 3 months, d_i is not observed yet. To address this issue, an AFT model is first built to predict d_i. Then the predicted duration is used in place of d_i in a GLM to predict y_{ic_i}.

Let $\tilde{\mathbf{x}}'_{it} = \left(\mathbf{x}'_{it}, y_{it}, \widehat{(d_i)_t}\right)$ represent the explanatory variable vector including covariates, the cumulative payment for claim i, and the predicted duration of the claim $\widehat{(d_i)_t}$, all as of time t. Let $\mathbf{w}'_{it} = (\mathbf{x}'_{it}, y_{it})$. Note $\tilde{\mathbf{x}}'_{it}$ and $\mathbf{w}'_{it}$ represent vectors of available predictors at t. The significant predictors in the models at different t may be different.

At each time $t = 1, 2, \cdots$, an AFT model A_t is fitted to predict the duration d_i, while a GLM M_t is built to predict the ultimate payment y_{ic_i}.

A_t: $d_i|\mathbf{w}'_{it} \sim$ Weibull$(\alpha^*_t, \theta^*_{it})$

$$\log d_i = \mathbf{w}'_{it}\boldsymbol{\beta_t}^* + \sigma^*_t\epsilon^*_{it}, \text{ where } \log(\theta^*_{it}) = \mathbf{w}'_{it}\boldsymbol{\beta_t}^* \text{ and } 1/\alpha^*_t = \sigma^*_t$$

M_t: $y_{ic_i}|\tilde{\mathbf{x}}'_{it} \sim \Gamma(\alpha_t, \theta_{it})$

$$\log E(y_{ic_i}) = \tilde{\mathbf{x}}'_{it}\boldsymbol{\beta_t}, \text{ so } \log(\alpha_t\theta_{it}) = \tilde{\mathbf{x}}'_{it}\boldsymbol{\beta_t}$$

where

[1] See definitions on page 2 unless otherwise defined in this chapter

- α_t^* and θ_{it}^* are Weibull distribution shape and scale parameters.
- α_t and θ_{it} are gamma distribution shape and scale parameters.
- $\boldsymbol{\beta_t}$ and $\boldsymbol{\beta_t}^*$ are the coefficients for the predictors.
- ϵ_{it}^* is a random disturbance term with 0 mean.

The Weibull density is

$$\frac{\alpha_t^*}{\theta_{it}^*}\left(\frac{d_i}{\theta_{it}^*}\right)^{\alpha_t^*-1}\exp\left(-\left(\frac{d_i}{\theta_{it}^*}\right)^{\theta_{it}^*}\right) \tag{3.1.1}$$

Since the available predictors for duration are the same for claim severity, it is often asked why not just build one model M_t at each t to predict y_{ic_i} directly without building A_t. This is because, although the available predictors are the same, the significant ones for d_i may not be the same as those for y_{ic_i}. Therefore, using $\widehat{(d_i)_t}$ to predict y_{ic_i} may allow predictors that would be otherwise insignificant on their own in M_t to become significant via $\widehat{(d_i)_t}$, which is a function of significant predictors for duration. Example 3 demonstrates this idea.

Example 3. Without loss of generality, assume Y and X_1 are two random variables with $\sigma_Y = \sigma_{X_1} = 1$ and $\sigma_{Y,X_1} = 0.3$. Further, assume $X_2 = Y - X_1$ so that $\sigma_{Y,X_2} = 0.7$. Next let us generate a random sample of N observations, (y_i, x_{1i}, x_{2i}), $i = 1, 2, \cdots, N$. Suppose we are to fit a multiple linear regression (MLR) model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$ with all usual assumptions. Then depending on the threshold, we may only select one significant variable, X_2.

If, however, we create $X_3 = X_1 + X_2 = Y$ so that $\sigma_{Y,X_3} = 1$ and include it as a candidate predictor for the MLR model above, we will select X_3 as the only

significant variable because *(a)* it has a stronger correlation with Y and *(b)* it avoids multicollinearity issues present when both X_1 and X_2 are included in the model.

In conclusion, a function, a linear combination in this case, of predictors improves the model by being a more significant predictor than each original predictors and avoids potential multicollinearity.

To handle right-censoring in GLM based on the gamma distribution with a log link for open claims in M_t, a corresponding AFT model is fitted based on Theorems 1 and 2 as follows. See Section 2.1.2.

A_t: $d_i|\mathbf{w}'_{it} \sim \text{Weibull}(\alpha^*_t, \theta^*_{it})$

$$\log d_i = \mathbf{w}'_{it}\boldsymbol{\beta_t}^* + \sigma^*_t\epsilon^*_{it}, \text{ where } \log(\theta^*_{it}) = \mathbf{w}'_{it}\boldsymbol{\beta_t}^* \text{ and } 1/\alpha^*_t = \sigma^*_t$$

M_t: $y_{ic_i}|\tilde{\mathbf{x}}'_{it} \sim \Gamma(\alpha_t, \theta_{it})$

$$\log y_{ic_i} = \tilde{\mathbf{x}}'_{it}\boldsymbol{\beta_t} + \sigma_t\epsilon_{it}, \text{ where } \log(\theta_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta_t}$$

3.2 Data Results

Two sets of models are built for 3 months and 6 months respectively using SAS PROC LIFEREG.

3.2.1 3-Month Models

The duration model, A_3, is an AFT model based on the Weibull distribution as described in Table 3.2.1 on page 30.

Table 3.2.1: 3-Month Duration Model Information

The LIFEREG Procedure

Model Information	
Dependent Variable	Log(dur_day)
Censoring Variable	cls_ind
Censoring Value(s)	0
Number of Observations	756
Noncensored Values	644
Right Censored Values	112
Number of Parameters	5
Name of Distribution	Weibull
Log Likelihood	-625.595161
Number of Observations Read	768
Number of Observations Used	756
Missing Values	12

The candidate predictors are described on page 7, some of which are highly correlated. Table 3.2.2 shows the correlation matrix for cumulative numbers of bills, treatments and office visits. The positive correlations among them make sense as the more a claimant visits a medical office, the more treatments he or she is likely to receive, resulting in more bills.

Table 3.2.2: Correlation Matrix for Select 3-Month Variables

Variable	Cum. NO. of Bills	Cum. NO. of Treatments	Cum. NO. of Office Visits
Cum. NO. of Bills	1	0.62334	0.50947
Cum. NO. of Treatments		1	0.62759
Cum. NO. of Office Visits			1

In order to build a relatively small model with good predictive performance, model selection criteria such as Bayesian Information Criteria (BIC), type III analysis, and multicollinearity diagnostics are applied. The final A_3 has following pre-

dictors:

- Age: claimant age at the beginning of the claim
- Cum_Chro: cumulative number of chiropractic treatments received by the end of 3 months
- Cum_Offc_Visit: cumulative number of visits to medical office

Table 3.2.3 on page 31 shows all three predictors are significant and have positive coefficient estimates.

Table 3.2.3: 3-Month Duration Model Results

Fit Statistics	
-2 Log Likelihood	1251.190
AIC (smaller is better)	1261.19
AICC (smaller is better)	1261.27
BIC (smaller is better)	1284.331

	Type III Analysis of Effects		
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	7.2305	0.0072
Cum_Chro	1	47.0483	<.0001
Cum_Offc_Vist	1	39.0152	<.0001

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5.1543	0.0587	5.0393	5.2694	7709.83	<.0001
Age	1	0.0032	0.0012	0.0009	0.0055	7.23	0.0072
Cum_Chro	1	0.0105	0.0015	0.0075	0.0135	47.05	<.0001
Cum_Offc_Vist	1	0.0452	0.0072	0.031	0.0594	39.02	<.0001
Scale	1	0.4949	0.0139	0.4684	0.523		
Weibull Shape	1	2.0205	0.0568	1.9122	2.135		

The payment model M_3 is an AFT model based on the gamma distribution. Note the default gamma distribution PROC LIFEREG uses is the 3-parameter gen-

eralized gamma distribution. Tables 3.2.4 on page 32 and 3.2.5 on page 33 show the results of the selected model based on the generalized gamma distribution.

Table 3.2.4: 3-Month Payment Model Information
3-Parameter Gamma

The LIFEREG Procedure

Model Information	
Dependent Variable	Log(CLMNT_Total_Paid)
Censoring Variable	cls_ind
Censoring Value(s)	0
Number of Observations	740
Noncensored Values	629
Right Censored Values	111
Number of Parameters	6
Zero or Negative Response	16
Name of Distribution	Gamma
Log Likelihood	-1031.670991
Number of Observations Read	768
Number of Observations Used	740
Missing Values	12

The significant predictors are

- Cum_Pay: cumulative payment made after 3 months
- Xb_Dur: linear predictor for claim duration from the 3-month duration model
- Cum_CT: cumulative number of CT scans by the end of 3 months.

They are quite different from those in A_3 with Xb_Dur, which is a linear combination of significant predictors for duration, being the most significant predictor. This provides empirical evidence for Example 3. The next most significant variable is Cum_Pay, not surprisingly, followed by a relatively weak but still significant Cum_CT.

To fit a 2-parameter gamma model in PROC LIFEREG, we need to use the technique described in Appendix B to search for a common scale and shape parameter

Table 3.2.5: 3-Month Payment Model Results
3-Parameter Gamma

Fit Statistics	
-2 Log Likelihood	2063.342
AIC (smaller is better)	2075.342
AICC (smaller is better)	2075.457
BIC (smaller is better)	2102.982

	Type III Analysis of Effects		
Effect	DF	Wald Chi-Square	Pr > ChiSq
Cum_Pay	1	98.7248	<.0001
Xb_Dur	1	227.4061	<.0001
Cum_CT	1	7.8615	0.005

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-7.5671	1.0809	-9.6857	-5.4485	49.01	<.0001
Cum_Pay	1	0.0746	0.0075	0.0599	0.0893	98.72	<.0001
Xb_Dur	1	2.8794	0.1909	2.5052	3.2537	227.41	<.0001
Cum_CT	1	0.0383	0.0137	0.0115	0.0651	7.86	0.005
Scale	1	1.0969	0.0321	1.0357	1.1617		
Shape	1	0.2722	0.0756	0.1241	0.4204		

defined by SAS. After rigorous search, that common MLE is found to be 0.9897 with the corresponding gamma shape parameter being 1.0209. This is very close in shape to an exponential distribution. Therefore, for familiarity in practice, 1 is selected to be the gamma shape parameter, which leads to $\hat{\delta} = \hat{\sigma} = 1/\sqrt{1} = 1$.

Tables 3.2.6 on page 34 and 3.2.7 on page 34 display results for M_3 based on the exponential distribution.

The overall log-likelihood, -1078.2812, is smaller than that based on the generalized gamma distribution, -1031.6710. This is expected, as the model based on the 3-parameter gamma has one more degree of freedom, which may result in better fit. The order of significant predictors remain the same, although Cum_CT is now marginally significant.

Table 3.2.6: 3-Month Payment Model Information
2-Parameter Gamma

The LIFEREG Procedure

Model Information	
Dependent Variable	Log(CLMNT_Total_Paid)
Censoring Variable	cls_ind
Censoring Value(s)	0
Number of Observations	740
Noncensored Values	629
Right Censored Values	111
Number of Parameters	4
Zero or Negative Response	16
Name of Distribution	Gamma
Log Likelihood	-1078.281226
Number of Observations Read	768
Number of Observations Used	740
Missing Values	12

Table 3.2.7: 3-Month Payment Model Results
2-Parameter Gamma

Fit Statistics	
-2 Log Likelihood	2156.562
AIC (smaller is better)	2164.562
AICC (smaller is better)	2164.617
BIC (smaller is better)	2182.989

	Type III Analysis of Effects		
Effect	DF	Wald Chi-Square	Pr > ChiSq
Cum_Pay	1	70.619	<.0001
Xb_Dur	1	132.7492	<.0001
Cum_CT	1	2.6686	0.1023

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-1.9977	0.9502	-3.86	-0.1353	4.42	0.0355
Cum_Pay	1	0.0685	0.0081	0.0525	0.0844	70.62	<.0001
Xb_Dur	1	1.9631	0.1704	1.6291	2.297	132.75	<.0001
Cum_CT	1	0.0208	0.0127	-0.0042	0.0457	2.67	0.1023
Scale	0	1	0	1	1		
Shape	0	1	0	1	1		

3.2.2 6-Month Models

Tables 3.2.8 on page 35 and 3.2.9 on page 36 are the results for the 6-month duration model A_6.

Table 3.2.8: 6-Month Duration Model Information

The LIFEREG Procedure

Model Information	
Dependent Variable	Log(dur_day)
Censoring Variable	cls_ind
Censoring Value(s)	0
Number of Observations	350
Noncensored Values	281
Right Censored Values	69
Number of Parameters	4
Name of Distribution	Weibull
Log Likelihood	-252.2708727
Number of Observations Read	356
Number of Observations Used	350
Missing Values	6

The significant predictors are

- Age: claimant age at the beginning of the claim
- Cum_Trmt: cumulative number of all treatments by the end of 6 months.

They are not all the same as those in A_3. Age continues to be significant yet the cumulative number of all treatments has replaced the cumulative numbers of chiropractics and medical office visits.

For the 6-month payment model M_3, Tables 3.2.10 and 3.2.11 show the results based on the generalized gamma distribution.

Next, Tables 3.2.12 on page 38 and 3.2.13 on page 39 contain the results for

Table 3.2.9: 6-Month Duration Model Results

Fit Statistics	
-2 Log Likelihood	504.542
AIC (smaller is better)	512.542
AICC (smaller is better)	512.658
BIC (smaller is better)	527.973

	Type III Analysis of Effects		
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	3.7342	0.0533
Cum_Trmt	1	19.7226	<.0001

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5.6242	0.0837	5.4602	5.7883	4513.19	<.0001
Age	1	0.0031	0.0016	0	0.0063	3.73	0.0533
Cum_Trmt	1	0.0007	0.0002	0.0004	0.001	19.72	<.0001
Scale	1	0.4142	0.0181	0.3801	0.4512		
Weibull Shape	1	2.4145	0.1056	2.2161	2.6306		

Table 3.2.10: 6-Month Payment Model Information
3-Parameter Gamma

The LIFEREG Procedure

Model Information	
Dependent Variable	Log(CLMNT_Total_Paid)
Censoring Variable	cls_ind
Censoring Value(s)	0
Number of Observations	348
Noncensored Values	279
Right Censored Values	69
Number of Parameters	7
Zero or Negative Response	2
Name of Distribution	Gamma
Log Likelihood	-333.4970682
Number of Observations Read	356
Number of Observations Used	348
Missing Values	6

Table 3.2.11: 6-Month Payment Model Results
3-Parameter Gamma

Fit Statistics	
-2 Log Likelihood	666.994
AIC (smaller is better)	680.994
AICC (smaller is better)	681.324
BIC (smaller is better)	707.96

	Type III Analysis of Effects		
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	14.6767	0.0001
Xb_Dur	1	58.4478	<.0001
Cum_Pay	1	139.0451	<.0001
Cum_ASTH	1	6.8502	0.0089

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-9.2864	2.3704	-13.9324	-4.6404	15.35	<.0001
Age	1	-0.0112	0.0029	-0.017	-0.0055	14.68	0.0001
Xb_Dur	1	3.1338	0.4099	2.3304	3.9372	58.45	<.0001
Cum_Pay	1	0.0488	0.0041	0.0407	0.0569	139.05	<.0001
Cum_ASTH	1	0.0737	0.0281	0.0185	0.1288	6.85	0.0089
Scale	1	0.6868	0.0298	0.6307	0.7478		
Shape	1	0.1535	0.1103	-0.0628	0.3697		

M_6 based on the 2-parameter gamma distribution. For details on fitting the 2-parameter gamma model, see Appendix B.

The significant predictors are

- Age: claimant age at the beginning of the claim
- Xb_Dur: linear predictor for claim duration from the 6-month duration model
- Cum_Pay: cumulative payment made after 6 months
- Cum_ASTH: cumulative number of anaesthesia treatments received by the

Table 3.2.12: 6-Month Payment Model Information
2-Parameter Gamma

The LIFEREG Procedure

Model Information	
Dependent Variable	Log(CLMNT_Total_Paid)
Censoring Variable	cls_ind
Censoring Value(s)	0
Number of Observations	348
Noncensored Values	279
Right Censored Values	69
Number of Parameters	5
Zero or Negative Response	2
Name of Distribution	Gamma
Log Likelihood	-343.798698
Number of Observations Read	356
Number of Observations Used	348
Missing Values	6

end of 3 months.

Not only are they different from the significant variables in A_6, but also they are from the significant ones in M_3. Moreover, the shape parameter of the gamma distribution underlying M_6 is 2.295, quite different from that for M_3, which is 1. See Table B.0.2 on page 95.

3.3 Pros and Cons

Comparing 3-month with 6-month models shows the flexibility of this method. Not only do the models at different time have different sets of significant predictors, but also they have different shape parameters. The independence of models at different time allows each model to use information available at each time to make the best prediction possible, or find local optimum.

Table 3.2.13: 6-Month Payment Model Results
2-Parameter Gamma

Fit Statistics	
-2 Log Likelihood	687.597
AIC (smaller is better)	697.597
AICC (smaller is better)	697.773
BIC (smaller is better)	716.858

	Type III Analysis of Effects		
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	5.9629	0.0146
Xb_Dur	1	37.657	<.0001
Cum_Pay	1	99.9577	<.0001
Cum_ASTH	1	11.6099	0.0007

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-4.7238	2.2102	-9.0558	-0.3919	4.57	0.0326
Age	1	-0.007	0.0029	-0.0126	-0.0014	5.96	0.0146
Xb_Dur	1	2.3591	0.3844	1.6056	3.1126	37.66	<.0001
Cum_Pay	1	0.0478	0.0048	0.0384	0.0572	99.96	<.0001
Cum_ASTH	1	0.1063	0.0312	0.0451	0.1674	11.61	0.0007
Scale	0	0.6601	0	0.6601	0.6601		
Shape	0	0.6601	0	0.6601	0.6601		

Because each claim only appears at most once in a given modelling sample in the series, this method bypasses the issue of dealing with correlated measurements of the same claim over time. The issue of predicting a fixed target with time-dependent variables also disappears, because when a predictor's value changes, we just build new models to obtain new predictions. In addition, since the models at different time are independent of each other, this method is not affected by whether the target is fixed or not (see Section 2.3 on page 24)

However, this method brings its own challenges. *(a)* each model is a snapshot

independent of the others, meaning knowledge gained from the 3-month models is not leveraged in the 6-month models; *(b)* the decreasing sample size over time increases the standard error of coefficient estimates; and *(c)* the cost of building a large number of models can be high. In theory, each time a time-dependent variable's value changes for any claim, two new models should be built. Potentially we could be building models for each month, week, or day, making the effort daunting.

Chapter 4

Method 2: Series of Bayesian Models

4.1 Methodology

Method 2 also builds sets of models at each t, much like Method 1, but these are Bayesian models and this method tries to make use of learning from previous time by incorporating information from earlier models into prior distributions for the current models. It is developed based on Professor Lynn Kuo's idea.

At each time $t = 1, 2, \cdots$, a model A_t (Level 1a) is fitted to predict the duration while a model M_t (Level 1b) is built to predict the ultimate payment. Coefficient estimates from the previous period $t-1$ are used as priors for the current period.

Level 1a: $d_i|\mathbf{w}'_{it} \sim$ Weibull$(\alpha^*_t, \theta^*_{it})$

$$\log d_i = \mathbf{w}'_{it}\boldsymbol{\beta_t}^* + \sigma^*_t\epsilon^*_{it}, \text{ where } \log(\theta^*_{it}) = \mathbf{w}'_{it}\boldsymbol{\beta_t}^* \text{ and } 1/\alpha^*_t = \sigma^*_t$$

Level 1b: $y_{ic_i}|\tilde{\mathbf{x}}'_{it} \sim \Gamma(\alpha_t, \theta_{it})$

$\log y_{ic_i} = \tilde{\mathbf{x}}'_{it}\boldsymbol{\beta_t} + \sigma_t\epsilon_{it}$, where $\log(\theta_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta_t}$

Level 2a: $\alpha_t^* \sim \pi_{\alpha_t^*}, \boldsymbol{\beta_t^*} \sim \pi_{\boldsymbol{\beta_t^*}}, \sigma_t^* \sim \pi_{\sigma_t^*}$

Level 2b: $\alpha_t \sim \pi_{\alpha_t}, \boldsymbol{\beta_t} \sim \pi_{\boldsymbol{\beta_t}}, \sigma_t \sim \pi_{\sigma_t}$

where $\tilde{\mathbf{x}}'_{it}$, $\mathbf{w}'_{it}$, α_t^*, α_t, θ_{it}^*, θ_{it}, $\boldsymbol{\beta_t}^*$, $\boldsymbol{\beta_t}$, and ϵ_{it}^* are defined as in Section 3.1, while Levels 2a and 2b represent the prior distributions of the parameters in M_t and A_t respectively.

4.1.1 Prior Distributions

Let us first consider power priors [2, Ibrahim and Chen] [14]. Let $\boldsymbol{\eta_t}$ represent the vector of all parameters at t. The power prior for $\boldsymbol{\eta_t}$ is then

$$\pi_{\boldsymbol{\eta_t}} \propto L\left(\mathbf{y}_{t-1}|D_{t-1}\right)^{a_0} \pi_{0\boldsymbol{\eta_t}} \tag{4.1.1}$$

where

- $L\left(\mathbf{y}_{t-1}|D_{t-1}\right)$ is the likelihood from data at $t-1$, D_{t-1}.
- $0 \leqslant a_0 \leqslant 1$ controls how much historic information is incorporated into the prior distribution for the current period.
- $\pi_{0\boldsymbol{\eta_t}}$ is what the prior would be without historic information.

This leads to the posterior distribution below

$$h\left(\boldsymbol{\eta_t}|D_t, D_{t-1}\right) \propto \left(L\left(\mathbf{y}_t|D_t\right) L\left(\mathbf{y}_{t-1}|D_{t-1}\right)^{a_0}\right) \pi_{0\boldsymbol{\eta_t}} \tag{4.1.2}$$

Equation (4.1.2) implies that D_t and D_{t-1} are independent so that $L\left(\mathbf{y}_t|D_t\right)$ and $L\left(\mathbf{y}_{t-1}|D_{t-1}\right)$ can be multiplied. It essentially augmented the current dataset with data from the previous period. However, observations in D_{t-1} and D_{t-1} may be measurements of the same claim and thus, not independent. Therefore the power prior method is not adopted in this case.

Alternatively, I considered the asymptotic posterior distribution [8, Lauritzen]. Let $l(\boldsymbol{\eta}_t) = \log L\left(\mathbf{y}_t|D_t\right)$, the log-likelihood function, and

$$\overline{l\left(\boldsymbol{\eta}_t\right)} = \frac{1}{N_t}\sum_{i=1}^{N_t} \log f\left(y_i|\boldsymbol{\eta}_t\right), \tag{4.1.3}$$

its sample average. Then we have

$$\begin{aligned} h\left(\boldsymbol{\eta}_t|D_t\right) &\propto L\left(\mathbf{y}_t|D_t\right)\pi_{\boldsymbol{\eta}_t} \\ &\propto e^{l(\boldsymbol{\eta}_t)}\pi_{\boldsymbol{\eta}_t} \\ &\propto e^{N_t\overline{l(\boldsymbol{\eta}_t)}}\pi_{\boldsymbol{\eta}_t} \end{aligned} \tag{4.1.4}$$

where $\pi_{\boldsymbol{\eta}_t}$ is the prior. Based on the Law of Large Numbers, as $N_t \to \infty$, $\overline{l\left(\boldsymbol{\eta}_t\right)} \to \boldsymbol{E}\left[\log f\left(y_i|\boldsymbol{\eta}_t\right)\right]$. And for large enough N_t, variation in $h\left(\boldsymbol{\eta}_t|D_t,\right)$ will be dominated by contribution from $e^{N_t\overline{l(\boldsymbol{\eta}_t)}}$.

Using a Taylor expansion of $l(\boldsymbol{\eta_t})$ around MLE $\hat{\boldsymbol{\eta}}_{\boldsymbol{t}}$ in $h\left(\boldsymbol{\eta_t}|D_t\right)$ leads to

$$\begin{aligned} h\left(\boldsymbol{\eta_t}|D_t\right) &\propto \pi_{\boldsymbol{\eta_t}} \exp\left(N_t\overline{l\left(\hat{\boldsymbol{\eta}}_{\boldsymbol{t}}\right)} + l'\left(\hat{\boldsymbol{\eta}}_{\boldsymbol{t}}\right)\left(\boldsymbol{\eta}_t - \hat{\boldsymbol{\eta}}_{\boldsymbol{t}}\right) - \frac{1}{2}\left(\boldsymbol{\eta}_t - \hat{\boldsymbol{\eta}}_{\boldsymbol{t}}\right)'\boldsymbol{I}\left(\hat{\boldsymbol{\eta}}_{\boldsymbol{t}}\right)\left(\boldsymbol{\eta}_t - \hat{\boldsymbol{\eta}}_{\boldsymbol{t}}\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\eta}_t - \hat{\boldsymbol{\eta}}_{\boldsymbol{t}}\right)'\boldsymbol{I}\left(\hat{\boldsymbol{\eta}}_{\boldsymbol{t}}\right)\left(\boldsymbol{\eta}_t - \hat{\boldsymbol{\eta}}_{\boldsymbol{t}}\right)\right) \end{aligned} \tag{4.1.5}$$

where

- $l'(\hat{\boldsymbol{\eta}}_{\boldsymbol{t}}) = 0$, because $\hat{\boldsymbol{\eta}}_{\boldsymbol{t}}$ is the MLE,
- $\boldsymbol{I}(\hat{\boldsymbol{\eta}}_{\boldsymbol{t}}) = -\boldsymbol{H}(\hat{\boldsymbol{\eta}}_{\boldsymbol{t}})$ is the observed information matrix and $\boldsymbol{H}$ is the Hessian matrix.

Therefore, for large enough N_t the posterior distribution of $\boldsymbol{\eta}_t$ is approximately normal. According to [8, Lauritzen], using the posterior mode of $\boldsymbol{\eta}_t$ in expression (4.1.5) achieves a more accurate approximation. If the sampling posterior distribution resembles the normal distribution, then the posterior mean should be close to the mode and can also be used in the approximation.

The approximate normal posterior distribution at t is adopted as the prior for $t+1$ in my analysis. Since some significant predictors in A_t do not exist in A_{t-1}, non-informative priors are used for them while MLEs from $t-1$ are used where applicable.

4.1.2 MCMC

Assume the prior distribution for $\boldsymbol{\beta_t}$ is $N\left(\widehat{\boldsymbol{\beta}}_{t-1}, \widehat{\boldsymbol{\Sigma}}_{t-1}\right)$, where $\widehat{\boldsymbol{\beta}}_{t-1}$ and $\widehat{\boldsymbol{\Sigma}}_{t-1}$ are the MLEs of the coefficient vector and covariance matrix from the previous model, M_{t-1}. Similarly, assume the priori distribution for α_t is $N\left(\hat{\alpha}_{t-1}, \hat{\psi}_{t-1}\right)$.

The likelihood function is

$$L\left(\mathbf{y}|\widetilde{\mathbf{X}}_t, \boldsymbol{\beta_t}, \alpha_t, \sigma_t\right) = \prod_{\mathbf{y}_{(1)}} f\left(y_{ic_i}|\widetilde{\mathbf{X}}_t, \boldsymbol{\beta}_t, \alpha_t, \sigma_t\right) \prod_{\mathbf{y}_{(0)}} P\left(Y_i \geq y_{i,t_i^+}|\widetilde{\mathbf{X}}_t, \boldsymbol{\beta}_t, \alpha_t, \sigma_t\right), \tag{4.1.6}$$

where $\mathbf{y}_{(1)}$ represents the set of closed claims, while $\mathbf{y}_{(0)}$ the open or censored

claims. Again, following AFT's parametrization based on the gamma distribution (see (2.1.3) on page 13) gives us

$$\begin{aligned} f\left(y_{ic_i}|\widetilde{\mathbf{X}}_t, \boldsymbol{\beta}_t, \alpha_t, \sigma_t\right) &= \frac{\exp\left(\alpha_t\sigma_t\epsilon_{it} - e^{\sigma_t\epsilon_{it}}\right)}{y_{ic_i}\Gamma(\alpha_t)} \\ &= \frac{1}{y_{ic_i}\Gamma(\alpha_t)}\exp\left(\alpha_t\sigma_t\frac{\log(y_{ic_i}) - \left(\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta_t}\right)}{\sigma_t} - \exp\left(\sigma_t\frac{\log(y_{ic_i}) - \left(\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta_t}\right)}{\sigma_t}\right)\right) \\ &= \frac{1}{\Gamma(\alpha_t)e^{\left(\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta_t}\right)\cdot\alpha_t}}y_{ic_i}^{\alpha-1}\exp\left(\frac{-y_{ic_i}}{e^{\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta_t}}}\right) \end{aligned} \tag{4.1.7}$$

and

$$P\left(Y_i \geq y_{i,t_i^+}|\widetilde{\mathbf{X}}_t, \boldsymbol{\beta}_t, \alpha_t, \sigma_t\right) = \frac{1}{\Gamma(\alpha_t)}\Gamma\left(\alpha_t, y_{i,t_i^+}/e^{\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta_t}}\right), \tag{4.1.8}$$

where $\Gamma\left(\alpha_t, y_{i,t_i^+}/e^{\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta_t}}\right)$ is the upper incomplete gamma function. Note σ_t is cancelled out in equations (4.1.7) and (4.1.8). So $L\left(\mathbf{y}|\widetilde{\mathbf{X}}_t, \boldsymbol{\beta_t}, \alpha_t, \sigma_t\right) = L\left(\mathbf{y}|\widetilde{\mathbf{X}}_t, \boldsymbol{\beta_t}, \alpha_t\right)$.

The full conditional density of $\boldsymbol{\beta}_t$ therefore is

$$\begin{aligned} h\left(\boldsymbol{\beta}_t|\mathbf{y}, \widetilde{\mathbf{X}}_t, \alpha_t\right) &\propto g\left(\boldsymbol{\beta}_t\right) L\left(\mathbf{y}|\widetilde{\mathbf{X}}_t, \boldsymbol{\beta_t}, \alpha_t\right) \\ &\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta_t} - \widehat{\boldsymbol{\beta}}_{t-1}\right)'\left(\widehat{\boldsymbol{\Sigma}}_{t-1}\right)^{-1}\left(\boldsymbol{\beta_t} - \widehat{\boldsymbol{\beta}}_{t-1}\right)\right\}\cdot \\ &\prod_{\mathbf{y}_{(1)}}\frac{1}{e^{\left(\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta_t}\right)\cdot\alpha_t}}\exp\left(\frac{-y_{ic_i}}{e^{\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta_t}}}\right)\prod_{\mathbf{y}_{(0)}}\Gamma\left(\alpha_t, y_{i,t_i^+}/e^{\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta_t}}\right) \end{aligned} \tag{4.1.9}$$

Since (4.1.9) does not appear to come from a tractable distribution for direct sampling, we construct the following Metropolis-Hastings (M-H) algorithm [12] . Let $\widehat{\boldsymbol{\beta}}_{ML,t}$ and $\widehat{\boldsymbol{\Sigma}}_{ML,t}$ be the MLEs for $\boldsymbol{\beta}_t$ and its covariance matrix from the current time t.

Then at step j, the proposed value $\boldsymbol{\beta_{pro}}$ is sampled from $N\left(\boldsymbol{\beta}_t^{(j-1)}, \widehat{\boldsymbol{\Sigma}}_{ML,t}\right)$. So the proposal density is

$$q\left(\boldsymbol{\beta_{pro}}|\boldsymbol{\beta}_t^{(j-1)}, \mathbf{y}, \widetilde{\mathbf{X}}_t, \alpha_t^{(j-1)}\right) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta_{pro}} - \boldsymbol{\beta}_t^{(j-1)}\right)' \left(\widehat{\boldsymbol{\Sigma}}_{ML,t}\right)^{-1} \left(\boldsymbol{\beta_{pro}} - \boldsymbol{\beta}_t^{(j-1)}\right)\right\}, \tag{4.1.10}$$

where $\boldsymbol{\beta}_t^{(0)} = \widehat{\boldsymbol{\beta}}_{ML,t}$. The acceptance probability, which is the probability with which $\boldsymbol{\beta_{pro}}$ becomes $\boldsymbol{\beta}_t^{(j)}$, is calculated as

$$\begin{aligned}\phi\left(\boldsymbol{\beta_{pro}}|\boldsymbol{\beta}_t^{(j-1)}, \mathbf{y}, \widetilde{\mathbf{X}}_t, \alpha_t^{(j-1)}\right) &= \min\left\{1, \frac{h\left(\boldsymbol{\beta}_{pro}|\mathbf{y}, \widetilde{\mathbf{X}}_t, \alpha_t^{(j-1)}\right)}{h\left(\boldsymbol{\beta}^{(j-1)}|\mathbf{y}, \widetilde{\mathbf{X}}_t, \alpha_t^{(j-1)}\right)} \frac{q\left(\boldsymbol{\beta}_t^{(j-1)}|\boldsymbol{\beta_{pro}}, \mathbf{y}, \widetilde{\mathbf{X}}_t, \alpha_t^{(j-1)}\right)}{q\left(\boldsymbol{\beta_{pro}}|\boldsymbol{\beta}_t^{(j-1)}, \mathbf{y}, \widetilde{\mathbf{X}}_t, \alpha_t^{(j-1)}\right)}\right\} \\ &= \min\left\{1, \frac{h\left(\boldsymbol{\beta}_{pro}|\mathbf{y}, \widetilde{\mathbf{X}}_t, \alpha_t^{(j-1)}\right)}{h\left(\boldsymbol{\beta}^{(j-1)}|\mathbf{y}, \widetilde{\mathbf{X}}_t, \alpha_t^{(j-1)}\right)}\right\}, \end{aligned} \tag{4.1.11}$$

as $q(\cdot)$ is symmetric. Assuming $U \sim$ Uniform $(0, 1)$, we can obtain $\boldsymbol{\beta}_t^{(j)}$'s value as follows:

$$\boldsymbol{\beta}_t^{(j)} = \begin{cases} \boldsymbol{\beta_{pro}}, & \text{if } U \leq \phi\left(\boldsymbol{\beta_{pro}} \mid \boldsymbol{\beta}_t^{(j-1)}, \mathbf{y}, \widetilde{\mathbf{X}}_t, \alpha_t^{(j-1)}\right) \\ \boldsymbol{\beta}_t^{(j-1)}, & \text{otherwise} \end{cases} \tag{4.1.12}$$

Next we sample α_{pro} from $N\left(\alpha_t^{(j-1)}, \hat{\sigma}_{ML,t}\right)$, where $\hat{\alpha}_{ML,t}, \hat{\sigma}_{ML,t}$ are the MLEs of α_t and its standard deviation, and $\alpha_t^{(0)} = \hat{\alpha}_{ML,t}$. The proposal density is therefore

$$q'\left(\alpha_{pro}|\boldsymbol{\beta}_t^{(j)}, \mathbf{y}, \widetilde{\mathbf{X}}_t, \alpha_t^{(j-1)}\right) \propto \exp\left\{-\frac{1}{2}\left(\frac{\alpha_{pro} - \alpha_t^{(j-1)}}{\hat{\sigma}_{ML,t}}\right)^2\right\}. \tag{4.1.13}$$

And the acceptance probability is

$$\phi'\left(\alpha_{pro}|\boldsymbol{\beta}_t^{(j)}, \mathbf{y}, \widetilde{\mathbf{X}}_t, \alpha_t^{(j-1)}\right) = \min\left\{1, \frac{h\left(\alpha_{pro} \mid \mathbf{y}, \widetilde{\mathbf{X}}_t, \boldsymbol{\beta}_t^{(j)}\right)}{h\left(\alpha_t^{(j-1)} \mid \mathbf{y}, \widetilde{\mathbf{X}}_t, \boldsymbol{\beta}_t^{(j)}\right)}\right\}. \tag{4.1.14}$$

Assuming $U' \sim$ Uniform $(0, 1)$, we obtain $\alpha_t^{(j)}$'s value as follows:

$$\alpha_t^{(j)} = \begin{cases} \alpha_{pro}, & \text{if } U' \leq \phi'\left(\alpha_{pro}|\boldsymbol{\beta}_t^{(j)}, \mathbf{y}, \widetilde{\mathbf{X}}_t, \alpha_t^{(j-1)}\right) \\ \alpha_t^{(j-1)}, & \text{otherwise} \end{cases} \tag{4.1.15}$$

Step j completes when we obtain $\left(\boldsymbol{\beta}_t^{(j)}, \alpha_t^{(j)}\right)$. It is noted that α needs to be positive, being the shape parameter of a gamma distribution. If a negative value is sampled, skip it and sample another value until a positive value is obtained. Then the steps repeat. Sampling of posterior densities for $(\boldsymbol{\beta}_t^*, \alpha_t^*)$ in A_t follows the same algorithm.

4.2 Data Results

PROC MCMC in SAS is used for analysis. Information from A_3 and M_3 are used as priors for A_6 and M_6.

4.2.1 6-Month Duration Model

All duration models are based on the Weibull distribution with various combinations of set-up for the MCMC simulation. Table 4.2.1 on page 48 shows the

configurations of several representative simulation runs.

Table 4.2.1: 6-Month Duration Model MCMC Set-up

Model	Tuning	Burn-In	MCMC	Thin	Prior	Starting
1	1000	2000	50000	5	Non-Informative	A_6 MLE
2	1000	2000	50000	5	A_3 MLE	A_6 MLE
3	1000	2000	50000	5	Non-Informative	Not MLE
4	1000	5000	80000	8	Non-Informative	Not MLE
5	10000	50000	50000	5	Non-informative	Not MLE
6	10000	50000	50000	5	Non-informative	A_6 MLE

where

- Tuning means the number of iterations SAS runs to tune the proposal density.
- Burn-in means the number of burn-in iterations.
- MCMC is the sampling runs after burn-in iterations.
- Thinning = j means only every jth sample in the MCMC runs is kept, reducing autocorrelation between samples. All the models above end up with a posterior sample of $10,000$ units.
- For comparison, both non-informative priors and MLEs from the 3-month model are tried.
- Starting values are either the MLEs from the current time or other values. Experience shows that simulation runs starting at MLEs usually converge faster.

In general, the higher the numbers of tuning, burn-in, MCMC, and thinning runs, the better the sampling results are but also the longer the simulation takes.

Therefore, when selecting the final model, several factors are considered including Deviance Information Criteria (DIC) [4], autocorrelation between sampling steps, and the computing resources it consumes.

Table 4.2.2 on page 49 shows the comparison of 3 best Bayesian models in terms of DIC as well as the MLE model from Chapter 3.

Table 4.2.2: 6-Month Duration Model MCMC Comparison

Rankings In DIC	Model	CPU Time	Real Time	DIC	-2LL
1	5	24.4	24.77	3656.858	504.631
2	6	11.85	12.15	3663.308	512.659
3	1	5.69	6.05	3663.443	512.743
	MLE	0.04	0.05		504.542

Here,

- DIC measures the goodness of fit by comparing the log likelihood based on the Bayesian parameter estimates and the saturated log likelihood based on the data, while penalizing models with too many predictors, like Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The difference is that DIC uses posterior densities. For more details, see [4]. The smaller DIC, the more desirable the model is.

- CPU time and Real Time represent how long the simulation runs in seconds for a model. The longer it takes, the less preferable it is.

- -2LL means -2 Log Likelihood of the data based on estimated parameters of a model. Smaller numbers are better.

We can see from Table 4.2.2 that Model 5 achieves the lowest DIC and an almost as low -2LL as the MLE model does, which means this model is comparable to the MLE model in terms of goodness of fit. However, Model 5 also consumes a lot more computing resources than the MLE model as it takes 61 times as long in CPU time as the latter.

Focusing on Model 5, Figures 4.2.1 (a), (b), (c), (d) show the MCMC runs are stable with autocorrelation falling off quickly between sampling steps for all parameters, and the posterior sampling distributions resemble the normal distributions.

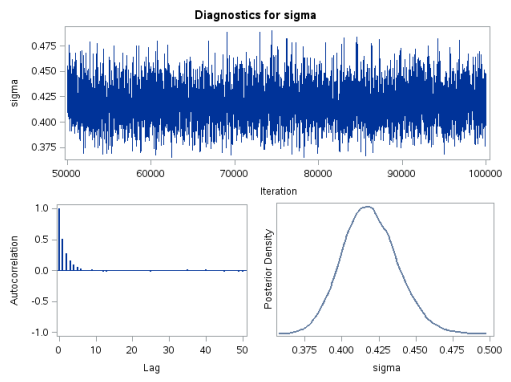
Further, the MCSE/SD ratios in Table 4.2.3(a) are very small, indicating only a small portion of the posterior variability is due to the Monte Carlo simulation. Table 4.2.3(b) shows by lag 10, autocorrelations have dropped to a very low level. Finally, Table 4.2.3(c) shows efficiency of the simulation is relatively high with good effective sample sizes.

The diagnostics indicate the simulation for Model 5 is satisfactory and we may use the posterior estimates. Tables 4.2.4(a) and 4.2.4(b) show the posterior statistics. Note the parameter estimates in Table 4.2.4(a) are very close to the MLEs in Table 3.2.9.

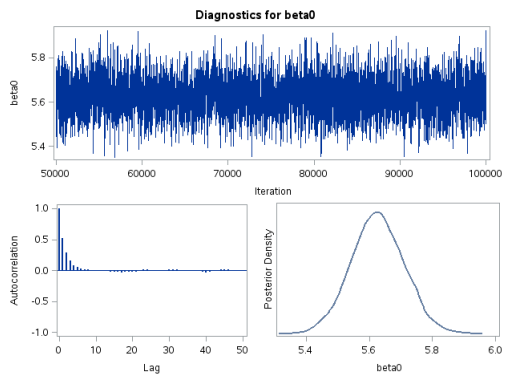
Since the best Bayesian model and the MLE model are similar in goodness of fit and parameter estimates, but the Bayesian model's cost of computing is a lot higher, the recommendation here is to just use the MLE model A_6 from Method 1.

Figure 4.2.1: 6-Month Duration Bayesian Model 5 Diagnostics Graphs

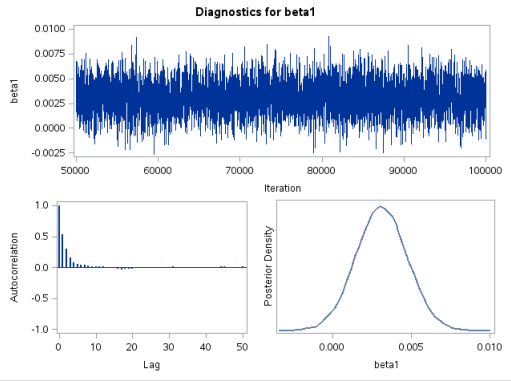
(a) SEV Scale



(b) Intercept



(c) Age



(d) Cum_Trmt

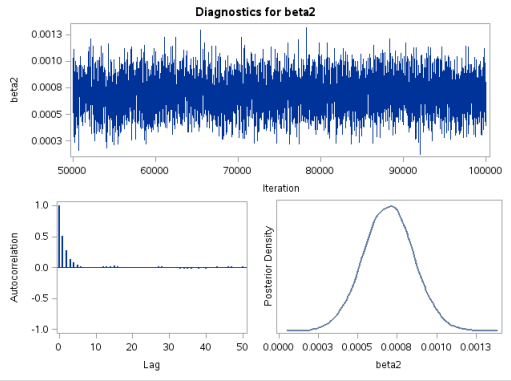


Table 4.2.3: 6-Month Duration Bayesian Model 5 Diagnostics

(a) Monte Carlo Standard Errors

Parameter	MCSE	Standard Deviation	MCSE/SD
sigma	0.00033	0.0184	0.018
beta0	0.00152	0.0838	0.0181
beta1	0.000031	0.00163	0.0189
beta2	0.000002792	0.000159	0.0176

(b) Posterior Autocorrelations

Parameter	Lag 1	Lag 5	Lag 10	Lag 50
sigma	0.5134	0.0532	-0.0102	-0.0194
beta0	0.5224	0.0538	0.0073	0.0005
beta1	0.5296	0.0559	0.0167	0.0127
beta2	0.5088	0.0426	-0.0023	0.0114

(c) Effective Sample Sizes

Parameter	ESS	Autocorrelation Time	Efficiency
sigma	3103.2	3.2224	0.3103
beta0	3036.3	3.2934	0.3036
beta1	2787.1	3.5879	0.2787

Table 4.2.4: 6-Month Duration Bayesian Model 5 Output

(a) Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
sigma	10000	0.4191	0.0184	0.4061	0.4183	0.4311
beta0	10000	5.6269	0.0838	5.5709	5.6258	5.6833
beta1	10000	0.00311	0.00163	0.00203	0.00311	0.00418
beta2	10000	0.000696	0.000159	0.000589	0.000696	0.000802

(b) Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
sigma	0.05	0.3849	0.4569	0.3841	0.4556
beta0	0.05	5.4598	5.7925	5.4577	5.7883
beta1	0.05	-0.0001	0.0063	-0.00019	0.00621
beta2	0.05	0.000383	0.00101	0.000379	0.001

4.2.2 6-Month Payment Model

The results presented here are based on the generalized gamma distribution as parametrized in Equation (4.1.7). Table 4.2.5 shows the set-ups of select models.

Table 4.2.5: 6-Month Payment Model MCMC Set-Up

Model	Tuning	Burn-In	MCMC	Thin	Prior	Starting	$\widehat{(d_i)_t}$
1	10000	50000	100000	10	Non-Informative	Not MLE	xb_dur_b
2	10000	50000	50000	5	Non-Informative	Not MLE	xb_dur_b
3	10000	50000	50000	5	M_3 MLE, N_7	M_6 MLE	xb_dur_b
4	1000	2000	50000	5	M_3 MLE, N_5	M_6 MLE	xb_dur_b
5	1000	2000	50000	5	M_3 MLE, N_7	M_6 MLE	xb_dur
6	1000	2000	50000	5	M_3 MLE, N_5	M_6 MLE	xb_dur

The same column headings in Table 4.2.5 and Table 4.2.1 have same meanings. Note,

- N_7 means the vector of all 7 parameters in M_6 is assumed to have a multivariate normal distribution, while N_5 means the vector of parameters in the linear predictor is assumed to have a 5-dimensional normal distribution, while the scale and shape parameters (see page 92 for details) are assumed to have prior distributions independent of N_5.

- The last column, $\widehat{(d_i)_t}$, indicates which linear predictor for duration from A_6 is used in M_6. "xb_dur" is the linear predictor based on MLEs from Method 1, while "xb_dur_b" is based on the posterior parameter estimates from Section 4.2.1 above. It is found that everything else being equal, the MCMC results do not differ much by the type of linear predictor used in this case. This is not surprising given the Bayesian estimates are very close to the MLEs (see Tables 4.2.4 and 3.2.9).

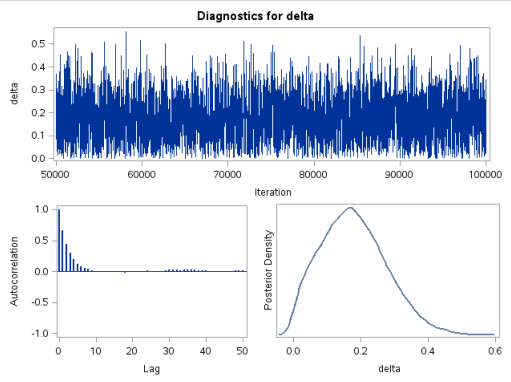
Table 4.2.6: 6-Month Payment Model MCMC Comparison

Rankings In DIC	Model	CPU Time	Real Time	DIC	-2LL
1	2	3:44	3:46	5781.653	667.17
2	1	4:34	4:40	5781.796	667.162
3	4	1:38	1:39	5832.05	724.457
4	6	1:37	1:38	5832.219	724.63
5	3	3:55	3:56	5835.542	727.921
6	5	1:34	1:34	5836.085	728.61
	MLE	0.02	0.02		666.994

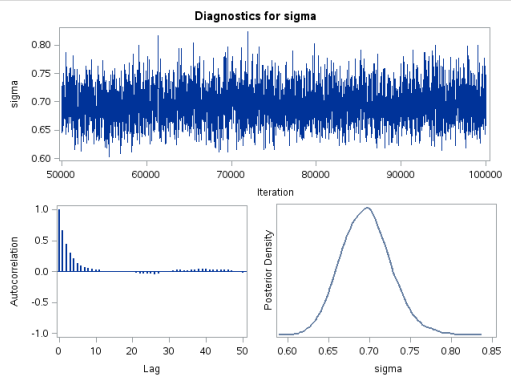
It is interesting to see from Table 4.2.6 that Model 2 has the lowest DIC but Model 1 actually has a sightly lower -2LL, although both statistics are very close between the two models. Since Model 1 takes a lot longer to run, Model 1 is con-

Figure 4.2.2: 6-Month Payment Bayesian Model 2 Diagnostics Graphs I

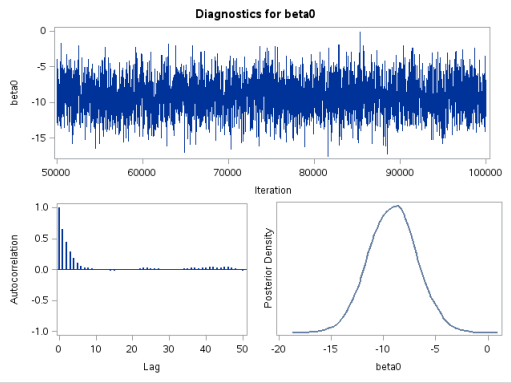
(a) Delta



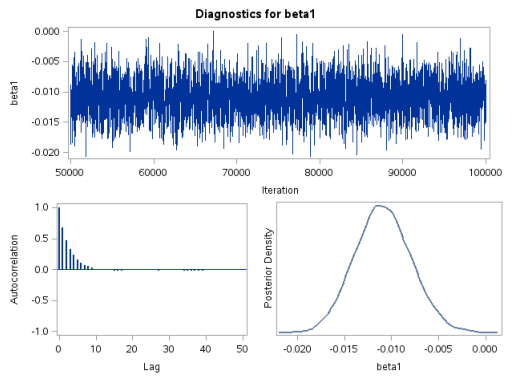
(b) Sigma



(c) Intercept



(d) Age



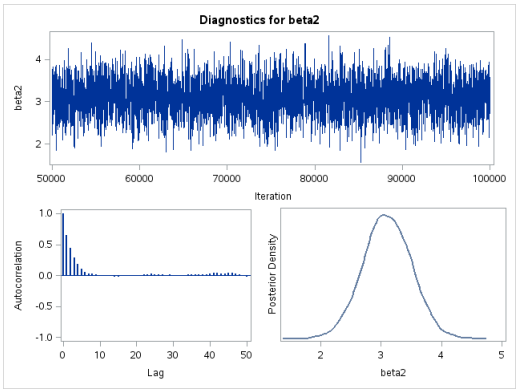
sidered the best Bayesian model here. However, the MLE model from Method 1 runs massively faster than even the fastest MCMC simulation, Model 5.

The diagnostic graphs in Figures 4.2.2 and 4.2.3 show the posterior sampling distribution (p.s.d.) of delta is a bit skewed and non-normal looking, while the p.s.d. for other parameters all resemble normal. The sampling runs look stable and autocorrelations drop off quickly.

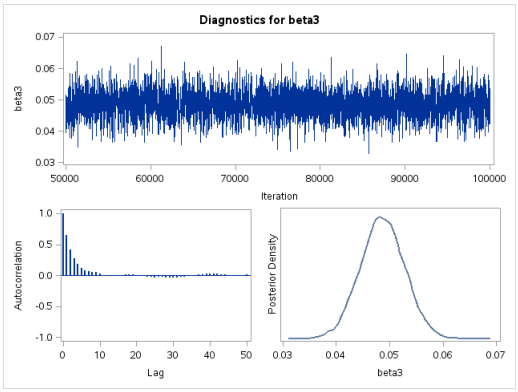
Table 4.2.7 contains details that support Figure 4.2.2. Simulations for all param-

Figure 4.2.3: 6-Month Duration Bayesian Model 2 Diagnostics Graphs II

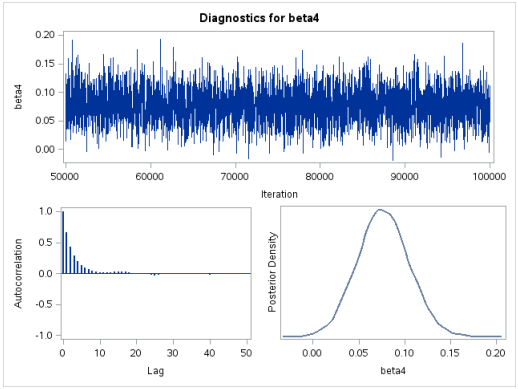
(a) Xb_Dur



(b) Cum_Pay



(c) Cum_Asth



eters are satisfactory. It is worth noting that Model 2 for M_6 has exactly the same set-up as Model 5 for A_6 in Section 4.2.1 but the M_6's simulation results are not as strong as A_6's.

The posterior parameter estimates in Table 4.2.8 are very close to their corresponding MLEs in Table 3.2.11. This is expected since we know from Table 4.2.5 that Model 2 uses non-informative priors and does not start at MLEs (see 4.1.5 on page 43). This indicates the posterior inference is not sensitive to the priors and mostly influenced by the data itself.

Considering the similarity in parameter estimates and goodness of fit but the superior speed in fitting the MLE model as compared to the Bayesian model, the MLE model is recommended for M_6.

4.3 Pros and Cons

The intention of Method 2 is to enhance Method 1 models with historical information. However, the best Bayesian models use non-informative priors and end up with posterior parameter estimates that are very close to their corresponding MLEs anyway for both A_6 and M_6. This result could be due to the potential lack-of-memory property of the claim adjustment process. That is, the most important information in predicting the ultimate payment at Month 6 is data available as of Month 6. What has happened in the past is no longer relevant.

Similar to method 1, this method circumvents the issues of correlated observations and time-dependent predictors. Using historical priors effectively augments the sample size at each t. However, it still treats t as a discrete variable and

Table 4.2.7: 6-Month Payment Bayesian Model 2 Diagnostics

(a) Monte Carlo Standard Errors

Parameter	MCSE	Standard Deviation	MCSE/SD
delta	0.0021	0.0957	0.0219
sigma	0.000679	0.0302	0.0225
beta0	0.0494	2.3109	0.0214
beta1	0.000068	0.00299	0.0229
beta2	0.00854	0.3997	0.0214
beta3	0.000091	0.00415	0.022
beta4	0.000642	0.0281	0.0228

(b) Posterior Autocorrelations

Parameter	Lag 1	Lag 5	Lag 10	Lag 50
delta	0.6601	0.1207	0.0076	0.0155
sigma	0.6617	0.1339	0.0306	-0.0193
beta0	0.653	0.1007	0.0023	-0.0261
beta1	0.6818	0.1543	0.004	0.0018
beta2	0.6538	0.1006	0.0013	-0.0262
beta3	0.6494	0.1231	0.0253	0.0189
beta4	0.6608	0.1321	0.0183	0.0062

(c) Effective Sample Sizes

Parameter	ESS	Autocorrelation Time	Efficiency
delta	2086.4	4.7931	0.2086
sigma	1976.9	5.0585	0.1977
beta0	2192.1	4.5619	0.2192
beta1	1909.3	5.2376	0.1909
beta2	2190.7	4.5648	0.2191
beta3	2057.1	4.8611	0.2057
beta4	1916.2	5.2187	0.1916

Table 4.2.8: 6-Month Payment Bayesian Model 2 Output

(a) Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
delta	10000	0.1786	0.0957	0.1074	0.1718	0.2416
sigma	10000	0.6954	0.0302	0.6743	0.6947	0.7149
beta0	10000	-9.1103	2.3109	-10.6793	-9.0965	-7.5681
beta1	10000	-0.0109	0.00299	-0.0129	-0.0109	-0.00891
beta2	10000	3.1018	0.3997	2.8326	3.0984	3.3739
beta3	10000	0.0487	0.00415	0.0459	0.0486	0.0514
beta4	10000	0.0768	0.0281	0.058	0.0764	0.0953

(b) Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
delta	0.05	0.018	0.3826	0.000159	0.3461
sigma	0.05	0.6394	0.7584	0.6372	0.7552
beta0	0.05	-13.6401	-4.5871	-13.5292	-4.5063
beta1	0.05	-0.0166	-0.00493	-0.0167	-0.00506
beta2	0.05	2.3229	3.8789	2.3218	3.875
beta3	0.05	0.0407	0.0568	0.0408	0.0569
beta4	0.05	0.0227	0.1322	0.0222	0.131

could lead to building a large number of models. Due to the Bayesian simulation, it tends be even more computationally intensive than method 1, as the time the MCMC simulation consumes far exceeds their MLE counterparts (see Tables 4.2.1 and 4.2.5).

Chapter 5

Method 3: Cumulative Longitudinal Model

This chapter explores ways to handle the three challenges (see Chapter 1, page 2) in a single model versus a series of models in Methods 1 and 2, which means all observations of each claimant are included in the same modelling sample.

5.1 Methodology

5.1.1 CLM

In my proposed Cumulative Longitudinal Model (CLM), the dependent variable is y_{it}, the cumulative payment for claimant i as of t. Right censoring in GLM based on the gamma distribution and the log link is again handled using the survival function method (Section 2.1.2) and the nested observations are handled using random effects. Specifically, we have

- Level 1: $y_{it} \sim \Gamma(\alpha, \theta_{it})$, $\log(\theta_{it}) = \mathbf{x}'_{it}\bar{\boldsymbol{\beta}} + \mathbf{w}'_{it}\boldsymbol{\zeta}_i$
- Level 2: $\boldsymbol{\zeta_i} \sim \text{Normal}(0, \boldsymbol{\Psi}) \Rightarrow \log(\theta_{it}) \sim \text{Normal}\left(\mathbf{x}'_{it}\bar{\boldsymbol{\beta}}, \mathbf{w}'_{it}\boldsymbol{\Psi}\mathbf{w}_{it}\right)$. So θ_{it} has a log-normal distribution.

where

- $\mathbf{x}'_{it}$ represents fixed effects with coefficients $\bar{\boldsymbol{\beta}}$.
- $\mathbf{w}'_{it}$ represents random effects with coefficients $\boldsymbol{\zeta_i}$, which reflects the nesting of observations within groups, or claimants.
- ζ_{ik} and $\zeta_{ik'}$ are assumed to be independent. So $\boldsymbol{\Psi}$ is a diagonal matrix.

Note $\mathbf{x}'_{it}$ usually includes $\mathbf{w}'_{it}$ with both time-dependent and non-time dependent covariates. Similar to (4.1.7), the conditional p.d.f. of y_{it} is

$$\begin{aligned} f(y_{it}|\theta_{it}, \alpha) &= f\left(y_{it}|\mathbf{x}'_{it}, \bar{\boldsymbol{\beta}}, \mathbf{w}'_{it}, \boldsymbol{\zeta}_i, \alpha\right) \\ &= f\left(y_{it}|\mathbf{x}'_{it}, \bar{\boldsymbol{\beta}}, \boldsymbol{\zeta}_i, \alpha\right) \\ &= \frac{1}{\Gamma(\alpha) e^{\left(\mathbf{x}'_{it}\bar{\boldsymbol{\beta}} + \mathbf{w}'_{it}\boldsymbol{\zeta}_i\right)\cdot\alpha}} y_{it}^{\alpha-1} \exp\left(\frac{-y_{it}}{e^{\mathbf{x}'_{it}\bar{\boldsymbol{\beta}} + \mathbf{w}'_{it}\boldsymbol{\zeta}_i}}\right) \end{aligned} \tag{5.1.1}$$

Let $\boldsymbol{Y}_i = (y_{i1}, y_{i2}, \cdots, y_{it}, \cdots)'$, $\boldsymbol{X_i} = (\mathbf{x_{i1}}, \mathbf{x_{i2}}, \cdots, \mathbf{x_{it}}, \cdots)'$, $\boldsymbol{W_i} = (\mathbf{w_{i1}}, \mathbf{w_{i2}}, \cdots, \mathbf{w_{it}}, \cdots)'$, and $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \cdots, \theta_{it}, \cdots)'$. The joint distribution of $\boldsymbol{Y_i}$ is developed as follows.

$$\begin{aligned} f_{\mathbf{I}}(\mathbf{Y}_i|\boldsymbol{\theta}_i, \alpha) &= f_{\mathbf{I}}\left(y_{i1}, y_{i2}, \cdots, y_{it}, \cdots|\boldsymbol{X_i}, \bar{\boldsymbol{\beta}}, \boldsymbol{W_i}, \boldsymbol{\zeta}_i, \alpha\right) \\ &= f_{\mathbf{I}}\left(y_{i1}, y_{i2}, \cdots, y_{it}, \cdots|\boldsymbol{X_i}, \bar{\boldsymbol{\beta}}, \boldsymbol{\zeta}_i, \alpha\right) \\ &= f_{i1}\left(y_{i1}\right) f^*_{i2}\left(y_{i2}|y_{i1}\right) f^*_{i3}\left(y_{i3}|y_{i2}, y_{i1}\right) \cdots \end{aligned} \tag{5.1.2}$$

As discussed in Section 2.2.2 on page 23, we cannot easily assume the nested observations are multivariate normal or conditionally independent. I propose the following method to generate the joint likelihood.

If we assume that the conditional distribution of y_{it} given $y_{i1}, y_{i2}, \cdots, y_{i,t-1}$ depends only on the previous measurement $y_{i,t-1}$, then (5.1.2) can be simplified as

$$f_{\mathbf{I}}\left(\mathbf{Y}_i|\boldsymbol{X_i}, \bar{\boldsymbol{\beta}}, \boldsymbol{\zeta}_i, \alpha\right) = f_{i1}\left(y_{i1}\right) f_{i2}^{*}\left(y_{i2}|y_{i1}\right) f_{i3}^{*}\left(y_{i3}|y_{i2}\right) \cdots f_{it}^{*}\left(y_{it}|y_{i,t-1}\right) \cdots \tag{5.1.3}$$

This assumption means that the distribution of the cumulative payment at t depends on the cumulative payment as of the previous time $t-1$, but how the cumulative payment has reached that level in even earlier time periods no longer matters. Further, due to the cumulative nature of y_{it}, we know $y_{it} \geqslant y_{i,t-1}$. If we assume the density of y_{it} depends on $y_{i,t-1}$ only through the value of $y_{i,t-1}$, then

$$\begin{aligned} f_{it}^{*}\left(Y_{it} = y_{it}|Y_{i,t-1} = y_{i,t-1}\right) &= f_{it}^{*}\left(Y_{it} = y_{it}|Y_{i,t} \geqslant y_{i,t-1}\right) \\ &= \frac{f_{it}\left(Y_{it} = y_{it}\right)}{P\left(Y_{i,t} \geqslant y_{i,t-1}\right)} \\ &= \frac{f_{it}\left(y_{it}\right)}{S_{it}\left(y_{i,t-1}\right)} \end{aligned} \tag{5.1.4}$$

where $S_{it}\left(\cdot\right)$ is the survival function. So the effect of $y_{i,t-1}$ on y_{it} is to truncate the latter's domain. Therefore,

$$f_{\mathbf{I}}\left(\mathbf{Y}_i|\boldsymbol{X_i}, \bar{\boldsymbol{\beta}}, \boldsymbol{\zeta}_i, \alpha\right) = f_{i1}\left(y_{i1}\right) \frac{f_{i2}\left(y_{i2}\right)}{S_{i2}\left(y_{i1}\right)} \frac{f_{i3}\left(y_{i3}\right)}{S_{i3}\left(y_{i2}\right)} \cdots \frac{f_{it}\left(y_{it}\right)}{S_{it}\left(y_{i,t-1}\right)} \cdots \tag{5.1.5}$$

Similar to Equation (4.1.7), it can be shown $S_{it}\left(y_{it}\right) = \frac{1}{\Gamma(\alpha)}\Gamma\left(\alpha, y_{it}/e^{\mathbf{x}_{it}'\bar{\boldsymbol{\beta}}+\mathbf{w}_{it}'\boldsymbol{\zeta}_i}\right)$.

To obtain the likelihood of the entire sample and because $\boldsymbol{\zeta}_i$ is an unobserved random variable, we first integrate $\boldsymbol{\zeta}_i$ out of the joint density $f_{\mathbf{I}}\left(\mathbf{Y}_i|\boldsymbol{X_i}, \bar{\boldsymbol{\beta}}, \boldsymbol{\zeta}_i, \alpha\right)$ to get the marginal density of $\boldsymbol{Y}_i$.

$$\int Q_i g\left(\boldsymbol{\zeta}_i|\boldsymbol{\Psi}\right) d\boldsymbol{\zeta}_i \tag{5.1.6}$$

where Q_i equals (5.1.5). Let $\boldsymbol{Y} = \left(\boldsymbol{Y}'_1, \boldsymbol{Y}'_2, \cdots, \boldsymbol{Y}'_k\right)'$ represent all claimants in the sample, the joint likelihood of $\boldsymbol{Y}$ is

$$\prod_{i=1}^{k} \int Q_i g\left(\boldsymbol{\zeta}_i|\boldsymbol{\Psi}\right) d\boldsymbol{\zeta}_i \tag{5.1.7}$$

Above is the basic set-up of the cumulative longitudinal model, which combines techniques of survival analysis and the first-order Markov chain to model grouped observations with a cumulative continuous dependent variable.

To adapt CLM for modelling the ultimate payment, one idea is to adjust the likelihood to reflect that the ultimate payment y_{ic_i} is always $\geq y_{it}$ at any t.

So if the ultimate payment is observed, the joint likelihood of all measurements of claimant i is

$$\begin{aligned}
&f_{\mathbf{I}}\left(Y_{i1} \geqslant y_{i1}, Y_{i2} \geqslant y_{i2}, \cdots, Y_{ic_{i-1}} \geqslant y_{ic_{i-1}}, Y_{ic_i} = y_{ic_i}|\boldsymbol{X_i}, \bar{\boldsymbol{\beta}}, \boldsymbol{\zeta}_i, \alpha\right) \\
&= \int \cdots \int_{y_{ic_{i-1}} \cdots y_{i1}}^{\infty \cdots \infty} f_{\mathbf{I}}\left(\mathbf{Y}_i|\boldsymbol{X_i}, \bar{\boldsymbol{\beta}}, \boldsymbol{\zeta}_i, \alpha\right) dY_{i1} \cdots dY_{ic_{i-1}} \cdot \frac{f_{ic_i}\left(y_{ic_i}\right)}{S_{ic_i}\left(y_{ic_{i-1}}\right)} \\
&= \int_{y_{i1}}^{\infty} f_{i1}\left(Y_{i1}\right) dY_{i1} \int_{y_{i2}}^{\infty} \frac{f_{i2}\left(Y_{i2}\right)}{S_{i2}\left(y_{i1}\right)} dY_{i2} \cdots \int_{y_{ic_{i-1}}}^{\infty} \frac{f_{ic_{i-1}}\left(Y_{ic_{i-1}}\right)}{S_{ic_{i-1}}\left(y_{ic_{i-2}}\right)} dY_{ic_{i-1}} \cdot \frac{f_{ic_i}\left(y_{ic_i}\right)}{S_{ic_i}\left(y_{ic_{i-1}}\right)} \\
&= S_{i1}\left(y_{i1}\right) \frac{S_{i2}\left(y_{i2}\right)}{S_{i2}\left(y_{i1}\right)} \cdots \frac{S_{ic_{i-1}}\left(y_{ic_{i-1}}\right)}{S_{ic_{i-1}}\left(y_{ic_{i-2}}\right)} \frac{f_{ic_i}\left(y_{ic_i}\right)}{S_{ic_i}\left(y_{ic_{i-1}}\right)}
\end{aligned} \tag{5.1.8}$$

If the claim is still open when the last cumulative payment is observed, then the joint likelihood is

$$\begin{aligned} & f_{\mathbf{I}}\left(Y_{i1} \geqslant y_{i1}, Y_{i2} \geqslant y_{i2}, \cdots, Y_{ic_{i-1}} \geqslant y_{ic_{i-1}}, Y_{ic_i} \geqslant y_{ic_i} | \boldsymbol{X_i}, \bar{\boldsymbol{\beta}}, \boldsymbol{\zeta}_i, \alpha\right) \\ & = S_{i1}\left(y_{i1}\right) \frac{S_{i2}\left(y_{i2}\right)}{S_{i2}\left(y_{i1}\right)} \cdots \frac{S_{ic_{i-1}}\left(y_{ic_{i-1}}\right)}{S_{ic_{i-1}}\left(y_{ic_{i-2}}\right)} \frac{S_{ic_i}\left(y_{ic_i}\right)}{S_{ic_i}\left(y_{ic_{i-1}}\right)} \end{aligned} \tag{5.1.9}$$

However, this modification creates a lot of artificially censored observations. In fact, every measurement of a claim at t before the last, t_+, is considered censored, even if the ultimate payment is observed. Also, since our objective is to predict the ultimate payment, which is included in the same sample as either the complete y_{ic_i} or the censored y_{i,t_i^+}, to only recognize that the ultimate total payment is $\geqslant y_{it}$ but not what it actually is seems to be not using all information.

An alternative adaptation attempts to remedy the above concerns above by directly using the last observed cumulative payment, y_{ic_i} or y_{i,t_i^+} as the dependent variable at each t. Specifically, let $\boldsymbol{Y}_i = (y_{ic_i}, y_{ic_i}, \cdots, y_{ic_i}, \cdots)'$, $\tilde{\boldsymbol{X}}_{\boldsymbol{i}} = (\tilde{\mathbf{x}}_{\mathbf{i1}}, \tilde{\mathbf{x}}_{\mathbf{i2}}, \cdots, \tilde{\mathbf{x}}_{\mathbf{it}}, \cdots)'$, where $\tilde{\mathbf{x}}_{it}' = (\mathbf{x}_{it}', y_{it})$, and $\boldsymbol{W_i}$, $\boldsymbol{\theta}_i$ are defined the same as in (5.1.2). Then (5.1.5) is modified to be

$$f_{\mathbf{I}}\left(\mathbf{Y}_i | \boldsymbol{X_i}, \bar{\boldsymbol{\beta}}, \boldsymbol{\zeta}_i, \alpha\right) = \frac{f_{i1}\left(y_{ic_i}\right)}{S_{i1}\left(y_{i1}\right)} \frac{f_{i2}\left(y_{ic_i}\right)}{S_{i2}\left(y_{i2}\right)} \frac{f_{i3}\left(y_{ic_i}\right)}{S_{i3}\left(y_{i3}\right)} \cdots \frac{f_{it}\left(y_{ic_i}\right)}{S_{it}\left(y_{it}\right)} \cdots \tag{5.1.10}$$

Censoring and integration of random effects would proceed similarly as in steps (5.1.8) to (5.1.7). However, the MCMC simulation based on Equation (5.1.10) is not satisfactory with strong autocorrelations remaining even after 500 lags. This is mainly due to the target variable being fixed over time.

A third variation is an indirect approach, which models the remaining payment at t, say $\bar{y}_{it}$, instead of the ultimate total payment directly. We have $\bar{y}_{it} = y_{ic_i} - y_{it}$ for non-censored observations and $\bar{y}_{it} = y_{i,t_i^+} - y_{it}$ for censored ones. Then $\widehat{y_{ic_i}} = \widehat{\bar{y}_{it}} + y_{it}$. Equation (5.1.5) is updated to

$$f_{\mathbf{I}}\left(\mathbf{Y}_i|\boldsymbol{X_i}, \bar{\boldsymbol{\beta}}, \boldsymbol{\zeta}_i, \alpha\right) = f_{i1}\left(\bar{y}_{i1}\right) \frac{f_{i2}\left(\bar{y}_{i2}\right)}{S_{i2}\left(y_{i1}\right)} \frac{f_{i3}\left(\bar{y}_{i3}\right)}{S_{i3}\left(y_{i2}\right)} \cdots \frac{f_{it}\left(\bar{y}_{it}\right)}{S_{it}\left(y_{i,t-1}\right)} \cdots \tag{5.1.11}$$

for un-censored observations and

$$f_{\mathbf{I}}\left(\mathbf{Y}_i|\boldsymbol{X_i}, \bar{\boldsymbol{\beta}}, \boldsymbol{\zeta}_i, \alpha\right) = S_{i1}\left(\bar{y}_{i1}\right) \frac{S_{i2}\left(\bar{y}_{i2}\right)}{S_{i2}\left(y_{i1}\right)} \frac{S_{i3}\left(\bar{y}_{i3}\right)}{S_{i3}\left(y_{i2}\right)} \cdots \frac{S_{it}\left(\bar{y}_{it}\right)}{S_{it}\left(y_{i,t-1}\right)} \cdots \tag{5.1.12}$$

for censored ones, where $\bar{y}_{it} \sim \Gamma(\alpha, \bar{\theta}_{it})$, $\log(\bar{\theta}_{it}) = \mathbf{x}'_{it}\bar{\boldsymbol{\beta}} + \mathbf{w}'_{it}\boldsymbol{\zeta}_i$, and y_{it} still stands for the cumulative payment as of t. The p.d.f. of $\bar{y}_{it}$ is

$$\begin{aligned} f(\bar{y}_{it}|\theta_{it}, \alpha) &= \frac{1}{\Gamma(\alpha) e^{\left(\mathbf{x}'_{it}\bar{\boldsymbol{\beta}} + \mathbf{w}'_{it}\boldsymbol{\zeta}_i\right)\cdot\alpha}} (\bar{y}_{it})^{\alpha-1} \exp\left(\frac{-\bar{y}_{it}}{e^{\mathbf{x}'_{it}\bar{\boldsymbol{\beta}} + \mathbf{w}'_{it}\boldsymbol{\zeta}_i}}\right) \\ &= \frac{1}{\Gamma(\alpha) e^{\left(\mathbf{x}'_{it}\bar{\boldsymbol{\beta}} + \mathbf{w}'_{it}\boldsymbol{\zeta}_i\right)\cdot\alpha}} (y_{ic_i} - y_{it})^{\alpha-1} \exp\left(-\frac{y_{ic_i} - y_{it}}{e^{\mathbf{x}'_{it}\bar{\boldsymbol{\beta}} + \mathbf{w}'_{it}\boldsymbol{\zeta}_i}}\right) \\ &= f(y_{ic_i}|y_{it}, \theta_{it}, \alpha). \end{aligned} \tag{5.1.13}$$

So Equation (5.1.13) is equivalent to assuming the distribution y_{ic_i} is gamma with a location parameter, y_{it}. The corresponding MCMC results are much improved.

5.1.2 GLMM

From Chapter 4, we know the claim adjustment process may possess the lack-of-memory property (see Section 4.3 on page 57), which implies the remaining payment at t, $\bar{y}_{it}$, could be independent of its value at the previous period, $\bar{y}_{it-1}$, conditional on the group effect. Hence, Equation 5.1.11 can be simplified as

$$f_{\mathbf{I}}\left(\mathbf{Y}_i|\boldsymbol{X_i},\bar{\boldsymbol{\beta}},\boldsymbol{\zeta}_i,\alpha\right)=f_{i1}\left(\bar{y}_{i1}\right)f_{i2}\left(\bar{y}_{i2}\right)f_{i3}\left(\bar{y}_{i3}\right)\cdots f_{it}\left(\bar{y}_{it}\right)\cdots \tag{5.1.14}$$

which is exactly the expression for the joint likelihood of grouped observations in GLMM when there is no censoring. When there is right censoring, we update Equation 5.1.12 as follows

$$f_{\mathbf{I}}\left(\mathbf{Y}_i|\boldsymbol{X_i},\bar{\boldsymbol{\beta}},\boldsymbol{\zeta}_i,\alpha\right)=S_{i1}\left(\bar{y}_{i1}\right)S_{i2}\left(\bar{y}_{i2}\right)S_{i3}\left(\bar{y}_{i3}\right)\cdots S_{it}\left(\bar{y}_{it}\right)\cdots \tag{5.1.15}$$

The rest of the steps remain the same as Equations (5.1.6) and (5.1.7). The results presented in the next section are based on the remaining payment being the target variable. For comparison, results for both CLM and the conditional independence-based GLMM are included.

5.2 Data Results

The CLM method described from (5.1.2) to (5.1.7) is not restricted to particular distributions. It can be shown when y_{it} has an exponential distribution, $\boldsymbol{\zeta_i}$ contains only one random intercept ζ_0, and e^{ζ_0} has an inverse gamma distribution, (5.1.6) can be expressed in analytic form. Then maximum likelihood estimation may pro-

ceed as usual. More research is needed on what combinations of distributions for $\boldsymbol{Y}_i$ and $\boldsymbol{\zeta_i}$ will lead to closed-form marginal likelihood (ML) for $\boldsymbol{Y}_i$.

In general, however, (5.1.7) cannot be computed analytically and needs to be approximated. SAS PROC GLIMMIX [3] uses Laplace approximation or adaptive quadrature to approximate (5.1.7), but they require conditional independence among $y_{it}|\boldsymbol{\zeta_i}$ and only accomodates normally-distributed random effects. PROC NLMIXED [5] allows the user to program the log-likelihood in addition to the prebuilt distributions. It also uses adaptive Gauss-Hermite quadrature to approximate the ML of $\boldsymbol{Y}_i$ and requires normal random effects. PROC GLIMMIX does not appear to handle censoring while PROC NLMIXED only accepts log-likelihood that can be expressed analytically, which is challenging to do for the survival function of the gamma distribution. More research is needed.

There are alternatives to likelihood-based approaches. Generalized Estimation Equation (GEE) does not evaluate (5.1.7), although it does involve some integration as it needs to estimate the first two moments of $\boldsymbol{Y}_i|\boldsymbol{\zeta_i}$. As is in [20, Zeger, Liang], it does not handle censoring.

Hierarchical GLM or HGLM [9, Lee, Nelder] also does not evaluate the ML of $\boldsymbol{Y}_i$. But it implies conditional independence and does not handle censoring. Whether GEE and HGLM can be modified to handle CLM needs more research.

Note conditional independence, normal random effects and no censoring are all special cases of CLM. For example, changing $S_{it}\left(y_{i,t-1}\right)$ in (5.1.4) to $S_{it}\left(0\right)$ for all t means $y_{it} \perp y_{it'}|\boldsymbol{\zeta_i}$.

As seen in Section 4.1.1, when the sample size is large enough, the asymptotic posterior distribution of $\bar{\boldsymbol{\beta}}$ approaches the normal distribution censored on its MLE. So MCMC is again used here.

5.2.1 CLM

Table 5.2.1 shows the set-up of select MCMC stimulations of the CLM variation based on Equations (5.1.11) and (5.1.12) on page 66. Note the running time is measured in minutes here.

Table 5.2.1: CLM MCMC Set-up

Model	Tuning	Burn-In	MCMC	Thin	Predictors	Prior
1D	50,000	150,000	200,000	20	Age, Cum_Pay3	Non-Informative
1C	50,000	150,000	200,000	20	Age, Cum_Trmt	Non-Informative
1B	10,000	50,000	50,000	5	Age, Cum_Trmt	Var$(\beta_j) = 1000$

Note: *(a)* For 1B, a smaller number is chosen for Var(β_j) to help the simulation converge faster and therefore not non-informative; *(b)* Cum_Pay3 is a scaled version of Cum_Pay. The objective is to make its coefficient estimate more readable.

Table 5.2.2 shows model 1D has the lowest DIC but not the longest running time, which is desirable. For more details on Model 1D, see Appendix D Section D.1.

Table 5.2.2: CLM MCMC Comparison

Rankings In DIC	Model	CPU Time	Real Time	DIC
1	1D	6:09.26	6:11.41	2821.16
2	1C	7:46.09	7:48.45	2899.772
3	1B	2:14	2:16.39	2902.629

5.2.2 GLMM

To recap, this approach assumes conditional independence among nested observations and the likelihood is based on Equations (5.1.14) and (5.1.15). Table 5.2.3 displays the set-up of relevant select MCMC simulations. Again, time is measured in minutes here.

Table 5.2.3: GLMM MCMC Set-up

Model	Tuning	Burn-In	MCMC	Thin	Predictors
2B	10,000	50,000	50,000	5	Age, Cum_Trmt
2C	50,000	150,000	200,000	20	Age, Cum_Trmt
2D	50,000	150,000	200,000	20	Age, Cum_Pay3
2E	50,000	200,000	500,000	50	Age, Cum_Pay3

Note: *(a)* All models have the same priors, which are included in the SAS code in Appendix D on page 110; *(b)* Cum_Pay3 is a scaled version of Cum_Pay. The objective is to make its coefficient estimate more readable.

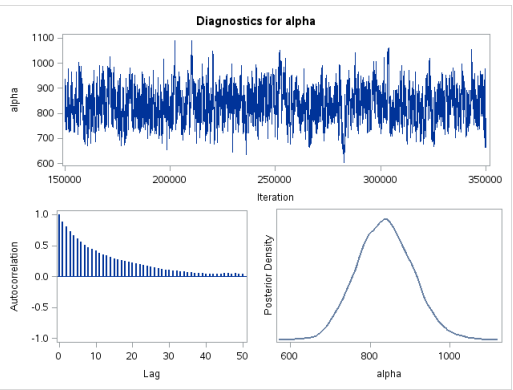
In comparison, Table 5.2.4 shows 2E has the lowest DIC but also took the longest time to run. For more details on Model 2E, see Appendix D Section D.2.

Table 5.2.4: GLMM MCMC Comparison

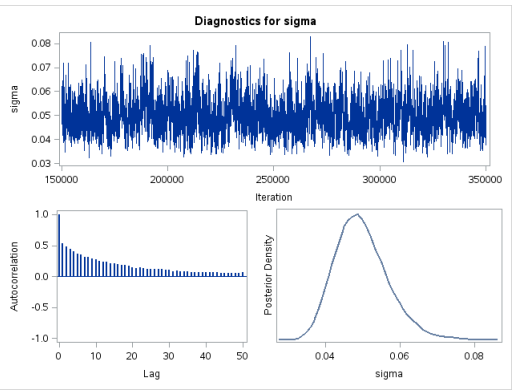
Rankings In DIC	Model	CPU Time	Real Time	DIC
1	2B	1:12.05	1:12.69	2792.402
2	2C	4:33.16	4:34.29	2789.696
3	2D	11:28.54	11:31.05	2736.559
4	2E	15:59.09	16:02.62	2735.557

Figure 5.2.1: CLM 1D MCMC Diagnostics Graphs I

(a) Alpha



(b) Sigma



Note: *(a)* Alpha is the common gamma shape parameter; *(b)* Sigma is the common standard deviation of the random effects.

5.2.3 Comparison

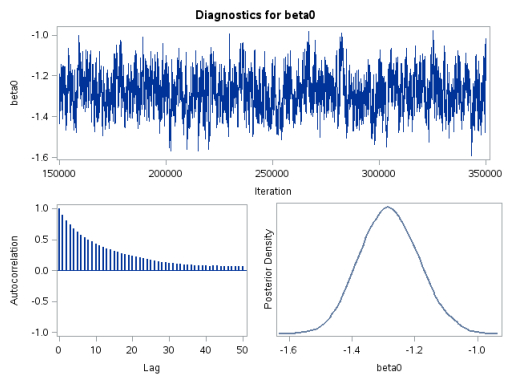
Since there are $1,339$ distinct claimants or groups with $26,285$ observations in the entire data set, MCMC runs using all the data have proven too time consuming and had to be terminated after running for hours. Two smaller samples are used in the exploration in previous sections separately. Below are the results of the best model from each section using the same sample, which has 20 groups with 220 observations.

Figures 5.2.1 and 5.2.2 show that the sampling posterior distributions resemble the normal distributions and the trace plots are generally stable. But the autocorrelations decline slowly even up to 50 lags.

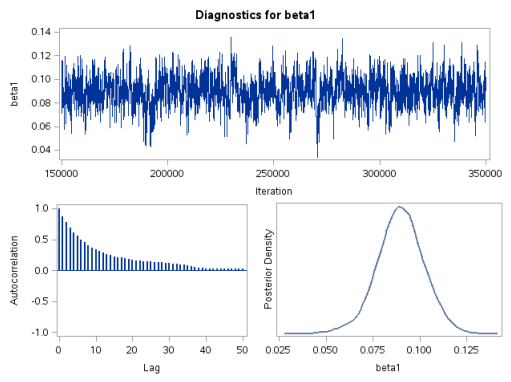
Table 5.2.5 show the MCSE/RD ratios are small but larger than in Chapter 4. Autocorrelations are definitely declining as lags increase but still sizeable at Lage10, though small at Lag 50. The effective sample size (ESS) and the efficiency

Figure 5.2.2: CLM 1D MCMC Diagnostics Graphs II

(a) Intercept



(b) Age



(c) Cum_Pay

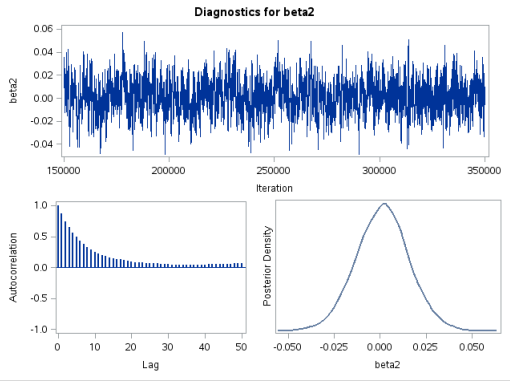
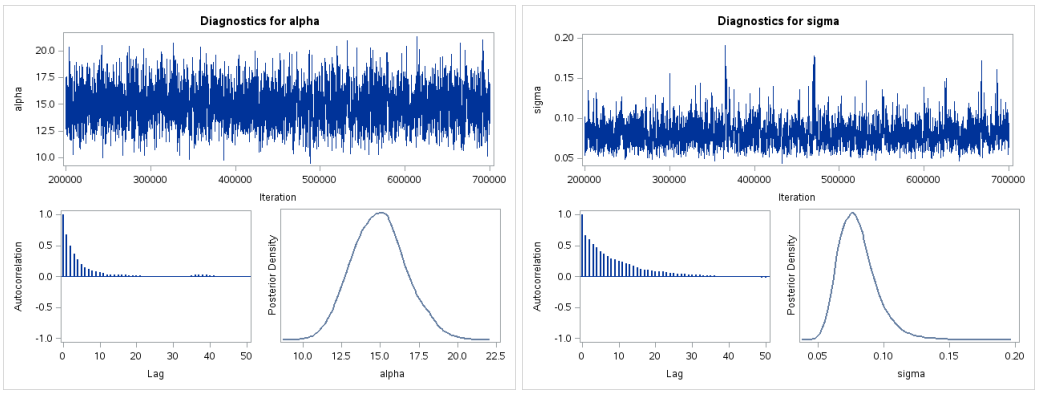


Figure 5.2.3: GLMM 2E MCMC Diagnostics Graphs I

(a) Alpha (b) Sigma



Note: *(a)* Alpha is the common gamma shape parameter; *(b)* Sigma is the common standard deviation of the random effects.

are low.

Table 5.2.6 (b) shows the HPD interval for Cum_Pay3 covers 0, suggesting the variable is not significant.

Next let us examine the results of GLMM Model 2E on the same sample. Diagnostics in Figures 5.2.3 and 5.2.4 are much better than those for CLM 1D with very stable trace plots, rapidly dropping autocorrelations and normal-resembling posterior sampling distributions.

The diagnostic details in Table 5.2.7 are generally good as well.

In Table 5.2.8 both predictors are significant with their HPD intervals not including 0 at the $\alpha = 0.05$ level. Also, the coefficient estimate for Cum_Pay3 is negative as expected.

Considering the MCMC simulation diagnostic results are better for GLMM 2E than CLM 1D and the fact that coefficient estimates in 2E have expected signs, 2E

Table 5.2.5: CLM MCMC Diagnostics

(a) Monte Carlo Standard Errors

Parameter	MCSE	Standard Deviation	MCSE/SD
beta0	0.00514	0.0935	0.055
beta1	0.000616	0.0125	0.0492
beta2	0.000637	0.0144	0.0444
alpha	3.6378	69.6058	0.0523
sigma	0.000313	0.00702	0.0446

(b) Posterior Autocorrelations

Parameter	Lag 1	Lag 5	Lag 10	Lag 50
beta0	0.8927	0.6254	0.4312	0.0627
beta1	0.8763	0.5534	0.3374	0.0312
beta2	0.8645	0.492	0.2497	0.0668
alpha	0.8871	0.6063	0.4102	0.0457
sigma	0.5289	0.3616	0.2788	0.0612

(c) Effective Sample Sizes

Parameter	ESS	Autocorrelation Time	Efficiency
beta0	330.6	30.2443	0.0331
beta1	412.8	24.2253	0.0413
beta2	508.1	19.681	0.0508
alpha	366.1	27.3143	0.0366
sigma	502.2	19.911	0.0502

Table 5.2.6: CLM 1D MCMC Output

(a) Posterior Summaries

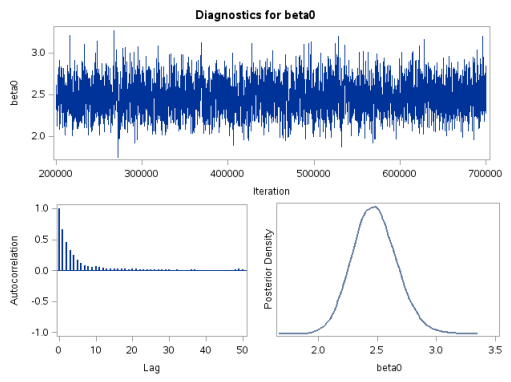
Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
beta0	10000	-1.2807	0.0935	-1.3453	-1.2822	-1.2168
beta1	10000	0.0897	0.0125	0.0818	0.0898	0.0977
beta2	10000	0.00131	0.0144	-0.00825	0.00149	0.0108
alpha	10000	837.1	69.6058	789.5	836.2	883.5
sigma	10000	0.0498	0.00702	0.0448	0.0491	0.054

(b) Posterior Intervals

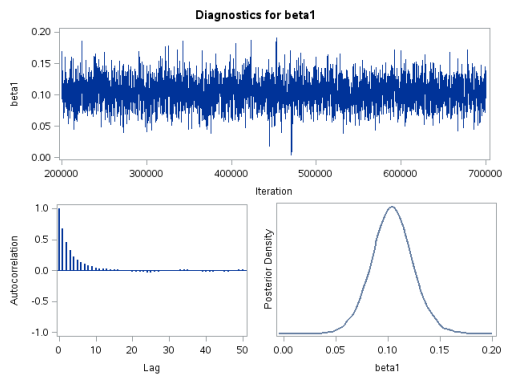
Parameter	Alpha	Equal-Tail Interval		HPD Interval	
beta0	0.05	-1.4595	-1.0934	-1.4605	-1.0953
beta1	0.05	0.0635	0.1138	0.066	0.1161
beta2	0.05	-0.0273	0.0297	-0.0269	0.0299
alpha	0.05	705.8	977.2	702.2	971.3
sigma	0.05	0.038	0.0652	0.037	0.0637

Figure 5.2.4: GLMM 2E MCMC Diagnostics Graphs II

(a) Intercept



(b) Age



(c) Cum_Pay

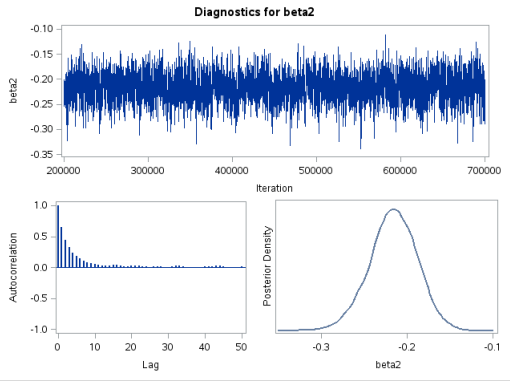


Table 5.2.7: GLMM 2E MCMC Diagnostics

(a) Monte Carlo Standard Errors

Parameter	MCSE	Standard Deviation	MCSE/SD
beta0	0.00468	0.1863	0.0251
beta1	0.000485	0.0201	0.0241
beta2	0.000709	0.0282	0.0252
alpha	0.0444	1.7347	0.0256
sigma	0.000549	0.015	0.0366

(b) Posterior Autocorrelations

Parameter	Lag 1	Lag 5	Lag 10	Lag 50
beta0	0.6565	0.1745	0.0598	0.016
beta1	0.6741	0.1714	0.0463	0.0151
beta2	0.6556	0.18	0.0564	0.0098
alpha	0.6756	0.1968	0.069	- 0.0099
sigma	0.663	0.4055	0.247	- 0.0247

(c) Effective Sample Sizes

Parameter	ESS	Autocorrelation Time	Efficiency
beta0	1582.7	6.3181	0.1583
beta1	1718.1	5.8205	0.1718
beta2	1580.4	6.3274	0.158
alpha	1527.6	6.546	0.1528
sigma	744.6	13.43	0.0745

Table 5.2.8: GLMM 2E MCMC Output

(a) Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
beta0	10000	2.4769	0.1863	2.35	2.4721	2.5986
beta1	10000	0.104	0.0201	0.091	0.1037	0.1169
beta2	10000	-0.2163	0.0282	-0.2341	-0.2157	-0.197
alpha	10000	14.9337	1.7347	13.7088	14.8893	16.0649
sigma	10000	0.0802	0.015	0.0698	0.0782	0.0881

(b) Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
beta0	0.05	2.13	2.8564	2.1191	2.8436
beta1	0.05	0.0639	0.1443	0.0657	0.1457
beta2	0.05	-0.2744	-0.1634	-0.2736	-0.1627
alpha	0.05	11.7183	18.4754	11.6405	18.3834
sigma	0.05	0.0577	0.1157	0.0555	0.1108

is preferred for this particular sample. However, this is a small sample and the results may be volatile. In fact, simulation for 1D on some other samples have generated significant coefficient estimates with the expected sign (see Appendix D Section D.1).

In order to generate more stable MCMC results, bootstrapping is suggested. Bootstrapping divides the entire sample into several smaller sub-samples, then runs MCMC simulations on each of them and obtains the sampling distributions of the posterior sampling statistics such as mean, percentiles and so on. That is, it generates the sample mean of the posterior sample mean.

5.3 Pros and Cons

The cumulative longitudinal model (CLM) is a special class of generalized linear mixed models (GLMM) where the dependent variable is a cumulative continuous variable. By proposing a new way of generating the joint likelihood for nested observations, the CLM expanded the GLMM, which typically assumes conditional independence for the aforementioned. Combined with the survival function method for handling right censoring, this method theoretically can handle all three challenges mentioned in Chapter 1 on page 2. Moreover, although my example is based on the gamma distribution, CLM is a general approach and can accommodate other distributions.

In practice, however, because the difficulty associated with integrating over random effects especially when the joint likelihood involves incomplete gamma functions, I have not yet found a good way to fit the maximum-likelihood (ML)

model. Although MCMC simulation can be used to approximate MLEs, it involves a huge amount of computing and takes a long time, mainly because of the large number of random effects or groups. Therefore, a trade-off must be made. We can either run simulations using a large sample which may strain the resources severely and even fail, or run a small sample each time and use bootstrapping to summarize the results and select variables, which may also take a long time. So the usefulness of the technique is restricted by available computing power.

The parameter estimates obtained in this method are global optima taking into account information over the entire claim cycle, while the estimates obtained in Methods 1 and 2 are local optima. So CLM/GLMM may help one learn more about the overall claim adjustment process. On the other hand, it is not as flexible as the other two methods as the latter may include different significant variables at different time.

Chapter 6

Summary

6.1 Recap

Both Methods 1 and 2 build a series of models at select time points of the claim cycle. This type of approach bypasses the issues of nested observations and predicting a fixed target with time-dependent covariates with independent samples and models at each t. More importantly, this approach is flexible allowing for different significant predictors at different time to make the best prediction possible given information available at a particular t.

Method 2 intends to improve Method 1 by integrating learning from previous time to the current period. It uses asymptotic posterior distributions from $t-1$ as the basis of prior distributions for t. In my data analysis, Method 2 generates similar parameter estimates as Method 1 does and thus does not improve the ML prediction from Method 1 much. However, at this point, whether this is due to my data on hand or a universal result needs more study.

In terms of limitations, the method of a series of models treats time as a discrete variable. Potentially, we may need to build a large number of models depending on the number of selected points to make predictions. It could be monthly, weekly, daily, hourly or even more frequent, placing a high demand on time and computing resources. In addition, the decreasing sample size over time may also increase the standard errors of estimates.

Method 3 introduces the cumulative longitudinal model (CLM) which attempts to tackle the three challenges (see page 2) in a single model. It uses random effects to represent the group effect and a joint likelihood resembling the first-order Markov chain. Rather than fits the ultimate total payment directly, it models the remaining payment equivalently. Method 3 has the advantage of simplicity with only one model to fit. It treats time as a continuous variable, meaning one model can make predictions at any time, not just pre-selected points. Finally, it seeks global optima that may help understand the overall claim adjustment process as compared to snapshots in Methods 1 and 2.

However, fitting the ML model using Method 2 proves difficult. As a result, MCMC simulation is deployed that takes a very long time to run and can only handle a small number of groups each time.

All methods use the survival function technique to handle right-censoring as in AFT.

6.2 Contributions

6.2.1 GLM with Right Censoring

Using the survival function method to handle right censoring in GLM, I introduced a likelihood-based approach to address this issue, which is an addition to the non-parametric methods such as the Buckley-James method or other Kaplan-Meier estimator based methods. This method is general. It applies to all distributions admissible to GLM and a continuous cumulative dependent variable.

In the case of a GLM with right censoring based on the gamma distribution and a log link, I proved the equivalence between GLM and AFT. I have also shown the relationship is not universal. That is not all distributions can be used in both GLM and AFT. And even if they can, the MLEs from GLM and AFT may not have the same relationship as it does in the case of gamma and a log link. They may not have a relationship at all.

6.2.2 Functions of Predictors

In contrast to many popular variable selection techniques such as forward, backward, stepwise, and best-subset selection which focuses on the significance of individual predictors, I have shown empirically in Example 3 that linear combinations of individually insignificant predictors could be significant itself and improve the goodness of fit of the model. This observation can be expanded to functions of predictors in general and broadened the scope of variable selection in models with a linear predictor.

6.2.3 CLM

My proposed cumulative longitudinal model (CLM) expanded the traditional GLMM which assumes conditional dependence among the nested observations. The CLM generates the joint likelihood for nested observations by assuming the dependent variable at t only depends on its value at $t-1$, analogous to a first-order Markov chain. This concept can also be expanded to situations where the dependent variable is not cumulative such as blood pressure over time.

6.3 Future Work

6.3.1 Non-Cumulative Target

Buckley-James, Kaplan-Meier, and the survival function method originated in survival analysis where the dependent variable is time. That is, they assume the censored variable's ultimate value is greater than or equal to the censored value. That is the reason they can be adapted to models with other cumulative dependent variables.

However, some variables are not cumulative over time. For example, if we are to predict the blood pressure at the time a patient is cured where right censoring means the patient exits the study before being cured, or the stock price when a company goes bankrupt where censoring means the company has not gone bankrupt when the study ends, none of the methods mentioned here would apply. How to handle censoring of non-cumulative dependent variables in GLM needs more research.

6.3.2 Other Distributions in GLM

GLM fits the mean while AFT fits the scale parameter with a linear predictor. When the GLM is based on the gamma distribution with a log link, the coefficient estimates for the linear predictor are equivalent to those in AFT. However, when the distribution is not gamma, whether or what relationship there is between GLM and AFT MLEs is not clear.

6.3.3 Variable Selection

It would be interesting to develop variable selection algorithms to not only evaluate individual predictors but also functions of them, starting with linear combinations of predictors.

6.3.4 CLM

Properties of CLM need more study. In my simulations, GLMM based on conditional independence actually outperformed CLM with lower autocorrelations. I would like to research if this is due to the peculiarity of the data or this is universally true, that is, GLMM is inherently superior to the CLM.

Another important area of research is to develop a way to fit ML models. This would involve approximating and programming the incomplete gamma function in analytic expressions. Also the MCMC simulation for CLM takes too long and could only handle a small number of observations at a time. How to improve the running time of the MCMC simulation is an important task.

6.4 Business Applications

Method 1 was implemented in a major Property and Casualty carrier in 2007. Based on the actuary's estimate, it generated significant financial benefits, which are detailed as follows.

1. More accurate estimation means less over reserve, which means the company could save interest expenses on the fund designated for reserve.

2. The saved fund could be used in investment to generate more income.

3. More accurate estimating also allows the company to underwrite a wider spectrum of risks knowing that the same amount of fund may now serve more claims, thus resulting in more written premium.

Although claim reserving is the motivation for developing my methods, my research has potential applications in many business areas. Here are some examples beyond claim adjustment.

- The total revenue a company receives from a customer by the time the customer relationship ends, for example, through cancellation or death. The customer could be an insurance policy holder, a cable TV company subscriber, a telecommunication company customer (e.g., cell phone), or a security brokerage firm client. This is an important component of customer lifetime value.

- Total loss or expenses incurred by a customer by the time the customer relationship terminates. This is the other component of customer lifetime value.

- Total medical expenses a long-term patient incurs by the time treatment ends either due to recovery or death. For example, a patient recovering from long-term disability or cancer treatment. This may be of interest to doctors, hospitals and insurance companies.

- Total expenditure (e.g., marketing, repair, loan interest, etc.) by the time a house is sold. This is of interest to both agents and sellers when the house is still in the market and decisions need to be made about further spending.

In all cases above, the interest lies in predicting the ultimate value of a cumulative target at the time a future event happens, and the prediction needs to be updated as new information comes in. When the event will happen is not certain but it does not depend on the target of interest. An example where the timing of the event does depend on the target of interest is total number of reported fire or explosion incidents when a cell phone is recalled. In this case, the recall is the direct result of the incidents. After the number of incidents reaches a certain level, a decision instead of prediction needs to be made about the recall.

When the target variable is not cumulative, Methods 1 and 2 will still apply but Method 3 won't. For example, predicting the savings or asset value by the time someone retires or dies for financial planning. Clearly savings might fluctuate over time.

6.4.1 Evaluation

Business evaluation of a procedure does not always align with the statistical one. In likelihood-based models, likelihood-based criteria such as Akaike Infor-

mation Criterion (AIC), Bayesian Information Criterion (BIC), and Deviance Information Criteria (DIC) are commonly used to assess the goodness of fit.

Yet, business customers may be more interested in knowing how close the predictions are to the actual, represented by a metric like mean squared error (MSE). It is important to note that MSE does not take censoring into account. It is also prone to being influenced by outliers, especially when the underlying distribution is skewed like the gamma distribution. So it is not an ideal criterion in this case.

A potential way is to compare the claim reserve process before and after the model is implemented to see if the reserve amounts have become closer to the actual payments over time. Here is one framework.

1. Get a sample of claims before the model is implemented; record their reserves when the claims are 3, 6, 9, ..., months old.

2. Monitor the claims and record their ultimate payment amounts when they are eventually all closed or censored after a very long time.

3. Compare the reserves at 3, 6, 9, ..., months to the ultimate payment to see if the variation has reduced over time.

4. Select a comparable sample after the model is implemented and repeat steps 1 to 3 above.

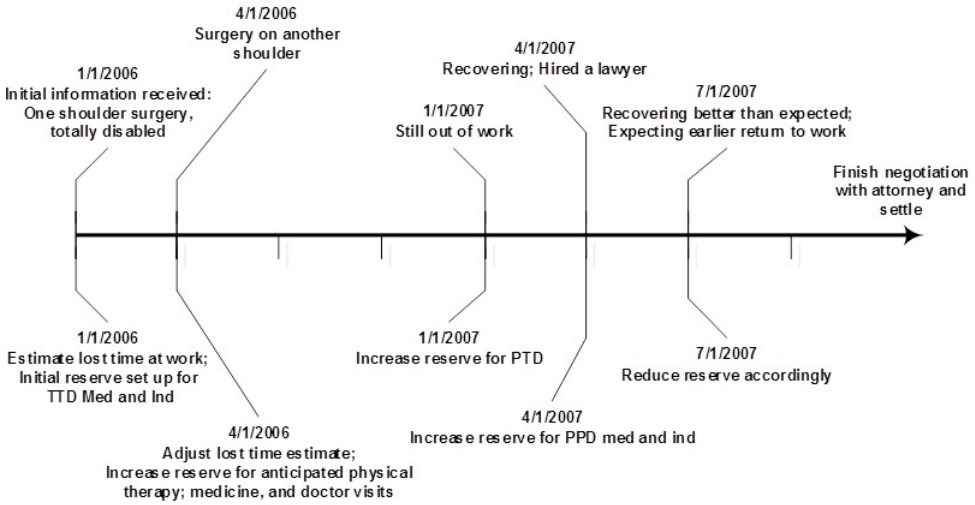
5. Compare results from step 4 with step 3 to see if the model has made significant difference.

Care must be taken in carrying out the approach above. First, the before and after samples must be comparable. For example, if the before sample is collected

during normal times while the after is collected during hurricane seasons, the nature of the claims may be different, compounding the model effect with the hurricane effect. Second, censoring will bring bias. We should wait as long as possible for the claims to close and only censor the data as the last resort in this test. Third, the sample size needs to be sufficiently large to enable us to separate the signal from the noise.

Appendix A

Time Dependence in Data



Appendix B

Fitting Gamma-Based AFT Using PROC LIFEREG

Equation (2.1.1) on page 11 shows the p.d.f. of a random variable of the familiar 2-parameter gamma distribution. The corresponding p.d.f. for a generalized gamma-distributed random variable is

$$\frac{p/a^d}{\Gamma(d/p)} y_i^{d-1} e^{-(y_i/a)^p} \tag{B.0.1}$$

where $y_i > 0$, $a > 0$, $d > 0$, and $p > 0$. When $p = 1$, the generalized gamma distribution becomes the gamma distribution.

PROC LIFEREG in SAS parametrizes B.0.1 differently as follows,

$$\frac{\delta^{1-\frac{2}{\delta^2}}}{\Gamma\left(\frac{1}{\delta^2}\right)} v_i^{\frac{1}{\delta}-1} e^{-\left(v_i \big/ \delta^{\frac{2}{\delta}}\right)^{\delta}} \tag{B.0.2}$$

where $v > 0$, $\delta > 0$.

Moreover, $\log y_i = \mu_i + \sigma \log v_i$ where $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}^*$. So $y_i = e^{\mu_i} v_i^{\sigma}$, $v_i = \exp\left(\frac{\log y_i - \mu_i}{\sigma}\right)$ and $\frac{\partial}{\partial y_i} v_i = \frac{v_i}{y_i \sigma}$. The p.d.f. of y_i is therefore

$$\begin{aligned} g(y_i) &= \frac{v_i}{y_i \sigma} \frac{\delta}{v_i \Gamma\left(\frac{1}{\delta^2}\right)} \left(\frac{v_i^{\delta}}{\delta^2}\right)^{1/\delta^2} e^{-\left(v_i \big/ \delta^{2/\delta}\right)^{\delta}} \\ &= \frac{\delta}{\sigma \Gamma\left(\frac{1}{\delta^2}\right)} y_i^{\frac{1}{\sigma\delta} - 1} e^{-\left(\frac{y_i}{e^{\mu_i} \delta^{2\sigma/\delta}}\right)^{\delta/\sigma}} \left(\frac{1}{\delta^2}\right)^{\frac{1}{\delta^2}} \left(\frac{1}{e^{\mu_i}}\right)^{\frac{1}{\sigma\delta}} \end{aligned} \tag{B.0.3}$$

Comparing Equation (B.0.3) to (B.0.1), we have , $p = \delta/\sigma$, $d = (\sigma\delta)^{-1}$, $a = e^{\mu_i} \delta^{2\sigma/\delta}$. When $p = 1$ or $\delta = \sigma$, $g(y_i)$ becomes the p.d.f. of the 2-parameter gamma distribution

$$\frac{1}{\Gamma\left(\frac{1}{\delta^2}\right) \left(e^{\mu_i} \delta^2\right)^{\frac{1}{\delta^2}}} y_i^{\frac{1}{\delta^2} - 1} e^{-y_i \big/ \left(e^{\mu_i} \delta^2\right)} \tag{B.0.4}$$

where $\alpha = 1/\delta^2$ is the shape parameter and $\theta_i = e^{\mu_i} \delta^2$ is the usual scale parameter.

However, PROC LIFEREG calls $\delta = 1/\sqrt[2]{\alpha}$ the shape parameter and σ the scale parameter. To fit an AFT based on the 2-parameter gamma, therefore, we need to use these model statement options *dist=gamma noshape1 shape1=*$1/\sqrt{\alpha}$ *noscale scale=*$1/\sqrt{\alpha}$ [16, Tsiatis and Zhang, P. 118] and search for the common value of δ and σ that maximizes the likelihood.

Using the 3-month payment model in Chapter 3 as an example, we first find the best 3-parameter gamma model (Tables 3.2.4 and 3.2.5) with MLEs $\hat{\delta} = 0.2722$ and $\hat{\sigma} = 1.0969$ and then search for the common MLE $\hat{\delta} = \hat{\sigma}$ in the vicinity of $[0.2722, 1.0969]$ for the 2-parameter gamma.

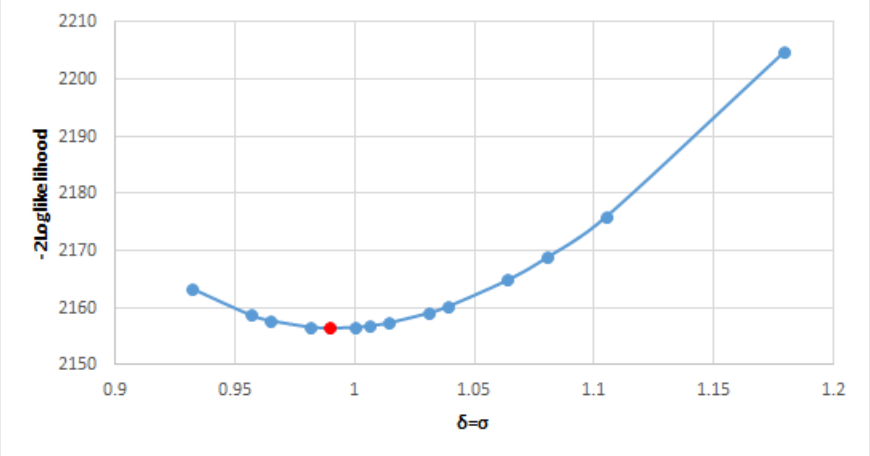
From Table B.0.1 on page 93 and Figure B.0.1 on page 93, we see that the -2Loglikelihood based on the 2-parameter gamma distribution is maximized when $\hat{\delta} = \hat{\sigma} = 0.9897$, where the corresponding gamma shape parameter is 1.0209, very

close to that of the exponential distribution.

Table B.0.1: -2Loglikelihood By $\hat{\delta} = \hat{\sigma}$ in gamma Dist.
3-Month Payment Model

Gamma Shape	$\hat{\delta} = \hat{\sigma}$	-2LL
0.719	1.1793	2204.724
0.8188	1.1051	2175.944
0.8567	1.0804	2168.8
0.8835	1.0639	2164.834
0.9262	1.0391	2160.213
0.941	1.0309	2159.054
0.9718	1.0144	2157.345
0.9877	1.0062	2156.814
1	1	2156.562
1.0209	0.9897	2156.437
1.0381	0.9815	2156.609
1.0739	0.965	2157.727
1.0926	0.9567	2158.689
1.1512	0.932	2163.346

Figure B.0.1: Search for MLE $\hat{\delta} = \hat{\sigma}$ in Gamma Dist.
3-Month Payment Model



For the 6-month payment model M_3, the results based on the generalized gamma distribution are included in Tables 3.2.10 on page 36 and 3.2.11 on page 37 .

For the 2-parameter gamma distribution, Table B.0.2 and Figure B.0.2 show the MLE for $\hat{\delta} = \hat{\sigma}$ is 0.6601, corresponding to a gamma shape of 2.295. This is different from the selected gamma shape parameter for M_3, which is 1.

Figure B.0.2: Search for MLE $\hat{\delta} = \hat{\sigma}$ in Gamma Dist.
6-Month Payment Model

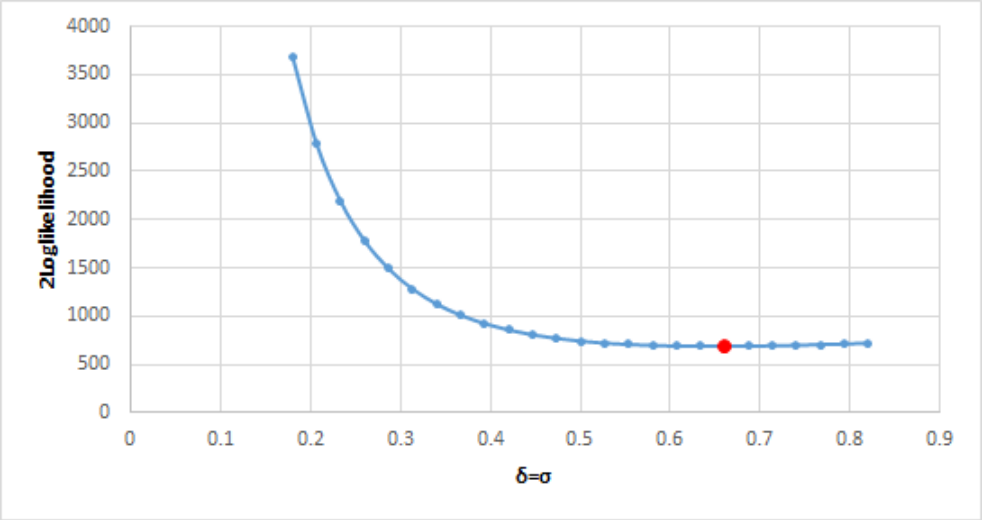


Table B.0.2: -2Loglikelihood By $\hat{\delta} = \hat{\sigma}$ in Gamma Dist.
6-Month Payment Model

Gamma Shape	$\hat{\delta} = \hat{\sigma}$	-2LL
30.8171	0.1801	3686.0471
23.3825	0.2068	2784.1209
18.3465	0.2335	2189.085
14.7781	0.2601	1780.3162
12.1579	0.2868	1490.786
10.1775	0.3135	1280.8867
8.6443	0.3401	1126.0287
7.4332	0.3668	1010.3142
6.4598	0.3935	923.1105
5.6658	0.4201	857.0989
5.0097	0.4468	807.1156
4.4613	0.4734	769.4391
3.9983	0.5001	741.3361
3.6038	0.5268	720.7661
3.2649	0.5534	706.1835
2.9716	0.5801	696.4021
2.7162	0.6068	690.5007
2.4923	0.6334	687.7567
2.295	0.6601	687.5974
2.1203	0.6868	689.5655
1.9648	0.7134	693.2924
1.8257	0.7401	698.4794
1.701	0.7667	704.8825
1.5886	0.7934	712.301

Appendix C

PROC MCMC Topics

C.1 6-Month Duration Model

Below is the code that runs the MCMC simulation for Model 5 of A_6.

```
1 title "Model 5: 6-Month Duration, Weibull Using Lifereg
    Parameterization";
2 title2 "Non-Informative Prior for Beta, not starting at
    current MLEs";
3 title3 "ntu=10000 nbi=50000 nmc=50000 thin=5";
4
5 ods graphics on;
6 ods output postsummaries=pst_smry_dur6m_5;
7 Proc MCMC data=prdt_6m outpost=pst_smpl_dur6m_5 seed=1 ntu
    =10000 nbi=50000 nmc=50000 thin=5 propcov=quanew monitor
```

```
 =(_parms_) DIC;
8  parms sigma 0.01 beta0 0 beta1 0 beta2 0;
9  **the starting values for the parameters are MLEs from the
       6-month model **;
10 **They are the initial values for sampling from the
     proposal densities **;
11 **sigma is the scale parameter of SEV distribution.
                          **;
12
13 prior beta: ~ normal(0, var=10000);
14 **non-informative priors **;
15 prior sigma ~ normal (0, sd=10000, lower=0);
16 **non-informative priors **;
17
18 beginnodata;
19   alpha = 1/sigma;
20 **alpha= 1/sigma is the shape parameter of the
     corresponding Weibull distribution **;
21 endnodata;
22
23 mu = beta0 + beta1*age + beta2*cum_trmt2;
24
25 **The definition below assumes independence among
     observations and thus additivity of log-likelihood of
```

```
      the sample.    **;
26    ll = cls_ind*(log(alpha) + (alpha-1)*log(dur_day) -alpha*
      mu) -(dur_day**alpha)*exp(mu*(-1)*alpha);
27
28    model general(ll);
29  Run;
30  ods graphics off;
31  ods output close;
32  Title;
```

The following code uses the posterior estimates from Model 5 above to generate fitted values in Proc Lifereg to obtain the log-likelihood value.

```
1   title "Model 5: 6-Month Duration Scored Using Bayesian
      Coefficient Estimates";
2   Data ini5;
3     intercept = 5.6269;
4     age       = 0.00311;
5     cum_trmt2 = 0.000696;
6     _scale_   = 0.4191;
7   Run;
8
9   *ods output ParameterEstimates=para_dur_6m;
10  Proc Lifereg data=prdt_6m outest=b_est_dur6m_5 inest=ini5;
11    model dur_day*cls_ind(0)=age  cum_trmt2 /dist=weibull covb
        corrb maxiter=0;
```

```
12    output out=prdt_6m_5 p=pred_dur_b xbeta=xb_dur_b quantiles
         =0.5;
13  Run;
14  title;
```

C.2 6-Month Payment Model

Here's the code for Model 2 MCMC simulation for M_6.

```
1   title "Model 2: 6-Month Payment, Gamma Using Lifereg
       Parameterization";
2   title2 "Non-Informative Prior for Beta, not starting at
       current MLEs";
3   title3 "ntu=10000 nbi=50000 nmc=50000 thin=5";
4
5   ods graphics on;
6   ods output postsummaries=pst_smry_pay6m_2;
7   Proc mcmc data=prdt_6m_5(where=(CLMNT_Total_Paid>0)) outpost
       =pst_smpl_pay6m_2
8             seed=1 ntu=10000 nbi=50000 nmc=50000 thin=5
                 propcov=quanew monitor=(_parms_) DIC;
9
10    Array beta[5] beta0-beta4;
11    Array data[5] 1 age xb_dur_b cum_pay3 cum_ASTH2;
12
```

```
13  parms delta 0.01 sigma 0.01 beta0-beta4 0;
14  **the starting values for the the parameters are MLEs from
        the 6-month model **;
15  **They are the initial values for sampling from the
       proposal densities **;
16  **sigma is the scale parameter of SEV distribution.
                            **;
17
18  prior beta: ~ normal(0, var=10000);
19  **non-informative priors **;
20  prior delta sigma ~ normal (0, sd=10000, lower=0);
21  **non-informative priors **;
22
23
24  beginnodata;
25    d = 1/delta;
26          a = 1/sigma;
27  endnodata;
28  **perform division here to reduce the number of divisions
       to be performed*;
29  **in ll later.


       *;
30
```

```
31   call mult(beta, data, mu);
32
33   **The definition below assumes idenpendence among
        observations and thus **;
34   **additivity of log-likihood of the sample.
                                        **;
35   ll = cls_ind*(log(delta) +log(a) - lgamma(d**2) + ((a*d)
        -1)*log(CLMNT_Total_Paid)
36                  - ( ((CLMNT_Total_Paid**(delta*a))*(d**2))/
                        exp(mu*delta*a) )  -(2*(d**2))*log(delta)
37                  - mu*a*d) +
38          (1-cls_ind)*(logsdf('GAMMA', (CLMNT_Total_Paid**(
               delta*a))*(d**2)/exp(mu*delta*a), d**2))
39          ;
40
41   model general(ll);
42 Run;
43 ods graphics off;
44 ods output close;
45 Title;
```

The code below generates fitted values in Proc Lifereg using the posterior parameter estimates from the Bayesian M_6.

```
1 Title "Model 2: 6-Month Payment Scored Using Bayesian
     Coefficient Estimates";
```

```
2  Data p_ini2;
3    Intercept      =        -9.1103;
4    age            =        -0.0109;
5    xb_dur_b       =        3.1018;
6    cum_pay3       =        0.0487;
7    cum_ASTH2      =        0.0768;
8    _Scale_        =        0.6954;
9    _Shape1_       =        0.1786;
10
11 Run;
12
13 *ods output ParameterEstimates=para_dur_6m;
14 Proc lifereg data=prdt_6m_5 outest=b_est_pay6m_3g2 inest=p_
     ini2;
15   model clmnt_total_paid*cls_ind(0)=age xb_dur_b cum_pay3
       cum_ASTH2 /dist=gamma covb CORRB maxiter=0;
16   output out=pay_6m_2 p=pred_pay_b xbeta=xb_pay_b quantiles
       =0.5;
17 Run;
18 title;
```

Appendix D

MCMC on CLM and GLMM with Censoring

Both simulations use the remaining payment as the target.

D.1 CLM

Here is the SAS code for the CLM Model 1D.

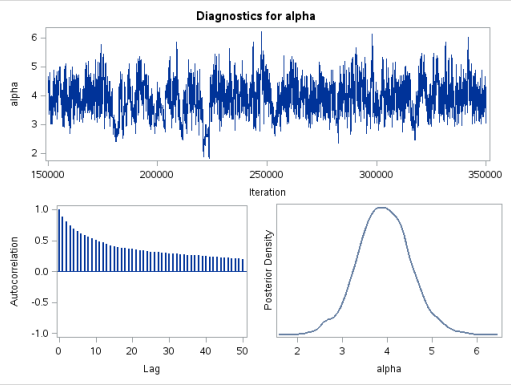
```
1 title "CLM Model 1d: Remaining Payment, CLM Likelihood";
2 title2 "age, cum_pay3";
3 title3 "ntu=50000 nbi=150000 nmc=200000 thin=20";ods
    graphics on;
4 ods output postsummaries=pst_smry_clm6m_1;
5 Proc mcmc data=pay_clm2 outpost=pst_smpl_clm6m_1
6           seed=1 ntu=50000 nbi=150000 nmc=200000 thin=20
```

```
            propcov=quanew monitor=(beta0-beta2 alpha sigma
            ) DIC;
7
8   Array beta[3] beta0-beta2;
9   Array zeta0[*] z0_1-z0_20;
10  Array zeta1[*] z1_1-z1_20;
11  Array zeta2[*] z2_1-z2_20;
12
13  **Array data[3] 1 age cum_pay3;
14
15  parms alpha 0.01  beta: 0 z: 0;
16  parms sigma 0.01;
17  **alpha is the common Gamma shape parameter**;
18  **sigma is the standard deviation for random effect z's
       distribution **;
19
20 beginnodata;
21 s2 = sigma**2;
22 ls = -log(sigma);
23 endnodata;
24
25 prior beta: ~ normal(0, var=100);
26 prior alpha ~ normal(0, sd=100, lower=0);
27 prior z: ~ normal(0, sd=sigma);
```

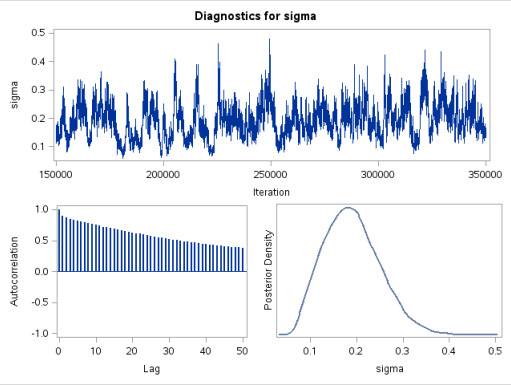
```
28 prior sigma ~ general(ls);
29 **sigma's prior is 1/sigma **;
30
31   **call mult(beta, data, mu);
32   mu = (beta0+zeta0[indx]) + (beta1+zeta1[indx])*age + (
        beta2+zeta2[indx])*cum_pay3;
33   *mu = (beta0+zeta0[indx]) + (beta1+zeta1[indx])*age + (
        beta2+zeta2[indx])*cum_trmt2;
34   theta=exp(mu);
35
36   **The definition below assumes idenpendence among
        observations and thus **;
37   **additivity of log-likihood of the sample.
                                    **;
38
39   ll = cls_ind*lpdfgamma(pay_remain, alpha, theta) +
40        (1-cls_ind)*logsdf('GAMMA', pay_remain, alpha, theta)
             - logsdf('GAMMA', cum_pay2, alpha, theta);
41
42   model general(ll);
43 Run;
44 ods graphics off;
45 ods output close;
46 Title;
```

Figure D.1.1: CLM MCMC Diagnostics Graphs I

(a) Alpha



(b) Sigma



Note: *(a)* Alpha is the common gamma shape parameter; *(b)* Sigma is the common standard deviation of the random effects.

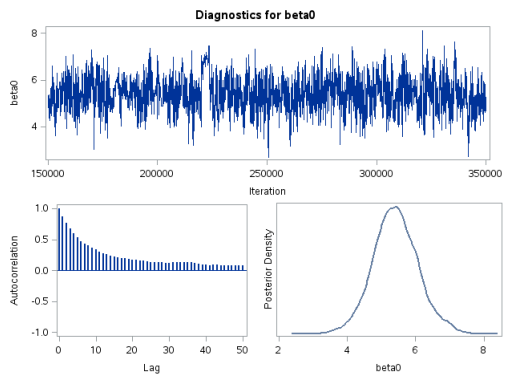
Figures D.1.1 and D.1.2 show although the sampling posterior distributions are starting to resemble the normal distributions, the autocorrelations are still high after 50 lags. Longer runs and more thinning might improve the results but will cause more strain of the computing resources, which are shared.

Table D.1.1 shares similar information. The MCSE/RD ratios are not too small while the effective sample size (ESS) and the efficiency are quite low.

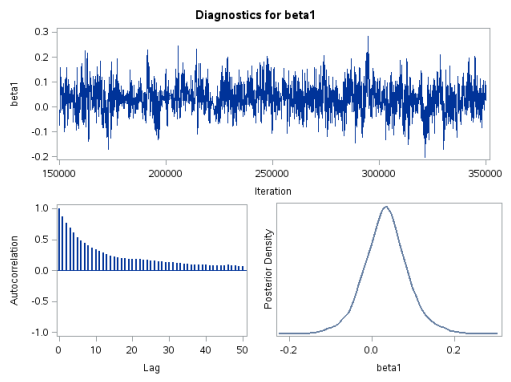
Table D.1.2 shows the HPD interval for Age contains 0, indicating it is not significant at the 0.05 level. In additional, the coefficient for Cum_Pay3 is negative. This is not surprising since we are now fitting the remaining payment, which goes down as time goes on and the more payment has already been made.

Figure D.1.2: CLM MCMC Diagnostics Graphs II

(a) Intercept



(b) Age



(c) Cum_Pay

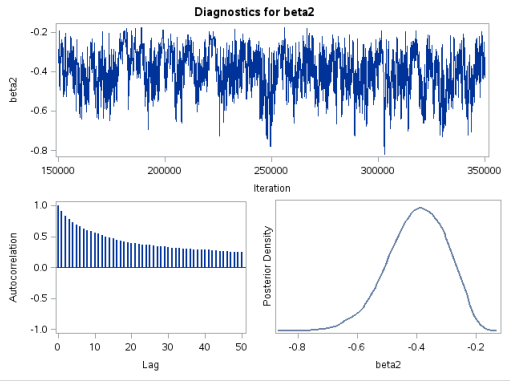


Table D.1.1: CLM MCMC Diagnostics

(a) Monte Carlo Standard Errors

Parameter	MCSE	Standard Deviation	MCSE/SD
beta0	0.0326	0.6594	0.0495
beta1	0.00277	0.0538	0.0515
beta2	0.00827	0.0982	0.0842
alpha	0.0435	0.5829	0.0747
sigma	0.00585	0.0597	0.0981

(b) Posterior Autocorrelations

Parameter	Lag 1	Lag 5	Lag 10	Lag 50
beta0	0.8695	0.5274	0.3343	0.0746
beta1	0.8758	0.5371	0.3396	0.0697
beta2	0.906	0.6908	0.5626	0.2439
alpha	0.8882	0.6487	0.5086	0.1998
sigma	0.8919	0.8181	0.7488	0.3808

(c) Effective Sample Sizes

Parameter	ESS	Autocorrelation Time	Efficiency
beta0	408.7	24.4653	0.0409
beta1	376.9	26.5355	0.0377
beta2	141	70.901	0.0141
alpha	179.2	55.8134	0.0179
sigma	104	96.1744	0.0104

Table D.1.2: CLM MCMC Output

(a) Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
beta0	10000	5.4123	0.6594	4.9829	5.4026	5.839
beta1	10000	0.0345	0.0538	0.00127	0.0341	0.0666
beta2	10000	-0.3987	0.0982	-0.4629	-0.3924	-0.326
alpha	10000	3.9217	0.5829	3.534	3.9187	4.3036
sigma	10000	0.1902	0.0597	0.1458	0.186	0.229

(b) Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
beta0	0.05	4.1157	6.7947	4.1335	6.8046
beta1	0.05	-0.073	0.1471	-0.073	0.1468
beta2	0.05	-0.6121	-0.2326	-0.5899	-0.2223
alpha	0.05	2.7392	5.1057	2.824	5.1673
sigma	0.05	0.0889	0.3179	0.0828	0.304

D.2 GLMM

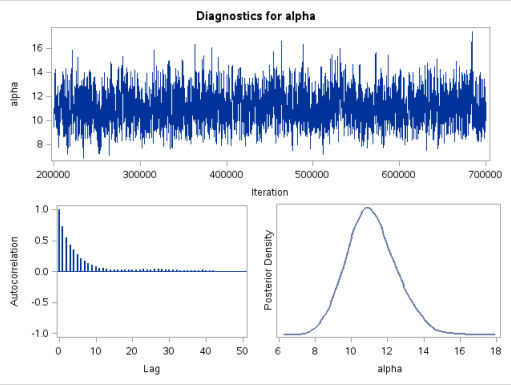
And below is the code for the GLMM Model 2E.

```
1  title "CLM Model 2e: Remaining Payment, GLMM Likelihood";
2  title2 "age, cum_pay3";
3  title3 "ntu=50000 nbi=200000 nmc=500000 thin=50";
4  ods graphics on;
5  ods output postsummaries=pst_smry_clm6m_1;
6  Proc mcmc data=pay_clm2 outpost=pst_smpl_clm6m_1 seed=1 ntu
       =50000 nbi=200000 nmc=500000 thin=50 propcov=quanew
       monitor=(beta0-beta2 alpha sigma) DIC;
7
8    Array beta[3] beta0-beta2;
9    Array zeta0[*] z0_1-z0_20;
10   Array zeta1[*] z1_1-z1_20;
11   Array zeta2[*] z2_1-z2_20;
12   **Array data[3] 1 age cum_pay3;
13
14   parms alpha 0.01  beta: 0 z: 0;
15   parms sigma 0.01;
16   **alpha is the common Gamma shape parameter  **;
17   **sigma is the standard deviation for random effect z's
        distribution**;
18
19 beginnodata;
```

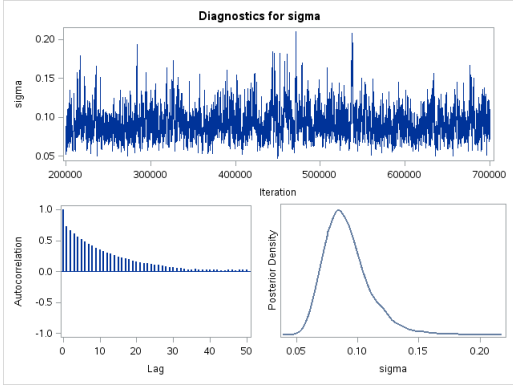
```
20 s2 = sigma**2;
21 ls = -log(sigma);
22 endnodata;
23
24   prior beta: ~ normal(0, var=100);
25   prior alpha ~ normal(0, sd=100, lower=0);
26   prior z: ~ normal(0, sd=sigma);
27   prior sigma ~ general(ls);
28   **sigma's prior is 1/sigma **;
29
30    **call mult(beta, data, mu);
31    mu = (beta0+zeta0[indx]) + (beta1+zeta1[indx])*age + (
          beta2+zeta2[indx])*cum_pay3;
32    *mu = (beta0+zeta0[indx]) + (beta1+zeta1[indx])*age + (
          beta2+zeta2[indx])*cum_trmt2;
33    theta=exp(mu);
34
35    ll = cls_ind*lpdfgamma(pay_remain, alpha, theta) +
36         (1-cls_ind)*logsdf('GAMMA', pay_remain, alpha, theta)
               /* - logsdf('GAMMA', cum_pay2, alpha, theta)*/;
37
38    model general(ll);
39 Run;
40 ods graphics off; ods output close; Title;
```

Figure D.2.1: GLMM MCMC Diagnostics Graphs I

(a) Alpha



(b) Sigma



Note: *(a)* Alpha is the common gamma shape parameter; *(b)* Sigma is the common standard deviation of the random effects.

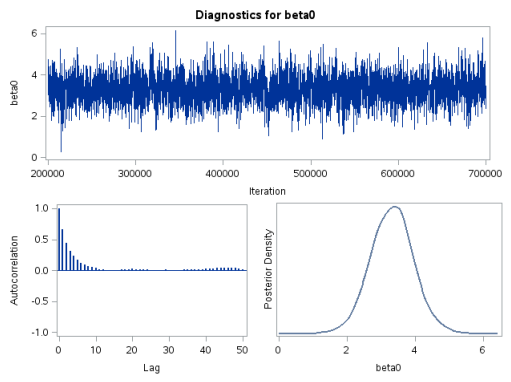
Diagnostics in Figures D.2.1 and D.2.2 are satisfactory with the trace plot being stable, autocorrelations declining quickly and the posterior sampling distributions resembling the normal distribution.

The diagnostic details in Table D.2.1 are generally good as well.

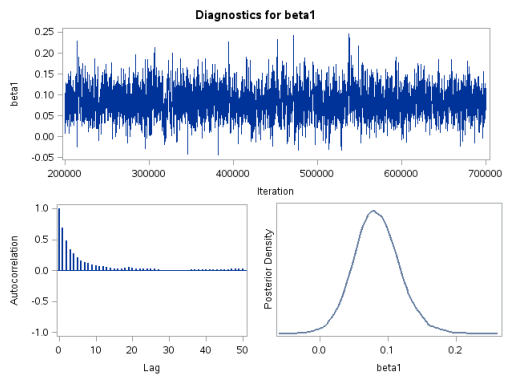
Table D.2.2 shows the signs for the coefficient estimates here are the same as in CLM simulation. Further, Age's HPD interval does not include 0, indicating it is significant.

Figure D.2.2: GLMM MCMC Diagnostics Graphs II

(a) Intercept



(b) Age



(c) Cum_Pay

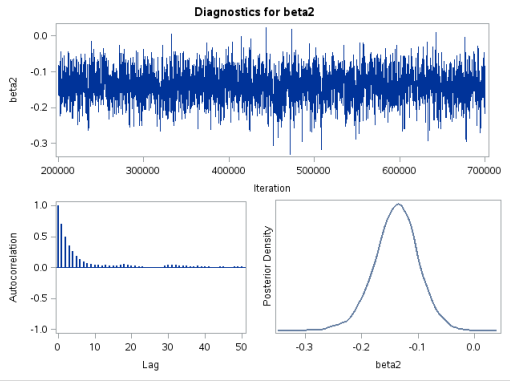


Table D.2.1: GLMM MCMC Diagnostics

(a) Monte Carlo Standard Errors

Parameter	MCSE	Standard Deviation	MCSE/SD
beta0	0.0149	0.6392	0.0233
beta1	0.000925	0.0341	0.0271
beta2	0.00102	0.0398	0.0256
alpha	0.0403	1.3582	0.0297
sigma	0.000839	0.0189	0.0443

(b) Posterior Autocorrelations

Parameter	Lag 1	Lag 5	Lag 10	Lag 50
beta0	0.6584	0.1688	0.0344	0.0196
beta1	0.6824	0.2094	0.0789	0.0259
beta2	0.6964	0.1885	0.0396	0.0207
alpha	0.7263	0.2758	0.0763	0.0003
sigma	0.7272	0.5143	0.3509	0.0255

(c) Effective Sample Sizes

Parameter	ESS	Autocorrelation Time	Efficiency
beta0	1835	5.4496	0.1835
beta1	1358	7.3638	0.1358
beta2	1522.8	6.5666	0.1523
alpha	1137.4	8.7917	0.1137
sigma	509.5	19.6255	0.051

Table D.2.2: GLMM MCMC Output

(a) Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
beta0	10000	5.4123	0.6594	4.9829	5.4026	5.839
beta1	10000	0.0345	0.0538	0.00127	0.0341	0.0666
beta2	10000	-0.3987	0.0982	-0.4629	-0.3924	-0.326
alpha	10000	3.9217	0.5829	3.534	3.9187	4.3036
sigma	10000	0.1902	0.0597	0.1458	0.186	0.229

(b) Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
beta0	0.05	4.1157	6.7947	4.1335	6.8046
beta1	0.05	-0.073	0.1471	-0.073	0.1468
beta2	0.05	-0.6121	-0.2326	-0.5899	-0.2223
alpha	0.05	2.7392	5.1057	2.824	5.1673
sigma	0.05	0.0889	0.3179	0.0828	0.304

Bibliography

[1] Jonathan Buckley and Ian James, *Linear regression with censored data*, Biometrika **66** (1979), 429–436, DOI :10.1093/biomet/66.3.429.

[2] J. G. Ibrahim and M.-H. Chen, *Power prior distributions for regression models*, Statistical Science **15** (2000), 4660.

[3] SAS Institute Inc., *Sas/stat(r) 9.22 user's guide*, `https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#glimmix_toc.htm`, November 2008, Retrieved 10/5/2016.

[4] ———, *Sas/stat(r) 9.2 user's guide*, `https://support.sas.com/documentation/cdl/en/statugbayesian/61755/PDF/default/statugbayesian.pdf`, February 2009, Retrieved 10/31/2016.

[5] ———, *Sas/stat(r) 14.1 user's guide*, `https://support.sas.com/documentation/onlinedoc/stat/141/nlmixed.pdf`, July 2015, Retrieved 11/7/2016.

[6] Zhezhen Jin, D. Y. Li, and Zhiliang Ying, *On least-squares regression with censored data*, Biometrika **93** (2006), 147–161.

[7] Tze Leung Lai and Zhiliang Ying, *Large sample theory of a modified Buckley-James estimator for regression analysis with censored data*, The Annals of Statistics **19** (Sept 1991), 1370–1402.

[8] Steffen Lauritzen, *Bayesian asymptotics*, `http://www.stats.ox.ac.uk/~steffen/teaching/bs2HT8/bayes.pdf`, May 2008, Lecture 8, BS2 Statistical Inference.

[9] Y. Lee and J.A. Nelder, *Hierarchical generalized linear models*, Journal of the Royal Statistical Society **58:4** (1996), 619–678, B (Methodological).

[10] P. McCullagh and J. A. Nelder, *Generalized linear model*, 2nd ed., vol. 37, Chapman & Hall/CRC, Boca Raton, 1989.

[11] Rupert Miller and Jerry Halpern, *Regression with censored data*, Biometrika **69** (1982), 521–31.

[12] S. James Press, *Subjective and objective Bayesian statistics: Principles, models, and applications*, 2nd ed., Wiley-Interscience, Hoboken, NJ, 2003.

[13] Nalini Ravishanker and Dipak K. Dey, *A first course in linear model theory*, Chapman & Hall/CRC, Boca Raton, 2001.

[14] SAS Global Forum 2009, *Bayesian modeling using the MCMC procedure*, Statistics and Data Analysis, no. 257-2009, 2009.

[15] Anders Skrondal and Sophia Rabe-Hesketh, *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*, Chapman & Hall/CRC, Boca Raton, 2004.

[16] Anastasios Tsiatis and Daowen Zhang, *Analysis of survival data*, `http://www4.stat.ncsu.edu/~dzhang2/st745/index.html`, Spring 2005, Lecture Notes, Chapter 5.

[17] Yanhua Wang, Shuyuan He, Lixing Zhu, and Kam C Yuen, *Asymptotics for a censored generalized linear model with unknown link function*, Probab. Theory Relat. Fields **138** (2007), 235–267, DOI : 10.1007/s00440-006-0022-5.

[18] L.J. Wei, *The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis*, Statistics in Medicine **11** (1992), 1871–1879.

[19] Lili Yu, Ruifeng Yu, and Liang Liu, *Quasi-likelihood for right-censored data in the generalized linear model*, Communications in Statistics - Theory and Methods **38:13** (2009), 2187–2200, DOI : 10.1080/03610920802499504.

[20] Scott L. Zeger, Kung-Yee Liang, and Paul S. Albert, *Models for longitudinal data: A generalized estimating equation approach*, Biometrics **44:4** (1988), 1049–1060.

[21] Mai Zhou and Gang Li, *Empirical likelihood analysis of the Buckley-James estimator*, Journal of Multivariate Analysis **99** (2008), 649–664, DOI :10.1016/j.jmva.2007.02.007.

Index